Interspeech 2022
18-22 September 2022, Incheon, Korea

Personal VAD 2.0: Optimizing Personal Voice Activity Detection for On-Device Speech Recognition

Shaojin Ding, Rajeev Rikhye, Qiao Liang, Yanzhang He,
Quan Wang, Arun Narayanan, Tom O'Malley, Ian McGraw

Google LLC, USA

{shaojinding, rvrikhye, wildstone, yanzhanghe, quanw}@google.com

Abstract

Personalization of on-device speech recognition (ASR) has seen explosive growth in recent years, largely due to the increasing popularity of personal assistant features on mobile devices and smart home speakers. In this work, we present Personal VAD 2.0, a personalized voice activity detector that detects the voice activity of a target speaker, as part of a streaming on-device ASR system. Although previous proof-of-concept studies have validated the effectiveness of Personal VAD, there are still several critical challenges to address before this model can be used in production: first, the quality must be satisfactory in both enrollment and enrollment-less scenarios; second, it should operate in a streaming fashion; and finally, the model size should be small enough to fit a limited latency and CPU/Memory budget. To meet the multi-faceted requirements, we propose a series of novel designs: 1) advanced speaker embedding modulation methods; 2) a new training paradigm to generalize to enrollment-less conditions; 3) architecture and runtime optimizations for latency and resource restrictions. Extensive experiments on a realistic speech recognition system demonstrated the state-of-the-art performance of our proposed method.

Index Terms: personal VAD, voice activity detection, speech recognition, on-device

1. Introduction

Personal Voice Activity Detection (VAD) [1] aims to detect the voice activity of a target speaker (or multiple target speakers) at the frame level. By rejecting frames which do not contain target user speech, Personal VAD is able to significantly reduce computational resources and also improves speech recognition. As a result, Personal VAD is an indispensable module for speech processing systems within personalized devices and services, such as smartphones, smart home speakers, and tablets.

Several studies [1, 2, 3, 4, 5] have demonstrated the viability of Personal VAD. Ding *et al.*[1] first proposed the idea of a Personal VAD through a proof-of-concept study on the frame-wise VAD accuracy. Following this, Medennikov *et al.* [2] and He *et al.* [3] investigated the use of Personal VAD as a speaker diarization technique in the "cocktail-party problem" setting [6] with one or more target speakers. Makishima *et al.* [4] proposed an enrollment-less training scheme for Personal VAD to avoid the need for an enrollment utterance during training. In addition, the authors of [5] extended Personal VAD for the purposes of start/end of speech detection, being more flexible in an ASR system.

However, there are still several critical challenges blocking the use of Personal VAD in production environments, especially on-device speech recognition systems [7, 8, 9, 10, 11]. Once deployed, Personal VAD will be an "always running" system that replaces standard VAD. Therefore, it should achieve satisfactory performance in both enrollment and enrollment-less scenarios. With enrollment, it should minimize the insertion error caused by non-target speech while avoiding deleting target speech. Without enrollment, it should perform at least as well as a standard VAD. From the perspective of deployment, Personal VAD should operate in a streaming fashion and have a small model size to have minimal latency and CPU/memory impacts.

The conventional Personal VAD [1, 2, 3, 4, 5] currently does not meet these requirements. First, the speaker embedding modulation approaches are usually based on concatenating the acoustic features with the speaker embedding, which we have found to result in sub-optimal VAD and ASR performance (Section 4.1). Second, conventional Personal VAD models assume an application scenario with at least one enrolled target speaker, which does not work when no enrolled speaker is present. Lastly, previous works are mostly proof-of-concept studies that do not consider any runtime optimizations, which is unfeasible for production environments.

To address the problems, we propose Personal VAD 2.0, an optimized Personal VAD model for on-device speech recognition. Our main contributions are outlined below:

- We propose two advanced methods for speaker embedding modulation, through a feature-wise transform (FiLM) layer [12] or through a speaker pre-net, instead of naively concatenating them to the input acoustic features.
- We propose a novel model training paradigm to extend personal VAD to behave as a standard VAD under enrollment-less conditions. During training, we randomly replace the target speaker embeddings with a zero vector and modify the labels of non-target speaker speech to target speaker speech.
- We also investigate model architecture and runtime optimizations, including using a streaming-friendly and more powerful Conformer backbone [13], and quantizing the model to 8-bit integer format to reduce model size to meet the strict on-device production requirements.
- We for the first time conduct experiments on an on-device ASR system with realistic speech traffic to evaluate the effectiveness of personal VAD models. Results show that our proposed model can significantly reduce ASR insertion errors in the enrollment scenario, while retaining the same performance as standard VAD in the enrollment-less scenario.

2. Methods

In this section, we start with a quick recap of the conventional Personal VAD model. Following this, we introduce our proposed methods of improving speaker embedding modulation,

Copyright © 2022 ISCA

10.21437/Interspeech.2022-856

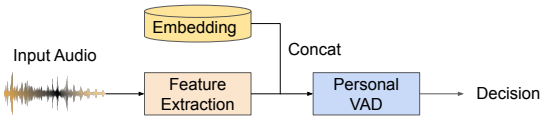


Figure 1: *Conventional Personal VAD model. The model takes acoustic features and speaker embedding as the inputs and produces a frame-wise prediction.*

generalizing to both enrollment and enrollment-less conditions, and architecture as well as runtime optimizations.

2.1. Recap: Personal VAD model

In conventional Personal VAD systems, users need to go through an enrollment process, where a pre-trained text-independent speaker recognition model [14] computes the target user's d-vector speaker embeddings from the target user's recordings, to encode their voice characteristics.

The diagram of a conventional Personal VAD is shown in Figure 1. The model first extracts acoustic features $\mathbf{x} \in \mathbb{R}^{T \times D}$ from input audio, where T and D denote the sequence length and feature dimension, respectively, and $\mathbf{x}_t$ represents the t-th frame of acoustic features. Each $\mathbf{x}_t$ is then concatenated with the speaker embedding from the target speaker $\mathbf{e}^{\text{target}}$:

$$\hat{\mathbf{x}}_t = [\mathbf{x}_t, \mathbf{e}^{\text{target}}]. \tag{1}$$

The model consumes the concatenated features as inputs and produces frame-wise decision among target speaker's speech (`tss`), non-target speaker's speech (`ntss`), and non-speech (`ns`):

$$\mathbf{p}_t = \text{PVAD}(\hat{\mathbf{x}}_t), \tag{2}$$

where $\mathbf{p}_t = [p_t^{\texttt{tss}}, p_t^{\texttt{ntss}}, p_t^{\texttt{ns}}]$ correspond to the posteriors of the three classes. Typically, previous studies implement the model PVAD with a couple of Bidirectional LSTM (BLSTM) layers (*e.g.*, three-layer BLSTM in [1, 2, 4]).

2.2. Improving speaker embedding modulation

As described in Section 2.1, vanilla Personal VAD models concatenate target speaker embedding with the acoustic feature to incorporate speaker specific information. However, acoustic features and speaker embeddings represent very different information, and they are extracted through entirely separated processes, leading to different distributions and magnitudes. Therefore, simply concatenating them and feeding them to the same layers may significantly limit the model capacity. To address this issue, we propose two novel speaker embedding modulation approaches: 1) through a FiLM layer, and 2) through a speaker pre-net.

2.2.1. Speaker embedding modulation through FiLM

A FiLM layer [15] applies a feature-wise affine transformation to its inputs, including scaling and shifting operations. This affine transformation generalizes concatenation-, biasing-, and scaling-based conditioning operators, which has been shown to be more expressive in learning conditional representations than using either of them individually [16, 17].

Figure 2(a) shows the diagram of Personal VAD model with FiLM layer. Suppose $\mathbf{h} = \text{Conformer}(\mathbf{x})$ is the input to FiLM layer. The FiLM generator first takes the target speaker embedding $\mathbf{e}^{\text{target}}$ as inputs, and produces a scaling vector $\gamma(\mathbf{e}^{\text{target}})$ and a shifting vector $\beta(\mathbf{e}^{\text{target}})$ (*i.e.*, FiLM parameters), with the same dimensions as $\mathbf{h}$. Following this, the input $\mathbf{h}$ is scaled and shifted by the corresponding vectors:

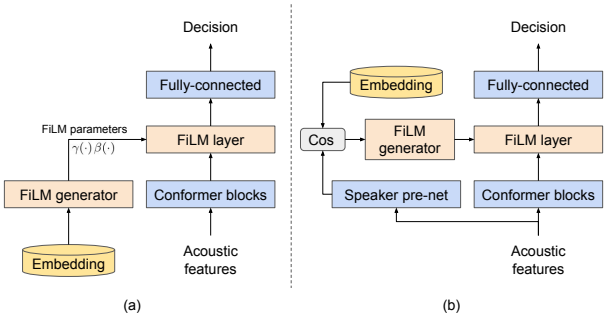


Figure 2: *(a) Personal VAD model with FiLM layer. $\gamma(\cdot)$ and $\beta(\cdot)$ are the scaling and shifting vectors of FiLM respectively. (b) Personal VAD model with speaker pre-net.*

$$\text{FiLM}(\mathbf{h}) = \gamma(\mathbf{e}^{\text{target}}) \cdot \mathbf{h} + \beta(\mathbf{e}^{\text{target}}) \tag{3}$$

The output of the FiLM layer is passed to the final fully-connected classifier.

2.2.2. Speaker embedding modulation through speaker pre-net

Alternatively, instead of directly conditioning speaker embeddings to the network, we propose adding a speaker pre-net to extract the speaker information from the acoustic features and produce a learned embedding that has the same dimensions as the speaker embedding. Following this, we compute the cosine similarity between the learned embedding and the target speaker embedding, which is then used for conditioning. Compared to direct speaker embeddings conditioning, this will provide more discriminative information to the classifier, which assists the model to better distinguish different speaker and provide more accurate decisions.

Figure 2(b) shows the diagram of Personal VAD with speaker pre-net. Formally, using our prior notations, the speaker pre-net consumes acoustic features $\mathbf{x}$ and produces a fixed-length embedding for each frame $\mathbf{e}^{\text{prenet}} \in \mathbb{R}^{T \times D_e}$:

$$\mathbf{e}^{\text{prenet}} = \text{PreNet}(\mathbf{x}) \tag{4}$$

where D_e is the dimension of the pre-net embedding (equal to that of speaker embedding). Then, we compute the cosine similarity score $\mathbf{s} \in \mathbb{R}^T$ between the two embeddings:

$$\mathbf{s} = \cos(\mathbf{e}^{\text{prenet}}, \mathbf{e}^{\text{target}}) \tag{5}$$

Lastly, we modulate the output of the conformer blocks with the cosine similarity score via FiLM:

$$\text{FiLM}(\mathbf{h}) = \gamma(\mathbf{s}) \cdot \mathbf{x} + \beta(\mathbf{s}) \tag{6}$$

As a straightforward extension, we also investigated modulating both speaker embedding and cosine similarity with FiLM (concatenating the two and then feed to FiLM layer) to consider both generative and discrimination speaker-related clues, which we will examine in the experiments.

2.3. Training the model for both enrollment and enrollment-less conditions

In practical on-device ASR use cases, the user may choose to either proceed with or skip the enrollment process. As an "always-running system", the Personal VAD model should be able to achieve satisfactory performance in both conditions. To achieve this, we propose a simple but effective training paradigm to guarantee satisfactory performance for both enrollment and enrollment-less conditions, as shown in Alg. 1. This guarantees the model learning the standard VAD behaviors while there is no speaker embedding. In practice, we set

$p_0 = 0.2$, which we found to provide reasonable performance for both conditions. Additionally, we add datasets that have no enrollment utterance to the model training, paired with a zero speaker embedding for each utterance. During inference, we also feed an zero speaker embedding to the model under the enrollment-less condition.

Algorithm 1 Training the model for both enrollment and enrollment-less conditions.

1: During each training epoch, sample a subset of training utterances with probability p_0
2: **for** each sampled training utterance **do**
3: Set the speaker embedding to zero vector: $\mathbf{e}^{\text{target}} = \overrightarrow{\mathbf{0}}$
4: Replace ground-truth labels of `ntss` to `tss`
5: **end for**

2.4. Conformer backbone: Streaming support & accuracy improvement

Transformer models [18] have been shown to substantially improve almost all speech modeling tasks over LSTMs in recent years [19, 13, 20]. Consequently, we investigate the use of Conformer backbone, an advanced Transformer architecture, for Personal VAD. The vanilla transformer uses the full sequence as the attention context, which leads to a very high latency in streaming systems that is similar to BLSTMs. To make the model stream-able and maximally reduce the latency, we set the model to have a limited left-context and no right-context, following [21, 22].

2.5. Model Quantization

Due to the limited resources on mobile devices, we adopt a runtime optimization for the on-device inference of Personal VAD. Instead of directly running TensorFlow graph with 32-bit float weights, we quantize the model weights to have 8-bit integers using dynamic range quantization [23, 24] and serialize the models to the TensorFlow Lite format. The quantized model has only 1/4 model size, and more importantly, achieves significant speedup with optimized hardware instructions of integer arithmetic, improving both memory usage and latency for Personal VAD.

3. Experimental setup

3.1. Datasets

We conduct our experiments using a vendor-collected dataset of realistic speech queries. The training set contains 2.6 million utterances (∼1,600 hours) from 6,923 speakers. As suggested by [1], there is no existing datasets that is ideal to Personal VAD evaluations, where each utterance contains natural speaker turns and it contains enrollment utterances for each individual speaker. To simulate the conversational speech, we create a set of synthetic data based on the vendor-collected dataset, following [1]. We first concatenate utterances from multiple speakers, and then we randomly select one of the speakers as the target speaker in the concatenated utterance. Accordingly, we extract the target speaker embedding from the utterance in an enrollment subset. In addition, we also apply two data augmentation techniques, MTR [25] and SpecAug [26], on our datasets to avoid domain overfitting and mitigate concatenation artifacts. Most of the models in the experiments are trained with concatenated dataset alone. Only when training the model for both enrollment and enrollment-less conditions, we used an additional internal training set [7] consisting of 35 million English *audio-text pairs* (∼27,500 hours) from multiple domains including YouTube and anonymized voice search traffic. The internal training set is only used for enrollment-less condition (where we do not perform any speaker-related operations), and our data handling abides by *Google AI Principles* [27]. We use the same datasets in our baseline standard VAD training for a fair comparison.

Table 1: *WERs (%) of the ablation study on the proposed components under the enrollment condition. Non-Concat test set shows the performance when only the target speaker is talking, while the concatenated test set showing the performance in a conversational speech. Model size is represented in MB. Runtime complexity is measure by FLOPs.*

Exp	Model	Speech Query test set		Size	FLOPs
		Non-Concat	Concat	(MB)	(M)
B0	Personal VAD [1]	17.9	41.0	5.8	3.54
E0	+ Conformer	15.3	31.5	2.8	9.51
E1	+ FiLM d-vector	11.3	29.5	2.8	9.58
E2	+ Speaker PreNet (cos)	11.7	27.5	4.0	9.51
E3	+ FiLM d-vector & cos	11.5	27.6	4.0	9.58
E4	+ 8-bit quantization	12.2	27.2	1.0	9.58

The test set of the vendor-collected dataset of realistic speech queries comprises 194,890 utterances from 1,241 speakers. These utterances are very similar to actual speech query traffic and can contain various ambient noise. As a result, we do not add extra reverberation or noise to it during evaluations. The non-concatenated test set shows the performance when only the target speaker is talking. Similarly, we create a concatenated test set to evaluate the performance of Personal VAD in conversational speech. To reduce the evaluation cost, we randomly samples a subset of 4,000 utterances from the original test set and the concatenated test set, respectively. In the evaluations under enrollment-less condition, we also use 14,000 Voice-search utterances (VS).

3.2. Implementation details

We use the same frontend feature as [7], a 128-dimensions log-Mel filterbank energies extracted with 32ms window and 10ms shift. We stack the features from 4 contiguous frames (512-dimensional) and subsample the sequence by a factor of 3. The LSTM-based models have 3 uni-directional LSTM layers with 256 units and a final fully-connected layer as the classifier. Our Conformer-based models have 4 Conformer layers in the block, each having a dimension of 64, attention head of 8, causal 7×7 convolution kernel, and 31 left-context. The speaker pre-net has two Conformer layers with the same hyper-parameters.

During inference, we compare the posterior of target speaker speech $p_t^{\texttt{tss}}$ to a pre-defined threshold (0.1 in all our experiments) to decide if a speech frame belongs to the target speaker. We discard non-target speech and non-speech frames and only pass target speech frames to downstream ASR model. We use a medium size (∼ 40M parameters) LSTM-based ASR model [8] among all our experiments, which is trained on the realistic voice search traffic. We use word error rate (WER) to measure the ASR performance with Personal VAD models.

4. Results

We conducted a series of ablation studies and comparisons to evaluate the performance of the proposed Personal VAD 2.0. Starting from the conventional Personal VAD model in [1], we incrementally investigate the performance gain (adding one technique at a time) obtained from Conformer backbone, novel speaker embedding modulation methods, model quantization, and the model training paradigm for enrollment and enrollment-less conditions. Finally, we made an overall comparison of the best performing Personal VAD 2.0 model to the baselines to emphasize the effectiveness of our proposed approach.

Table 2: *Overall comparisons to baseline standard VAD and Personal VAD, and the ablation study results of the proposed training paradigm for both enrollment and enrollment-less conditions. Under enrollment-less condition, the WERs on Concat testset are expected to be over 100%, so we do not show it here for simplicity.*

Exp	Method	Enrollment		Enrollment-less		Size (MB)	FLOPs (M)
		Non-Concat	Concat	VS	Non-Concat		
B1	LSTM Standard VAD	N/A	N/A	7.2	11.5	1.5	3.41
B2	Conformer Standard VAD	N/A	N/A	6.9	10.1	0.7	8.77
B0	Personal VAD [1]	17.9	41.0	$\geq$100.0	$\geq$100.0	5.8	3.54
E5	Personal VAD 2.0	12.4	32.7	7.0	10.1	1.0	9.58
E4	w/o our training paradigm	12.2	27.2	$\geq$100.0	$\geq$100.0	1.0	9.58
E5	w/ our training paradigm	12.4	32.7	7.0	10.1	1.0	9.58

4.1. Ablation studies

Conformer backbone. We first evaluate the efficacy of the Conformer backbone. As shown in Table 1, using Conformer backbone (*E0*) significantly improves both Non-Concat and Concat test sets by 2.6 and 9.5 absolute WERs, respectively. More importantly, the model size of the Conformer model is less than a half of the LSTM model size (*E0*: 2.8MB vs. *B0*: 5.4MB), suggesting that Conformer-based Personal VAD model can achieve much high parameter efficiency.

Proposed speaker embedding modulation methods. Next, we evaluate our proposed speaker embedding modulation methods: (*E1*) FiLM, (*E2*) Speaker pre-net (*i.e.*, FiLM cos similarity score), and (*E3*) the combination of both. Table 1 shows the performance of the three approaches. Compared to concatenating at the input (*E0*), all three approaches achieve essential performance gains on both Non-Concat ($\geq$3.6) and Concat ($\geq$2.0) test sets, indicating that the proposed modulation approaches have better capacity for speaker-related information. When comparing between the three approaches, we found that they perform similarly on Non-Concat, but *E2* and *E3* achieve further improvement on Concat, indicating the importance of modulating discriminative information to the models.

Model quantization. We also evaluate the influence of model quantization in final ASR WERs. We added 8-bit model quantization on the top of *E3*, as shown in system *E4* in Table 1. From the results, we found that the quantization hurts the WER on Non-Concat test set, but the degradation is marginal (0.7). Interestingly, the Concat WER is improved by 0.2, possibly due to the quantization avoid the model overfitting to some local noises. With the 75% model size reduce and the marginal performance change, model quantization proves its effectiveness and necessity in on-device deployment.

Training the model for both enrollment and enrollment-less conditions. Lastly, we explore the model training paradigm that is proposed to guarantee Personal VAD to have reasonable performance under enrollment-less condition. Results are shown in Table 2. Under enrollment-less condition, the model with it (*E5*) can achieve 7.0 and 10.1 on VS and Non-Concat test sets, respectively, which suggests that adopting the paradigm can successfully generalize Personal VAD to enrollment-less condition. Surprisingly, compared to *E4*, *E5* has a 4.5 WER increase under enrollment condition. To analyze this observation, we break the WER down to deletion, insertion, and substitution errors (*E4*: 3.7/14.8/8.9 vs. *E5*: 4.7/18.9/9.1). We noticed that the increased WER is mainly due to an increased insertion error rate. A possible explanation is that, with the joint training paradigm, the model is biased to provide a higher posterior for target speaker speech, as we use this dimension for all the speech frames when dealing with enrollment-less data.

4.2. Overall comparisons to baselines and discussions

After verifying the impact of each aspect of our proposed approach, we deliver the comparison results between the best performing Personal VAD 2.0 model and standard/Personal VAD baselines in Table 2. For a fair model size comparison, we also conduct 8-bit quantization to the standard VAD baseline, as they have been widely used in on-device products. Compared with the two standard VAD baselines, Personal VAD 2.0 can significantly reduce the WER (mostly insertion errors) on conversational speech scenarios (Concat test set), while retaining the performance on regular speech queries under either condition. Personal VAD 2.0 has higher FLOPs, due to the use of Conformer backbone. However, it does not have significant impacts to the overall ASR system latency since VAD has much less computations than ASR. More importantly, conformer backbone makes it possible to process several frames in a the same batch, which further speed up model inference and reduces the latency.

Compared with the conventional Personal VAD, Personal VAD 2.0 not only tremendously improves the performance under the enrollment condition but also generalizes well to enrollment-less conditions. More importantly, with the optimizations based on Conformer backbone and 8-bit weight quantization, we developed a light-weight Personal VAD model that can easily fit into any on-device systems. Meanwhile, we still notice an issue of the current Personal VAD 2.0 model – it performs slightly better on regular speech queries (Non-Concat) under enrollment-less condition than under enrollment condition. A potential reason is that when the target speaker and non-target speaker have very similar voice, the model is more likely to make erroneous predictions. The gap can be reduced by tuning the VAD decision threshold, at the cost of an increasing insertion error rate on conversational speech, resulting in a trade-off to be made according to the deployment scenarios (*e.g.*, for keyword spotting, we need to minimize deletion error, and it is less sensitive to insertion errors).

5. Conclusions

In this study, we have proposed Personal VAD 2.0, an optimized Personal VAD model for on-device speech recognition. Through a series of ablation studies, we evaluate the impact of our novel design choices and runtime optimizations. Specifically, in terms of speaker embedding modulation, we showed that using FiLM and speaker pre-net can significantly improve the model performance than simply concatenation at the input. We also confirmed that our proposed training paradigm can effectively generalize Personal VAD to enrollment-less inference setting, while retaining the performance under enrollment condition. Additionally, we verified that adopting Conformer backbone and adding 8-bit quantization to Personal VAD model can tremendously increase the parameter efficiency. Compared to baseline standard VADs and conventional Personal VAD, Personal VAD 2.0 achieves the state-of-the-art performance in an on-device speech recognition task. Future work will focus on mitigating the performance difference between enrollment and enrollment-less conditions, and investigating multi-user Personal VAD in on-device ASR scenarios.

6. References

[1] S. Ding, Q. Wang, S.-Y. Chang, L. Wan, and I.-L. Moreno, "Personal VAD: Speaker-conditioned voice activity detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 433–439.

[2] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny *et al.*, "Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario," in *Proc. Interspeech*, 2020.

[3] M. He, D. Raj, Z. Huang, J. Du, Z. Chen, and S. Watanabe, "Target-speaker voice activity detection with improved i-vector estimation for unknown number of speaker," *arXiv preprint arXiv:2108.03342*, 2021.

[4] N. Makishima, M. Ihori, T. Tanaka, A. Takashima, S. Orihashi, and R. Masumura, "Enrollment-less training for personalized voice activity detection," *arXiv preprint arXiv:2106.12132*, 2021.

[5] A. Jayasimha and P. Paramasivam, "Personalizing speech start point and end point detection in ASR systems from speaker embeddings," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 771–777.

[6] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[7] T. N. Sainath, Y. He, B. Li, A. Narayanan, R. Pang, A. Bruguier, S.-y. Chang, W. Li, R. Alvarez, Z. Chen *et al.*, "A streaming on-device end-to-end model surpassing server-side conventional model quality and latency," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6059–6063.

[8] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.

[9] J. Macoskey, G. P. Strimel, J. Su, and A. Rastrow, "Amortized neural networks for low-latency speech recognition," *arXiv preprint arXiv:2108.01553*, 2021.

[10] V. Joshi, R. Zhao, R. R. Mehta, K. Kumar, and J. Li, "Transfer learning approaches for streaming end-to-end speech recognition system," *Proc. Interspeech*, pp. 2152–2156, 2020.

[11] C. Wang, Y. Wu, L. Lu, S. Liu, J. Li, G. Ye, and M. Zhou, "Low latency end-to-end streaming speech recognition with a scout network," *Proc. Interspeech*, pp. 2112–2116, 2020.

[12] V. Dumoulin, E. Perez, N. Schucher, F. Strub, H. d. Vries, A. Courville, and Y. Bengio, "Feature-wise transformations," *Distill*, vol. 3, no. 7, p. e11, 2018.

[13] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *Proc. Interspeech*, pp. 5036–5040, 2020.

[14] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[15] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[16] E. Perez, H. De Vries, F. Strub, V. Dumoulin, and A. Courville, "Learning visual reasoning without strong priors," *arXiv preprint arXiv:1707.03017*, 2017.

[17] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[20] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[21] N. Moritz, T. Hori, and J. Le, "Streaming automatic speech recognition with the transformer model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6074–6078.

[22] B. Li, A. Gulati, J. Yu, T. N. Sainath, C.-C. Chiu, A. Narayanan, S.-Y. Chang, R. Pang, Y. He, J. Qin *et al.*, "A better and faster end-to-end model for streaming ASR," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5634–5638.

[23] R. Alvarez, R. Prabhavalkar, and A. Bakhtin, "On the efficient representation and execution of deep acoustic models," *arXiv preprint arXiv:1607.04683*, 2016.

[24] Y. Shangguan, J. Li, Q. Liang, R. Alvarez, and I. McGraw, "Optimizing speech recognition for the edge," *arXiv preprint arXiv:1909.12408*, 2019.

[25] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in Google Home," in *Interspeech*, 2017, pp. 379–383.

[26] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech*, pp. 2613–2617, 2019.

[27] Google, "Artificial Intelligence at Google: Our Principles." [Online]. Available: "https://ai.google/principles/"