UNIVERSAL PARALINGUISTIC SPEECH REPRESENTATIONS USING SELF-SUPERVISED CONFORMERS

Joel Shor[1]*, Aren Jansen*[2]*, Wei Han*[2]*, Daniel Park*[2]*, Yu Zhang*[2]

Verily Life Sciences, Boston, USA[1] and Mountain View, California, USA[2]
joelshor@verily.com

ABSTRACT

Many speech applications require understanding aspects beyond the words being spoken, such as recognizing emotion, detecting whether the speaker is wearing a mask, or distinguishing real from synthetic speech. In this work, we introduce a new state-of-the-art paralinguistic representation derived from large-scale, fully self-supervised training of a 600M+ parameter Conformer-based architecture. We benchmark on a diverse set of speech tasks and demonstrate that simple linear classifiers trained on top of our time-averaged representation outperform nearly all previous results, in some cases by large margins. Our analyses of context-window size demonstrate that, surprisingly, 2 second context-windows achieve 96% the performance of the Conformers that use the full long-term context on 7 out of 9 tasks. Furthermore, while the best per-task representations are extracted internally in the network, stable performance across several layers allows a single universal representation to reach near optimal performance on all tasks.

Index Terms— speech, representation learning, self-supervised learning, paralinguistics, transformer

1. INTRODUCTION

Powerful representations of data are useful in a number of ways. They improve model performance on small datasets by transferring data-driven insights from larger datasets. The models that create representations can also be used as pre-training for improved performance. If the model that generates the representation is non-reversible, then the representations can unlock applications in some privacy-sensitive scenarios. In this paper, we significantly improve state-of-the-art representations for paralinguistic speech tasks.

There are a number of promising data-driven speech representations. Some directions include self-supervised contrastive learning [1, 2, 3], predictive coding [4, 5], masked-unit prediction [6], multi-task learning [7], multimodal coincidence [8, 9], and intermediate representations from a supervised task [10, 11]. One of the most promising objectives for representation learning for speech recognition was proposed in the recent Wav2Vec 2.0 [12] framework, which combined Transformers [13] and a self-supervised contrastive learning objective [5]. The Wav2Vec 2.0 training objective was subsequently combined with more powerful Conformer architectures, producing large improvements in semi-supervised speech recognition applications [14, 15, 16]. This paper explores the use of these Conformer-based models to define fixed representations for non-ASR speech analysis and paralinguistics tasks. To fully evaluate the potential of these models, we evaluate several model sizes and pretraining datasets combinations.

Recent work to establish a common benchmark has made it possible to directly compare speech representations [1, 17]. In this work, we use the Non-Semantic Speech Benchmark (NOSS) [1], a collection of publicly available non-semantic speech tasks including speech emotion recognition, language identification, and speaker identification. Following [18], we include masked speech detection [19]. We also include three new tasks: synthetic speech detection [20], an additional speech emotion recognition dataset [21], and dysarthria classification [22]. Our work further establishes the usefulness of these embeddings over classical paralinguistic features, and can be used to improve other transfer-learning speech applications like voice imitation [23] and personalized ASR [24].

Finally, our work explores the impact of context window size on performance. We show that 2-second context windows are sufficient for nearly all tasks, but further context truncation can lead to large losses in performance. Furthermore, we analyzed the range of embeddings produced by the sequence of Conformer blocks that define the encoder, demonstrating stable performance over a large portion of the network regardless of architecture complexity. Using Centered Kernel Alignment (CKA) analysis [25, 26], we further demonstrate that the representations defined by this range of blocks are surprisingly similar, both within and (to lesser degree) across architectures.

The main contributions of this paper are:

1. Generate features for non-semantic speech tasks that set a new state-of-the-art (SoTA) performance on 7 of 9 tasks **using only time-averaged features and linear classification models**
2. Analyze the performance versus context window size tradeoff, and show that 2-second context windows are sufficient
3. Perform a more extensive embedding comparison than previously done, both in terms of downstream tasks and embeddings compared. Using a per-example analysis, we demonstrate that our embedding is strictly better than previous ones
4. Demonstrate that similarly-performing representations in different architectures are similar in the CKA-sense

2. CONFORMER-BASED REPRESENTATIONS

2.1. Architectures

Each of our proposed paralinguistic representations is defined using a *speech encoder* comprised of a stack of convolution-augmented Transformer blocks known as Conformers [14]. Each Conformer block inserts a small depthwise separable convolutional module between the Transformer's self-attention and MLP modules, which has been shown to be highly beneficial to many recognition applications. The input to this speech encoder is the output of a 3-layer 1-dimensional convolutional *feature encoder* that is applied to 80-bin log mel spectrogram features. The spectrogams come from 16kHz audio that is resampled if necessary. Two convolutional strides of

Table 1: Comparison of models. Resnetish50 [10]. MobileNetv3 [27]. RNN-T [28]. EfficientNet [29]. Conformer [14]. AudioSet [30]. YT-U [16], LL is Libri-Light [31]. *"RA" stands for "relative attention."

Name	Architecture	Params	Training data	Labels required
YAMNet [1]	MobileNetv1	3.7M	Audioset	Y
TRILL [1]	Resnetish50	24.5M	Audioset	N
FRILL [18]	MobileNetv3	10.1M	Audioset	N
COLA [2]	EfficientNetB0	4.0M	Audioset	N
ASR Emb [11]	RNN-T	122M	-	Y
Conformer XL (No) RA* YT (LL)	Conformer	608M	YT-U (LL)	N
Conformer XXL YT (LL)	Conformer	1.0B	YT-U (LL)	N
Conformer G	Conformer	8.0B	YT-U	N

two produce a vector time series that is downsampled by a factor of 4x, yielding a frame rate that is preserved throughout the entire speech encoder.

The models are trained using the Wav2Vec 2.0 contrastive loss [12]: we first extract encoded features from the feature encoder and then use masked features as inputs to the Conformer to create context vectors. These context vectors are trained to agree with the target context vectors, obtained by applying a linear layer to the initial encoded features, by a contrastive loss. Table 1 lists the various Conformer architectures considered in our evaluation. We consider three Conformer encoder complexities defined in the original study [16], including 608 million (24 layers/8 heads/1024D output), 1.0 billion (42 layers/8 heads/1024D output), and 8.0 billion parameters (36 layers/16 heads/3072D output). Also shown are corresponding details for five baseline representations that we include in our evaluation. These cover a range of model architectures, complexities, and training objectives.

2.2. Pre-training Datasets

We use two datasets for self-supervised training of the above architectures. The first is **YT-U**, a 900k hour dataset [16] derived from YouTube. YT-U is built by first randomly collecting 3 million hours of audio from "speech-heavy" YouTube videos. The results are then segmented, and the non-speech segments are removed to yield approximately 900k hours of unlabeled audio data.

The second is **Libri-Light** [31], which contains 60k hours of audio derived from open-source audio books in the LibriVox project. It is the largest publicly available, unlabeled semi-supervised dataset to date.

3. EXPERIMENTS

3.1. Tasks: The Non-Semantic Speech Benchmark (NOSS)

In order to fairly compare representations, we benchmark each representation on the same 9 tasks (Table 2). Our tasks include most of the original NOSS benchmark tasks [1], a mask-detection task used in representation benchmarking in [18], a fake speech detection task [20], an additional speech emotion recognition task [21], and a dysarthria classification task [22]. When a single scalar is necessary (e.g. to compare embeddings), we aggregate over the performances using the "Aggregate Embedding Score", which is the average accuracy of a model, averaged across tasks.

ASVSpoof2019: We introduce the ASVSpoof2019 [20] dataset as a new task in our benchmark. This task measures a model's ability to distinguish real from synthetic speech. We use the Logical Access (LA) portion of this dataset. The LA database contains bona fide and spoofed speech generated using 17 different text-to-speech and voice conversion systems. The task is especially challenging because spoofed speech in the test set is generated using techniques not seen in training.

Table 2: Downstream evaluation datasets. *Results in our study used a subset of Voxceleb filtered according to YouTube's privacy guidelines.

Dataset	Target	Classes	Samples	Avg length (s)
VoxCeleb* [32]	Speaker ID	1,251	12,052	8.4
VoxForge [33]	Language ID	6	176,438	5.8
Speech Commands[34]	Command	12	100,503	1.0
Masked Speech [19]	Mask wearing	2	36,554	1.0
ASVSpoof [20]	Synthetic or not	2	121,461	3.2
Euphonia [22]	Dysarthria	5	15,224	6.4
CREMA-D [35]	Emotion	6	7,438	2.5
IEMOCAP [21]	Emotion	4	5,531	4.5
SAVEE [36]	Emotion	7	480	3.8

IEMOCAP: The Interactive Emotional Dyadic Motion Capture database [21] is an acted, multimodal, and multispeaker database. We use the improvised scenarios portion with categorical emotion labels. To compare fairly with previous SoTA work [17], we only use the audio component and only 4 of the 10 labels (angry, happy, neutral, and sad).

Euphonia: The Euphonia dataset [22] is a large dysarthric speech dataset. Our task uses a 661 speaker subset of 29 identical phrases with manual dysarthria labels from speech-language pathologists on their overall intelligibility using a five-point Likert scale.

3.2. Benchmark Results

For our first set of experiments, we compute embeddings from our speech representation models (Table 1) and train simple models on the NOSS tasks using the same methodology from [1]. For each pair of benchmark task and embedding, we train and evaluate a number of simple linear classification techniques (logistic regression, balanced logistic regression, linear discriminant analysis) on top of clip-level average embeddings. We choose the best performing classifier (as determined on dev set) and use it to report test performance for that (task, embedding) pair.

Table 3 shows the results of the benchmark. Like recent studies we report the performance of the best (model, layer) pair on a per-task basis. However, we also aim to establish a single universal set of features that serve all downstream tasks. Thus, we also evaluate all intermediate representations and rank order them according to the Aggregate Embedding Quality on the dev set. We then report performance on the test set in final line of Table 3. It comes from layer 12/23 of the 600M parameter YT model, without relative attention. We call this model "Conformer Applied to Paralinguistics," or "**CAP**", and we refer to the best layer as "**CAP12**." We note that this representation **was within 6% accuracy of the per-task best layer on 7 of 9 tasks**. Figure 1 shows how the aggregate embedding quality varies in this model across intermediate layers.

Linear classifiers on Conformer representations set a new SoTA on 7/9 tasks: The "best per-task" row in Table 3 shows the test set results on the representations with the best dev-set performance. Linear models on these representations set a new SoTA on 7 of the tasks, often outperforming far more complex models. Furthermore, these linear models **outperform previous SoTA models that use more modalities than just speech** (CREMA, SAVEE).

Table 3: Test performance on the NOSS Benchmark and extended tasks. "Prev SoTA" are arbitrarily complicated models, but **all other rows are linear models on time-averaged input**. †Filtered according to YouTube's privacy guidelines. We omit previous SoTA results, since they used the entire dataset. ‡Task performance is reported using unweighted average recall [19] instead of accuracy. Also, test set labels are not available, so we report accuracy on the eval set. **Uses equal error rate [20]. #The only non-public dataset. We exclude it from aggregate scores. ††Included in the table but not aggregate score, since it's less than 1/10th the size of the next smallest dataset and results have high variance. *Audio and visual features used in previous SoTA. +Prev SOTA performed cross-fold validation. We hold out speakers M05 and F05 as test. ++YAMNet uses layer 10, as in [1]. §Best per-task results are computed by taking the model/layer with the best results of the dev set, and reporting those results on the test set. If the dev set performance is better but the test results are worse, "Best per-task" can be worse than "Best overall".

Model	Voxceleb1†	Voxforge	Speech Commands	Masked Speech‡	ASVSpoof 2019**	Euphonia#	CREMA-D	IEMOCAP	SAVEE††
Prev SoTA	-	95.4 [37]	**97.9** [38]	73.0 [39]	5.11 [17]	45.9 [11]	74.0* [40]	67.6+ [17]	84.0* [36]
Baselines									
YAMNet++ [1]	10.9	79.8	78.5	59.7	9.23	43.0	66.4	57.5	69.2
TRILL [1]	12.6	84.5	77.6	65.2	7.46	48.1	65.7	54.3	65.0
FRILL [18]	13.8	78.8	74.4	67.2	7.45	46.6	71.3	57.6	63.3
COLA [2]	11.7	71.0	60.6	65.0	4.58	47.6	69.3	63.9	59.2
ASR Emb [11]	5.2	98.9	96.1	54.4	11.2	**54.5**	71.8	65.4	85.0
Conformers									
Best per-task§ (model, layer #)	**53.5** (XXL-YT, 25)	**99.8** (G-YT, 19)	97.5 (CAP, 16)	**74.2** (XL-LL RA, 5)	**2.5** (CAP, 12)	53.6 (CAP, 13)	87.2 (G, 26)	**79.2** (CAP, 15)	**92.5** (CAP, 15)
Best CAP per task (layer #)	50.3 (11)	99.7 (14)	97.5 (16)	73.4 (10)	**2.5** (12)	53.6 (13)	**88.2**§ (12)	**79.2** (15)	92.5 (15)
Best single layer (CAP12)	51.0§	99.7	97.0	68.9	**2.5**	51.5	**88.2**§	75.0	81.7

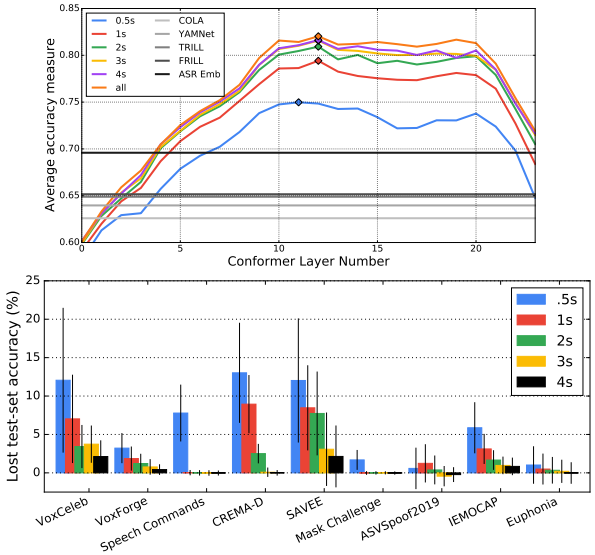

Fig. 1: **Upper)** Average test accuracy, averaged across tasks, for "CAP." X-axis is the network layer. Different lines are different chunking values. **Lower)** Absolute accuracy lost due to smaller context windows. Error bars are 1 standard deviation. Each bar is a mean over (models) x (layers) = **192 values**.

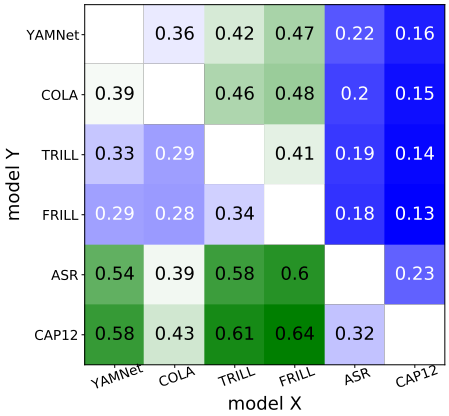

Fig. 2: Each square is the probability that Model Y correctly predicts an example given that Model X and Model Y disagree on the prediction. The result is averaged over task. Each task is an average over examples.

CAP12 significantly outperforms previous representations, especially on speech emotion recognition: CAP12 outperforms every other non-Conformer representation on every dataset we used with the lone exception of "ASR Emb" on SAVEE. Especially noteworthy are the results on CREMA-D and IEMOCAP, where **CAP12 outperforms previous embeddings by 16% and 9% respectively**.

CAP12 significantly outperforms previous single-model SoTA on ASVSpoof2019: Linear models on averaged CAP12 would've been the best single-model entry in the ASVSpoof2019 competition, and would've ranked 3rd overall [20].

CAP12 is strictly better than other representations: Since aggregate performance ignores patterns of errors, we investigate the agreement between predictions made from different embeddings on a per-example basis. Figure 2 show that when CAP12 and other embeddings disagree, CAP12 is correct 32%-64% of the time, while other embeddings are correct only 13%-23% of the time. With the exception of the supervised ASR Embedding, it is relatively uncommon for other embeddings to be correct when CAP12 is wrong.

3.3. Context window size

Our second experiment studies the role of context window size. Conformers, like Transformers, use the entire audio clip to generate embeddings, while CNN-based methods have fixed context-window sizes. To help understand how essential the large context window is for performance, we feed finite-window-sized inputs to the Conformer models, just like CNNs process input. We chunk the audio into fixed length sub-clips (e.g. 1 second), and have the Conformer

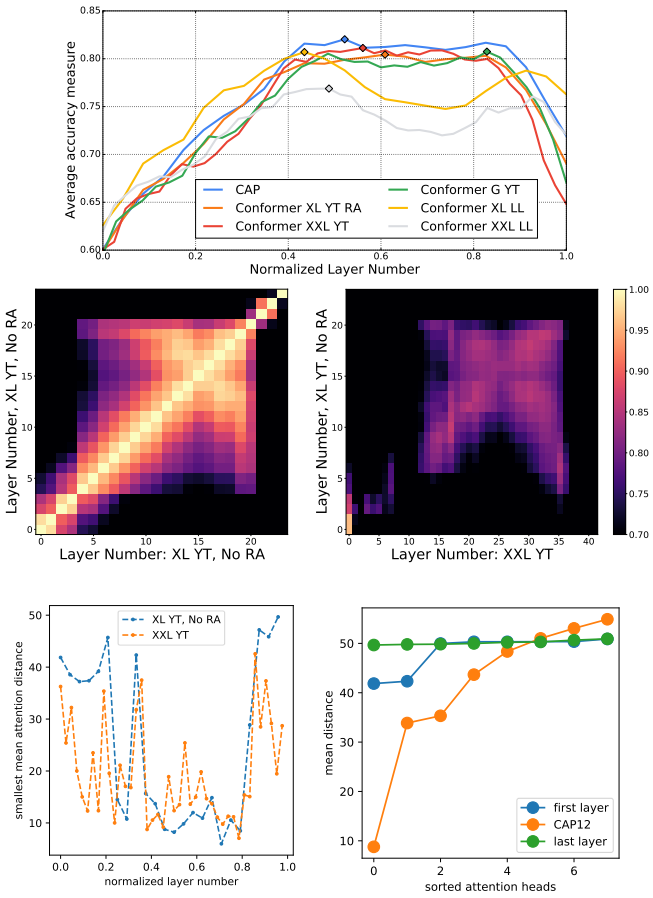

Fig. 3: **Upper)** Average test-set accuracy, averaged across tasks, for 6 different Conformer models as a function of layer index normalized to [0.0, 1.0] using (layer #) / (# of layers), where # of layers is different for different models. **Middle)** Linear CKA scores between all pairs of layers: (left) within the Conformer XL YT network and (right) across the top performing Conformer XL YT and XXL YT networks. The colormap is truncated at 0.7 as is common to both images. **Lower)** Mean attention distance: (left) the shortest attention head on every layer; (right) all attention heads on 3 representative layers of the CAP model.

model generate local embeddings independently. We then average over all local representations and evaluate the quality of the embedding on each NOSS task. We also average the loss in performance across models, tasks, and layers to determine the average effect that finite-context windows have on downstream performance. Results are shown in Figure 1.

2-second context windows are sufficient: The best-performing layer for 4 / 3 / 2 / 1 / 0.5 second context windows are 99% / 99% / 98% / 96% / 91% as accurate as the entire context window, averaged across tasks. This result is a function of the time-domain of the phenomena being studied, but is also a function of the fact that the datasets we use are known to include the signal we care about (average audio clips shown in Table 2).

3.4. Layerwise Analysis

The comparable performance of CAP12 relative to the optimal per-task embeddings identified in Table 3 suggests a high degree of representational stability across layers and architectures evaluated. Thus, our final set of experiments further probe the per-layer performance across layer and architecture. Figure 3(upper) plots the average accuracy on NOSS tasks as a function of layer for each architecture, but where layer index is normalized to a common [0, 1] scale. We observe an overall dependence on pretraining dataset, with YouTube-trained models clearly outperforming LibriLight-trained ones. However, within the YouTube models, we observe a surprisingly similar performance trajectory as we move through the normalized network position. Furthermore, for each of these models, we observe a wide performance plateau in the second half of each network.

To test whether this behavior arises from representational similarity in the models' shared performance plateau, we apply linear Centered Kernel Alignment (CKA) between pairs of layers both within and across networks, following the methodology of a recent vision Transformer study [26]. Briefly, CKA computes a $[0, 1]$-valued similarity between two Gram matrices (using an arbitrary kernel function, which we take to be linear) separately computed from two representations over the same sample of input examples (see [25] for details). Figure 3 (middle left) shows the pairwise layer similarity within the CAP network. While each layer is most similar to its neighbors, we observe a large block of similar layers in the second half of the network corresponding to the performance plateau in Figure 3 (upper). This indicates that the stable downstream performance is indeed fueled by a stable representation across these layers. While the overall similarity across XL and XXL networks is lower in Figure 3 (middle right), we again see a block of similar layers corresponding to shared performance plateau. This indicates similar characterization of paralinguistic properties in this stage of the network regardless of total network depth.

Finally, in Figure 3 (lower right), we plot mean attention distances of self attention layers to study how much temporal context each layer is aggregating over. Following [26], we compute the mean attention distance as the attention probability-weighted average temporal distance for each attention head, and average over 1k clips from [31]. We observe that higher and lower layers contain only global (long distance) attention heads, whereas middle layers have a mix of local and global ones. Interestingly, there is a clear correlation between the shortest attention distance on each layer (Figure 3, lower left) and its average accuracy on NOSS tasks (Figure 3, upper), which suggests the importance of local information for paralinguistic tasks.

4. CONCLUSION

In this paper, we introduce a class of Conformer-based self-supervised representation for speech. These representations set a new state-of-the-art performance on 7/9 paralinguistic speech tasks using only **embeddings averaged across time**, and using only **linear models on those embeddings.** Furthermore, these representations substantially outperform other speech representations despite not using labels for training. Even though the models use the entire context window to generate embeddings, we demonstrate that 2-second windows give 96% the performance of the full context window on 7 of 9 tasks, and that **these representations with 500ms context windows still outperform previous representations**. Finally, we show that Conformer models of different sizes and datasets learn comparable representations at similar parts of the network, indicating that our findings are fundamental to the problem and not a superficial artifact.

5. REFERENCES

[1] J. Shor *et al.*, "Towards learning a universal non-semantic representation of speech," in *Interspeech*, 2020, pp. 140–144.

[2] A. Saeed *et al.*, "Contrastive learning of general-purpose audio representations," in *ICASSP*, 2021, pp. 3875–3879.

[3] A. Jansen *et al.*, "Unsupervised learning of semantic audio representations," in *ICASSP*. IEEE, 2018, pp. 126–130.

[4] Y.-A. Chung *et al.*, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Interspeech*, 2019, pp. 146–150.

[5] A. van den Oord *et al.*, "Representation learning with contrastive predictive coding," 2019.

[6] W. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *CoRR*, vol. abs/2106.07447, 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[7] S. Pascual *et al.*, "Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks," in *Interspeech*, 2019, pp. 161–165.

[8] R. Arandjelovic *et al.*, "Look, listen and learn," in *ICCV*. IEEE Computer Society, 2017, pp. 609–617.

[9] A. Jansen *et al.*, "Coincidence, categorization, and consolidation: Learning to recognize sounds with minimal supervision," in *ICASSP*. IEEE, 2020, pp. 121–125.

[10] S. Hershey *et al.*, "Cnn architectures for large-scale audio classification," in *ICASSP*, 2017.

[11] S. Venugopalan *et al.*, "Comparing Supervised Models and Learned Speech Representations for Classifying Intelligibility of Disordered Speech on Selected Phrases," in *Interspeech*, 2021, pp. 4843–4847.

[12] A. Baevski *et al.*, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS*, vol. 33, 2020, pp. 12 449–12 460.

[13] A. Vaswani *et al.*, "Attention is all you need," in *NeurIPS*, I. Guyon *et al.*, Eds., vol. 30, 2017.

[14] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*. ISCA, 2020, pp. 5036–5040.

[15] Y. Zhang *et al.*, "Pushing the limits of semi-supervised learning for automatic speech recognition," 2020.

[16] ——, "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," 2021.

[17] S. wen Yang *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[18] J. Peplinski *et al.*, "FRILL: A Non-Semantic Speech Embedding for Mobile Devices," in *Interspeech*, 2021.

[19] B. Schuller *et al.*, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly emotion, Breathing & Masks," in *INTERSPEECH*, 2020.

[20] M. Todisco *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Interspeech*, 2019, pp. 1008–1012.

[21] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[22] R. L. MacDonald *et al.*, "Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia," in *Interspeech*, 2021, pp. 4833–4837.

[23] Y. Jia *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, S. Bengio *et al.*, Eds., vol. 31, 2018.

[24] J. Shor *et al.*, "Personalizing ASR for dysarthric and accented speech with limited data," *Interspeech*, Sep 2019.

[25] S. Kornblith *et al.*, "Similarity of neural network representations revisited," in *ICML*. PMLR, 2019, pp. 3519–3529.

[26] M. Raghu *et al.*, "Do Vision Transformers see like convolutional neural networks?" 2021.

[27] A. Howard *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1314–1324.

[28] Y. He *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP*, 05 2019, pp. 6381–6385.

[29] M. Tan *et al.*, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, ser. Proceedings of Machine Learning Research, K. Chaudhuri *et al.*, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114.

[30] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017, pp. 776–780.

[31] J. Kahn *et al.*, "Libri-Light: A benchmark for ASR with limited or no supervision," in *ICASSP*, 2020, pp. 7669–7673.

[32] A. Nagrani *et al.*, "Voxceleb: a large-scale speaker identification dataset," in *INTERSPEECH*, 2017.

[33] K. MacLean, "Voxforge," *Ken MacLean.[Online]. Available: http://www.voxforge.org/home.[Acedido em 2012]*, 2018.

[34] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *ArXiv e-prints*, Apr. 2018.

[35] H. Cao *et al.*, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, pp. 377–390, 2014.

[36] S. Haq *et al.*, "Speaker-dependent audio-visual emotion recognition." in *AVSP*, 2009, pp. 53–58.

[37] Sarthak *et al.*, "Spoken language identification using convnets," in *Ambient Intelligence*, I. Chatzigiannakis *et al.*, Eds. Springer International Publishing, 2019, pp. 252–265.

[38] D. Seo *et al.*, "Wav2KWS: Transfer learning from speech representations for keyword spotting," *IEEE Access*, vol. 9, pp. 80 682–80 691, 2021.

[39] J. Szep *et al.*, "Paralinguistic Classification of Mask Wearing by Image Classifiers and Fusion," in *Interspeech*, 2020, pp. 2087–2091.

[40] E. Ghaleb *et al.*, "Multimodal and temporal perception of audio-visual cues for emotion recognition," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 552–558.