ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095360

ERSAM: NEURAL ARCHITECTURE SEARCH FOR ENERGY-EFFICIENT AND REAL-TIME SOCIAL AMBIANCE MEASUREMENT

Chaojian Li[*,1], *Wenwan Chen*[*,2], *Jiayi Yuan*[*,2], *Yingyan (Celine) Lin*[1], *and Ashutosh Sabharwal*[2]

[1]Georgia Institute of Technology, Atlanta, GA
[2]Rice University, Houston, TX

ABSTRACT

Social ambiance describes the context in which social interactions happen, and can be measured using speech audio by counting the number of concurrent speakers. This measurement has enabled various mental health tracking and human-centric IoT applications. While on-device Socal Ambiance Measure (SAM) is highly desirable to ensure user privacy and thus facilitate wide adoption of the aforementioned applications, the required computational complexity of state-of-the-art deep neural networks (DNNs) powered SAM solutions stands at odds with the often constrained resources on mobile devices. Furthermore, only limited labeled data is available or practical when it comes to SAM under clinical settings due to various privacy constraints and the required human effort, further challenging the achievable accuracy of on-device SAM solutions. To this end, we propose a dedicated neural architecture search framework for **E**nergy-efficient and **R**eal-time **SAM** (ERSAM). Specifically, our ERSAM framework can automatically search for DNNs that push forward the achievable accuracy vs. hardware efficiency frontier of mobile SAM solutions. For example, ERSAM-delivered DNNs only consume 40 mW · 12 h energy and 0.05 seconds processing latency for a 5 seconds audio segment on a Pixel 3 phone, while only achieving an error rate of 14.3% on a social ambiance dataset generated by LibriSpeech. We can expect that our ERSAM framework can pave the way for ubiquitous on-device SAM solutions which are in growing demand.

Index Terms— social ambiance, neural arch search

1. INTRODUCTION

Social ambiance, which describes the context of an environment where social interactions are happening, can be measured via speech audio [1, 2]. Among the widely used social ambiance measure (SAM) solutions, counting the number of concurrent speakers around an individual has been verified to be associated with their mental health symptoms (e.g., depression and psychotic disorders) [1]. While deep neural networks (DNNs) have enabled accurate SAM in mental health tracking by counting concurrent speakers [1], it is still challenging to achieve continuous, real-time, and on-device DNN-based SAM, which is highly desirable for preserving user privacy and thus ensuring wide adoption of the aforementioned mental health tracking. First, powerful DNNs are often complex while mobile/wearable devices are very limited in both computational and memory resources, e.g., deploying the wav2vec2 DNN [3] on a Pixel 3 phone [4] for speech recognition requires 2200 mW [5] power consumption vs. only 700 mW [6] allowed by the Pixel 3 phone [4]. Second, training DNNs usually requires a large amount of labeled data to achieve a satisfactory task accuracy, whereas SAM users expect DNNs to be trained locally on a device with limited data (e.g., less than 8 hours of training data [7]), due to both the privacy concern and prohibitive human effort needed to collect clinical SAM audio data.

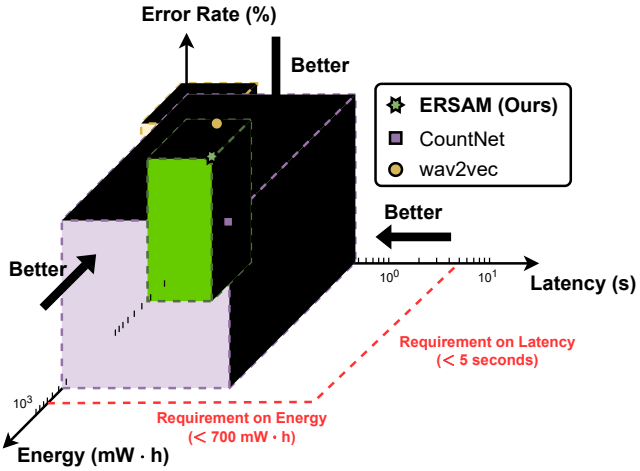

Fig. 1: DNNs searched by our ERSAM outperform state-of-the-art DNN-based SAM: achieving the best accuracy vs. hardware efficiency trade-offs while meeting the requirements of being energy-efficient and achieving real-time latency. Here a smaller volume of each data point's 3D cuboid corresponds to a better accuracy vs. hardware efficiency trade-off.

To close the aforementioned gap between the required computational complexity for DNN-based SAM and the limited resources on mobile/wearable devices, it is critical to develop compact DNNs that ensure (i) real-time latency (e.g., $\leq$ **5 seconds of processing time for an audio segment of 5 seconds** [1]) for timely usage of the generated SAM information (e.g., real-time intervention), (ii) constrained energy ($\leq$ **700mW · 12h [6]**), and (iii) an acceptable SAM accuracy under the aforementioned low-resource settings. Thanks to the recent success of both (i) **H**ard**w**are-aware **N**eural **A**rchitecture **S**earch (HW-NAS) [8] in developing DNN models with optimal accuracy vs. hardware efficiency trade-offs [8, 9, 10] and (ii) knowledge distillation in leveraging pre-trained giant models to boost the achievable accuracy of compact DNNs [11, 12], it is natural to consider them for continuous, real-time, and on-device DNN-based SAM solutions. However, existing HW-NAS and knowledge distillation techniques are not directly applicable for meeting the above

*Equal contribution.

We would like to acknowledge the funding support from the NSF SCH, Expedition, and RTML programs (Award ID: 1838873, 1730574, and 1937592) for this project.

requirements for the following reasons: (i) prior HW-NAS works mainly focus on computer vision or natural language processing tasks [8, 9], where the search space of existing HW-NAS works might not be optimal or even feasible for DNNs dedicated to SAM due to their differences from SAM in terms of 1) input modalities (e.g., images vs. audio), 2) commonly used operators (e.g., 2D convolution vs. 1D convolution), and 3) dataset size (e.g., 150 GB ImageNet [13] with $>$ 1 million images vs. 1 GB dataset with only 8 hours of audio); and (ii) the large cost of querying the giant models required by vanilla knowledge distillation techniques makes knowledge distillation impractical to be used on a local edge/mobile device.

Our key contribution is a new framework, called neural architecture search for Energy-efficient and Real-time SAM (ERSAM), which achieves continuous, real-time, and on-device DNN-based SAM, as elaborated below.

- We develop and validate a dedicated HW-NAS and knowledge distillation framework called ERSAM to develop DNN-based SAM solutions towards meeting SAM's real-world application driven requirements.
- ERSAM Enabler 1: The proposed ERSAM integrates a hardware-aware search space dedicated to SAM by leveraging the cost profiling observations from state-of-the-art DNN-based SAM models on mobile phones.
- ERSAM Enabler 2: We propose an efficient knowledge distillation scheme to be embedded into both the search and training processes of our ERSAM framework's HW-NAS engine, that enables the feasibility of developing compact yet effective DNNs for SAM on the local edge by only making necessary queries to the giant teacher model.
- ERSAM performance: As shown in Fig. 1, our ERSAM framework's delivered DNNs fulfill all the above requirements, achieving (i) **0.05 seconds ($\leq$ 5 seconds [1])** latency for an audio segment of 5 seconds, (ii) **40 mW $\cdot$ 12 h ($\leq$ 700 mW $\cdot$ 12 h [6])** energy consumption on a Pixel 3 phone, and (iii) 14.3% error rate ($\downarrow$ 3.6% than the state-of-the-art work of counting the number of speakers [14]) under the low-resource setting of **8 hours ($\leq$ 8 hours [7])** training data.

2. RELATED WORKS

2.1. SAM based on counting concurrent speakers

Speech audio can provide important information about social ambiance, and can be applied to finding social hot spots [15], event contexts [2][16], and children's language environments [17, 18]. Specifically, [2] measures social ambiance by inferring the occupancy and human chatter levels in local business scenarios; [16] characterizes social ambiance from overlapping conversational data in a crowded environment. More recently, the number of concurrent speakers around an individual has been verified to be associated with mental health symptoms [1]. The number of simultaneous speakers is leveraged as a proxy for the overall social activity in the associated environment based on the assumption that concurrent speakers provide fine-grained information of social ambiance [1].

Meanwhile, several recent studies [14, 19] have shown that DNNs can achieve state-of-the-art accuracy in predicting the number of concurrent speakers as compared to non-DNN solutions. However, they either (i) can only estimate up to five speakers, which cannot cover all social scenarios in daily life, or (ii) require more than 30 hours of training data, which can be challenging to acquire and label for SAM under clinical scenarios. In contrast, our proposed ERSAM can automatically deliver DNNs that perform energy-efficient and real-time SAM, based on a small-scale training dataset which favors preserving privacy and timely feedback to fulfill the requirements of SAM under clinical scenarios.

2.2. Hardware-aware neural architecture search

HW-NAS has been proposed with the aim of automating the search for efficient DNN structures under target hardware efficiency constraints (e.g., energy or latency on target devices) [20]. Besides the early works based on reinforcement learning [21, 22], [8, 23] develop differentiable HW-NAS following [24] to significantly improve search efficiency. More recently, [9, 10] propose to jointly train all sub-networks within the search space in a weight-sharing supernet and then locate the optimal architectures under different cost constraints without re-training or fine-tuning, thus reducing the cost of the whole search and training pipeline as compared to previous HW-NAS solutions. However, all these HW-NAS works are dedicated to computer vision or natural language processing applications and cannot be directly applied to SAM, due to the differences in input modalities, commonly used operators and dataset size. Note that although there exist prior works that apply NAS to speech recognition [25, 26, 27], none of them target developing continuous, real-time, and on-device DNN-based SAM that meets all the requirements mentioned in Sec. 1.

2.3. Efficient knowledge distillation

Although knowledge distillation [11] is commonly used to boost the achievable accuracy vs. inference efficiency trade-offs of compact DNNs, it requires more training time due to the extra overhead of querying larger teacher models [12, 28]. While there exist prior works that focus on efficient knowledge distillation, they either (i) only focus on the data efficiency [29, 30] instead of the efficiency of the training latency or energy, or (ii) are only designed for or verified on computer vision tasks with specially designed image operators [31]. Specifically, [30] leverages the insight that teacher models are not always perfect and bypasses the wrong predictions of teacher models to improve the accuracy of student models. Different from all the prior works mentioned above, our proposed efficient knowledge distillation scheme (i) presents a different insight, which is the improved accuracy of the larger teacher model over the small student model is caused by the case when the smaller models are wrong and uncertain while the larger models are correct and certain and (ii) targets compressing training/search cost.

3. THE PROPOSED ERSAM FRAMEWORK

Overview. As shown in Fig. 2, our ERSAM accepts training data and corresponding labels of limited size, hardware-aware network operators, and a costly teacher model as inputs, and then automatically generates a dedicated DNN to enable continuous, real-time, on-device DNN based SAM. ERSAM integrates two enablers: (i) a hardware-aware search space dedicated to SAM based on the cost profiling observations of state-of-the-art DNN based SAM models on mobile phones, and (ii) an efficient knowledge distillation scheme to be embedded in the search and training process by only making queries when the student model is uncertain under the given input data. Specifically, during each iteration, we sample one sub-network

Table 1: Macro-architecture of our designed hardware-aware hardware search space for SAM.

Operator Type	#Channels	#Repeats	Kernel Size	Stride
1D Conv.	16 - 32	{1}	10	5
1D Conv.	32 - 64	{1}	8	4
1D Conv.	64 - 128	{1, 2, 3}	4	2
1D Conv.	128 - 256	{1, 2, 3}	1	1
1D Conv.	128 - 256	{1, 2, 3}	{1, 2, 3}	1
1D Conv.	128 - 256	{1, 2, 3}	{4, 5, 6}	1
1D Conv.	128 - 256	{1, 2, 3}	{7, 8, 9}	1
1D Conv.	128 - 256	{1, 2, 3}	{10, 11, 12}	1

from the search space, and then pass the training data to the sampled sub-network. If the measured uncertainty based on the sampled sub-network's predictions is higher than a pre-defined threshold, the query to the teacher model will be enabled, and the corresponding back-propagation will be performed to incorporate an extra distillation loss [11]. Otherwise, the training will be the same as the standard training process without knowledge distillation. Note that ERSAM is built on top of a weight-sharing NAS [9, 10], which trains all the sub-networks in a weight-sharing supernet and then locates the optimal one under different cost constraints without re-training; thus, the search cost is shared by training the supernet.

3.1. Enabler 1: Hardware-aware search space design

As pointed out in prior works, even for DNNs that have the same width, depth, or input resolution, the real hardware-cost (e.g., latency and/or energy) can be quite different on a target device [33, 34]. Thus, a hardware-aware search space is critical for achieving our target continuous, real-time, and on-device DNN-based SAM. Specifically, we first analyze the latency of each operator in wav2vec [32], a state-of-the-art speech recognition model. As shown in Fig. 3, the bottleneck operator, marked with a red circle in Fig. 3, has a cost that is $> 35\%$ of the whole model's latency on a Pixel 3 phone. We find that the above cost-dominant operator features both a large kernel size (i.e., 8) and input sequence length (i.e., $0.2\times$ of the original input sequence length), while maintaining the same channel size as other operators (i.e., 512). As such, we design a hardware-aware search space for SAM that is abstracted from the model architecture of wav2vec [32], while avoiding cases where a large channel size, kernel size, and input sequence length simultaneously exist within the same operator. The developed search space in our ERSAM framework, inspired by our cost profiling observations on mobile phones, is summarized in Tab. 1.

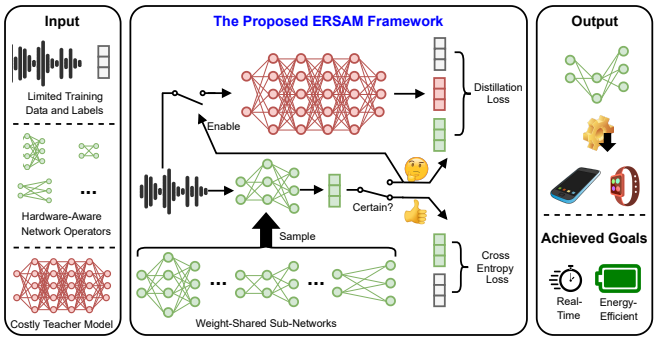

Fig. 2: Overview of our proposed ERSAM framework.

3.2. Enabler 2: Efficient knowledge distillation scheme

The proposed efficient knowledge distillation scheme is based on our observation that the improved accuracy of larger models over smaller models for SAM is a result of the case where the smaller models are wrong and uncertain but the larger models are correct and certain. Specifically, we compare the correct and wrong predictions of a larger model called wav2vec [32] and a smaller counterpart, a uniformly scaled wav2vec with only 6 ($0.3\times$) layers and 128 ($0.25\times$) channels, on the LibriSpeech-SAM dataset described in Sec. 4.1.

As shown in Fig. 4, we can observe that the more uncertain the smaller model is, the more improvement can be gained by switching to the larger model from the smaller one, as evidenced by the sharp increase in the red curve corresponding to the case where only the large model is correct. Such an observation is also consistent with recent findings in natural language processing tasks [35], i.e., the larger models' advantages over smaller ones appear when the latter are not uncertain about the input data. Leveraging the above observation, in the forward process of our proposed efficient knowledge distillation scheme, we first obtain the student model's classification probability distribution, and then compare it with the prior classification probability distribution to measure the uncertainty of the current input data, which is defined as follows:

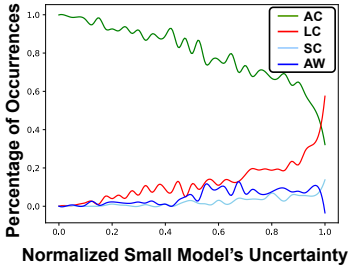

Fig. 4: The percentage of cases when varying the smaller model's uncertainties (AC: both correct, LC: only the larger model correct, SC: only the smaller model correct, and AW: both wrong).

$$S_{sample} = -\log\left(\frac{1}{n}\sum_{i=0}^{n-1}(Y_i - \hat{Y}_i)^2\right), \quad (1)$$

$$S_{batch} = \frac{1}{s}\sum_{j=0}^{s-1} S_{sample,j}, \quad (2)$$

where S_{sample} and S_{batch} denote the uncertainty score of one sample and a batch of samples, s is the batch size, Y_i is the output probability of the model for class i, $\hat{Y}_i$ is the prior probability of the labels for class i, and n is the number of classes. To simplify the measurement, all our experiments are based on the assumption that different classes have the same prior probability, i.e., $\hat{Y}_i = \hat{Y}_j$ for any i and j. As shown in Fig. 2, if S_{batch} is larger than the pre-defined threshold, then a query to the teacher model will be made; otherwise, this iteration will proceed with the standard training without knowledge distillation. The proposed efficient knowledge distillation scheme can achieve the same or even better accuracy compared with that of training with vanilla knowledge distillation [11] yet at a similar training speed as the standard training, as verified in Sec. 4.3.

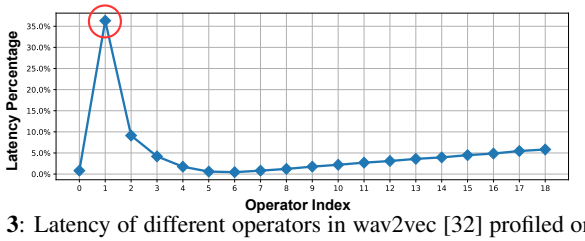

Fig. 3: Latency of different operators in wav2vec [32] profiled on a Pixel 3 [4] phone.

4. EXPERIMENTS RESULTS

4.1. Experiment settings

Dataset. To the best of our knowledge, there does not exist any realistic dataset labeled with the number of simultaneous speakers. Thus, we have synthesized a dataset with speech mixtures from LibriSpeech corpus[36]. To create a k-speaker mixture, where $k \in \{0, 1, 2, 3, 4, \geq 5\}$, we first generate a 15-30 minutes recording by concatenating all the sentences from each speaker, and then randomly select an audio segment from each recording. Finally, segments from k speakers are trimmed to 5 seconds and overlapped with each other to generate a speech mixture which is labeled as speaker count of k. Additionally, we include non-speech samples from the TUT Acoustic Scenes dataset [37] in our training data when k = 0 for scenarios where no speech is detected. Following the procedures above, 8 hours of speech mixtures are generated for training and validation. Meanwhile, the test subset with 2 hours of data is synthesized from a different set of speakers from LibriSpeech-test subset. The synthesized dataset is denoted as LibriSpeech-SAM and is used for all the experiments in this work. Specifically, to be fair with each possible speaker count, the dataset features balanced class distribution for both train and test subsets, i.e., $\sim 1,000$ samples for each class in the training set and ~ 300 samples in the testing set.

Baselines and evaluation metrics. We consider the following two models as our baselines: (i) (uniformly scaled) wav2vec [32], one of the most common models for speech recognition. Because the original wav2vec [32] exceeds the latency and energy budgets for being real-time on mobile devices, we scale it down by reducing its number of channels and number of layers uniformly to fulfill the requirements in Sec. 1; (ii) Countnet [14], the previous state-of-the-art for counting the number of speakers. For a fair comparison, we retrain it with the corresponding official implementation on our LibriSpeech-SAM dataset, because its accuracy reported in [14] is trained on a different dataset. Moreover, we leverage the following evaluation metrics in all our experiments: (i) the **error rate** on the test set of LibriSpeech-SAM; (ii) the **latency** of an audio segment of 5 seconds on a Pixel 3 Phone; (iii) the **energy consumption** on a Pixel 3 Phone when running all day (i.e., 12 h active time).

Experiment platforms. All the training and search processes are performed on a workstation with a commonly-used NVIDIA 2080 Ti GPU to match the settings of the training on the local edge (e.g., on personal desktops or laptops). The latency and energy consumption of DNN's inference are measured on a Pixel 3 Phone by running the model using the official tflite benchmark binary [38] and being monitored by the Qualcomm Snapdragon Profiler [39].

4.2. Compare with state-of-the-art

We summarize the comparison of our proposed ERSAM with the previous state-of-the-art in both Tab. 2 and Fig. 1. We can observe that ours can achieve the best error rate vs. hardware cost (e.g., latency or energy consumption) trade-offs, verifying our ERSAM's ability to deliver continuous, real-time, and on-device DNN-based SAM in an automatic manner without manually searching.

Table 2: Compare the DNN searched by our proposed ERSAM with the previous state-of-the-art.

Method	Error Rate on LibriSpeech-SAM (%)	Latency on Pixel 3 (ms)	Energy consumption on Pixel 3 (mW·h)
Countnet [14]	16.7	492	422
(Uniformly Scaled) wav2vec [32]	17.4	50	40
ERSAM	**14.3**	**47**	**38**

4.3. Ablation study on the proposed efficient knowledge distillation scheme

To better understand the efficacy and efficiency of our proposed efficient knowledge distillation scheme introduced in Sec. 3.2, we compare it with the vanilla knowledge distillation [11] and the standard training without any knowledge distillation in terms of the error rate of the finally delivered DNN and the corresponding search and training time. As summarized in Tab. 3, we can observe that (i) the vanilla knowledge distillation can achieve a lower error rate (e.g., $\downarrow 2.4\%$) but with $2.95\times$ of the search and training cost, as compared to standard training (i.e., ERSAM w/o KD); (ii) our proposed efficient knowledge distillation achieves the lowest error rate among all competitors while maintaining a similar (e.g., $1.05\times$) search and training cost with standard training. Such observations imply that our proposed efficient knowledge distillation scheme can be used as a plug-and-play module in HW-NAS, thanks to its advantage of achieving a lower error rate than vanilla knowledge distillation without requiring extra search and training costs.

Table 3: Comparison among ERSAM w/o knowledge distillation (KD), w/ vanilla KD [11], and w/ our proposed efficient KD.

Method	Error Rate on LibriSpeech-SAM (%)	Search Cost (GPU hours)
ERSAM w/o KD	16.9	2.1
ERSAM w/ vanilla KD [11]	14.5 ($\downarrow$ 2.4)	6.2 (2.95 $\times$)
ERSAM w/ our proposed KD	**14.3 ($\downarrow$ 2.6)**	**2.2 (1.05 $\times$)**

4.4. Generalize to real-world scenarios

To simulate speech audio collected from realistic scenarios, we (i) apply MUSAN noises [40] (12-18 dB) to each recording, (ii) include domain shifts caused by recording devices using LibriAdapt [41], and (iii) construct a few noisy SAM datasets following Sec. 4.1. As summarized in Tab. 4, we can observe that the proposed ERSAM can still achieve a low error rate under challenging settings and beat the uniformly scaled wav2vec [3] in terms of achieved error rate vs. hardware cost trade-offs (i.e., a $\downarrow 2.7\% \sim 5.0\%$ lower error rate under similar latency and energy consumption).

Table 4: Generalize the DNN searched by our proposed ERSAM and uniformly scaled wav2vec [3] to the dataset with noises and device diversity for simulating the scenarios of real-world application.

Error Rate (%)	(scaled) wav2vec [32]	**ERSAM**
LibriSpeech-SAM	17.4	**14.3** ($\downarrow$ 3.1)
LibriSpeech-SAM, Noisy	20.7	**17.1** ($\downarrow$ 3.6)
LibriAdapt-SAM on Pseye	30.4	**27.5** ($\downarrow$ 2.9)
LibriAdapt-SAM on Respeaker	28.6	**23.6** ($\downarrow$ 5.0)
LibriAdapt-SAM on Shure	28.7	**25.9** ($\downarrow$ 2.8)

5. CONCLUSION

In this work, we propose ERSAM, a dedicated framework for developing continuous, real-time, on-device DNN based SAM to meet all three requirements on the execution latency (e.g., $\leq$ 5 seconds), energy (e.g., $\leq$ 700mW $\cdot$ 12 h), and size of training data (e.g., audio recordings of $\leq$ 8 hours in total) towards practical adoption. This framework consists of: (i) a hardware-aware search space dedicated to SAM; and (ii) an efficient knowledge distillation scheme to be embedded in the search and training processes. The DNN model searched by our ERSAM achieves better accuracy vs. hardware efficiency trade-offs than previous state-of-the-art works and fulfills all the requirements towards continuous, real-time, on-device SAM.

6. REFERENCES

[1] Wenwan Chen et al., "Privacy-preserving social ambiance measure from free-living speech associate with chronic depressive and psychotic disorders," *Frontiers in psychiatry*, 2021.

[2] He Wang et al., "Local business ambience characterization through mobile audio sensing," in *Proceedings of the 23rd international conference on World wide web*, 2014.

[3] Alexei Baevski et al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[4] Google LLC., "Pixel 3," `https://g.co/kgs/pVRc1Y`, accessed 2020-09-01.

[5] Meta Inc., "Speech Recognition on Android with Wav2Vec2 ," `https://github.com/pytorch/android-demo-app/blob/master/SpeechRecognition`, accessed 2021-09-01.

[6] Sergey Zhidkov et al., "On smartphone power consumption in acoustic environment monitoring applications," *Applied System Innovation*, vol. 1, no. 1, 2018.

[7] Jin Xu et al., "Lrspeech: Extremely low-resource speech synthesis and recognition," in *Proceedings of the 26th ACM SIGKDD*, 2020.

[8] Bichen Wu et al., "Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search," in *Proceedings of the IEEE CVPR*, 2019.

[9] Han Cai et al., "Once-for-all: Train one network and specialize it for efficient deployment," *arXiv:1908.09791*, 2019.

[10] Jiahui Yu et al., "Bignas: Scaling up neural architecture search with big single-stage models," in *ECCV*. Springer, 2020.

[11] Geoffrey Hinton et al., "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015.

[12] Jianping Gou et al., "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, 2021.

[13] Jia Deng et al., "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on CVPR*, 2009.

[14] Fabian-Robert Stöter et al., "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM TASLP*, vol. 27, no. 2, 2018.

[15] Dave Makhervaks et al., "Combining acoustics, content and interaction features to find hot spots in meetings," in *ICASSP 2020-2020 IEEE ICASSP*. IEEE, 2020, pp. 8054–8058.

[16] Md Abdullah Al Hafiz Khan et al., "Sensepresence: Infrastructure-less occupancy detection for opportunistic sensing applications," in *2015 16th IEEE MDM*, 2015.

[17] Alejandrina Cristia et al., "A thorough evaluation of the language environment analysis (lena) system," *Behavior Research Methods*, vol. 53, no. 2, pp. 467–486, 2021.

[18] Nairán Ramírez-Esparza et al., "Look who's talking: Speech style and social context in language input to infants are linked to concurrent and future speech development," *Developmental science*, vol. 17, no. 6, pp. 880–891, 2014.

[19] Chao Peng et al., "Competing speaker count estimation on the fusion of the spectral and spatial embedding space.," in *INTERSPEECH*, 2020.

[20] Yongan Zhang et al., "Dna: Differentiable network-accelerator co-search," 2020.

[21] Mingxing Tan et al., "Mnasnet: Platform-aware neural architecture search for mobile," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[22] Andrew Howard et al., "Searching for mobilenetv3," in *Proceedings of the IEEE ICCV*, 2019.

[23] Han Cai et al., "Proxylessnas: Direct neural architecture search on target task and hardware," *arXiv:1812.00332*, 2018.

[24] Hanxiao Liu et al., "Darts: Differentiable architecture search," *arXiv:1806.09055*, 2018.

[25] Yi-Chen Chen et al., "Darts-asr: Differentiable architecture search for multilingual speech recognition and adaptation," *arXiv:2005.07029*, 2020.

[26] Tong Mo et al., "Neural architecture search for keyword spotting," *arXiv:2009.00165*, 2020.

[27] Jihwan Kim et al., "Evolved speech-transformer: Applying neural architecture search to end-to-end automatic speech recognition.," in *INTERSPEECH*, 2020.

[28] Rishi Bommasani et al., "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.

[29] Dongdong Wang et al., "Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model," in *Proceedings of the IEEE CVPR*, 2020.

[30] Zhong Meng, Jinyu Li, Yong Zhao, and Yifan Gong, "Conditional teacher-student learning," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6445–6449.

[31] Zhiqiang Shen et al., "A fast knowledge distillation framework for visual recognition," *arXiv:2112.01528*, 2021.

[32] Steffen Schneider et al., "wav2vec: Unsupervised pre-training for speech recognition," *arXiv:1904.05862*, 2019.

[33] Chaojian Li et al., "Hw-nas-bench: Hardware-aware neural architecture search benchmark," *arXiv:2103.10584*, 2021.

[34] Yonggan Fu et al., "Depthshrinker: A new compression paradigm towards boosting real-hardware efficiency of compact neural networks," *arXiv:2206.00843*, 2022.

[35] Taman Narayan et al., "Predicting on the edge: Identifying where a larger model does better," *arXiv:2202.07652*, 2022.

[36] Vassil Panayotov et al., "Librispeech: an asr corpus based on public domain audio books," 2015.

[37] Annamaria Mesaros et al., "Acoustic scene classification: An overview of DCASE 2017 challenge entries," in *2018 16th IWAENC*, September 2018.

[38] Google LLC., "Performance measurement," `https://www.tensorflow.org/lite/performance/measurement`, accessed 2021-05-21.

[39] Qualcomm Technologies, Inc., "Snapdragon Profiler," `https://developer.qualcomm.com/software/snapdragon-profiler`, accessed 2022-03-21.

[40] David Snyder et al., "Musan: A music, speech, and noise corpus," *arXiv:1510.08484*, 2015.

[41] Akhil et al Mathur, "Libri-adapt: a new speech dataset for unsupervised domain adaptation," in *ICASSP 2020-2020 ICASSP*. IEEE, 2020, pp. 7439–7443.