ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890314

AMSER: Accelerate Mobile Speech Emotion Recognition with Signal Compression

Yu Lu [†‡&], Ran Wang [†‡&], Dian Ding[†‡*], Han Zhang[†‡],
Liyun Zhang[†‡], Lanqing Yang[†‡], Yi-Chao Chen[†‡], Guangtao Xue[†‡*]
[†] Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[‡] Shanghai Key Laboratory of Trusted Data Circulation and Governance, and Web3
Email:{yulu01, wang_r, dingdian94, han_zhang, zhang_ly, yanglanqing, yichao, gt_xue}@sjtu.edu.cn

***Abstract*—Speech-based interaction systems are widely used in mobile devices like smartphones. With advances in deep neural networks, tasks such as speech emotion recognition (SER) enhance these systems' user-friendliness. However, deploying SER models on mobile devices is challenging due to their complexity and computational demands. While pruning can reduce complexity, it often compromises accuracy, and hardware accelerators like FPGAs are difficult to integrate into mobile devices. This paper proposes AMSER, a real-time speech emotion recognition framework using signal compression and task offloading. AMSER utilizes logarithmic Mel-filter bank coefficients (Fbank) and singular value decomposition (SVD) for feature extraction and compression. The compressed signal is only 6.25% of the original size, achieving 2.24x faster transfer rates and 55.35% energy savings compared to raw audio transmission. Despite the compression, the features preserve key audio information for text and emotion recognition, performed server-side. Experiments show a WER of 4.68% (Librispeech), 10.69% (CommonVoice), and 69.83% emotion recognition accuracy (IEMOCAP).**

***Index Terms*—Speech Emotion Recognition, Feature Compression**

I. INTRODUCTION

Speech is a prevalent interaction method in smartphones, stereos, and other IoT devices. The global speech and voice recognition market is expected to grow from $12.62 billion in 2023 to $59.62 billion by 2030 [1]. Unlike text, speech carries richer information such as emotion [2] and gender [3], [4]. Emotion recognition, in particular, enables intelligent systems to offer more personalized services [5]. For instance, customer service systems can adjust responses or assess business performance based on customer emotions.

Despite advancements in deep learning improving speech-related applications [6]–[8], the complexity of models and large parameter counts impose significant computational and storage demands. Mobile devices face limitations in processing power, energy consumption, and heat dissipation, making them unsuitable for such systems. Additionally, these interactive applications are highly sensitive to latency [9]–[11], which mobile devices struggle to meet. As a result, deploying real-time emotion recognition systems on mobile devices remains a critical challenge. Researchers have reduced model complexity on mobile devices using techniques like branch pruning [12], weight sharing [13], tensor quantization [14], and knowledge distillation [15]–[17], but these often reduce accuracy. Hardware solutions like GPUs [18], FPGAs [19], and ASICs [20], [21] improve computational capacity but are difficult to deploy on mobile devices due to size and power constraints.

We propose AMSER, a distributed speech emotion recognition framework using signal compression. Mobile devices handle speech acquisition and preprocessing through Mel-filter bank [22] coefficients (Fbank) and singular value decomposition (SVD) [23]. This reduces the processed sample size to 6.25% of the original, significantly lowering storage requirements. Deploying real-time speech applications on mobile devices faces several challenges. First, mobile devices have limited computing power, making it hard to support complex neural networks. Second, IoT devices like smart speakers lack storage for long-term audio data and large models. Lastly, current emotion recognition models rely solely on dataset knowledge, limiting their accuracy.

We propose AMSER to address these challenges by creating a real-time speech emotion recognition framework for mobile devices and servers. The system offloads deep neural network tasks to servers, reducing the computational and storage burden on mobile devices. It also compresses speech signals using Fbank features and SVD, minimizing storage needs. Finally, AMSER leverages the pre-trained RoBERTa model to incorporate external knowledge, enhancing emotion recognition accuracy.

Extensive experiments demonstrate the feasibility of deploying a real-time speech emotion recognition system on mobile devices. The key contributions of this paper are as follows:

- We propose AMSER, a speech emotion recognition system for edge mobile devices. Unlike traditional systems that offload all computations to the server, AMSER reduces transmission latency and optimizes resource usage on edge devices.
- We propose a feature extraction and compression module for audio signals, optimized for mobile devices. Using Fbank, the audio is converted into an acoustic spectrogram, with SVD applied to compress and filter out high-frequency redundant information.
- We constructed a neural network for speech emotion recognition based on the whisper [6] and RoBERTa [7] models.

[&] Both authors contributed equally to the research.
[*] Guangtao Xue and Dian Ding are the corresponding authors.

- Extensive experiments show that compared to direct raw audio transfer, AMSER improves transfer rates by 2.24x, reduces energy consumption by 55.35%, and achieves a 6.25% file compression ratio. On the IEMOCAP dataset, it achieves 69.83% accuracy and an F1-score of 0.698.

II. RELATED WORK

A. Deep Neural Network Deployment

Deploying DNN models on edge devices is a common challenge in AI fields like NLP and computer vision. Solutions such as Vigil [24], Reducto [25], Filter-Forward [26], and Glimpse [27] implement selective data offloading to minimize latency based on feature type, filtering thresholds, and content. Cracking open the DNN [28] enhances video analytics through joint camera-cloud inference and continuous online learning. Elf [9] improves mobile deep vision by distributing inference tasks to multiple servers. Remix [29] optimizes object detection on edge devices with image partitioning strategies under latency constraints. AMSER offers a real-time speech emotion recognition framework via compression and task offloading.

B. Speech Emotion Recognition

Recent research in Speech Emotion Recognition (SER) has leveraged deep learning techniques. Xu et al. [30] introduced an attention-based network that aligns textual and audio information for feature extraction. Yoon [31], [32] developed a dual RNN encoder model that integrates text and audio signals. Delbrouck et al. [33] proposed UMNOS, a transformer-based model for single-sentence emotion recognition and sentiment analysis.

III. PRELIMINARY STUDY

In speech recognition tasks, methods like MFCC or Fbank are commonly used to extract two-dimensional features from audio signals through windowed sampling. For example, OpenAI's Whisper [6] uses Fbank to extract acoustic spectrograms from audio, followed by a transformer-based encoder-decoder model to convert the spectrogram into text labels.

Features extracted through Fbank often contain redundant information, with high-frequency details offering limited utility in systems like Whisper. Similar to image compression, where high-frequency details can be removed without losing key information, we propose using the SVD algorithm to compress acoustic spectrograms. This preserves low-frequency features while reducing dimensionality for better identification and classification.

We verify the efficacy of SVD for compressing audio features within the Whisper speech recognition framework. In the Whisper framework, the speech signal $s \in \mathcal{R}^t$ undergoes extraction by Fbank to yield the acoustic spectrogram feature matrix $f \in \mathcal{R}^{m \times n}$:

$$f = Func_{Fbank}(s) \tag{1}$$

Let $k = min(m, n)$, then we compute the SVD of matrix f:

$$\begin{aligned} &f = Udiag(S)V^H \\ &U \in \mathcal{R}^{m\times k}, S \in \mathcal{R}^k, V \in \mathcal{R}^{n\times k} \end{aligned} \tag{2}$$

where $diag(S) \in \mathcal{R}^{k\times k}$, V^H is the conjugate transpose when V is complex, and the transpose when V is real-valued, and the matrices U, V are orthogonal in the real case, and unitary in the complex case. In this scenario, singular values S are sorted in descending order and are distinct. Denoting them as $\sigma_1 > \sigma_2 > \sigma_3 \cdots > \sigma_k$. Then f can be expressed as the following decomposition:

$$f = Udiag(S)V^H = \sum_{i=1}^{k} \sigma_i \begin{pmatrix} | \\ u_i \\ | \end{pmatrix} \begin{pmatrix} - & v_i & - \end{pmatrix} \tag{3}$$

where $U = (u_1, u_2, \ldots, u_k)$ and $V^H = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix}$.

Considering that the contribution of these singular values to the matrix shrinks sequentially, then according to the Eckhart-Young theorem [34],we can take the compression approximation of the acoustic spectrogram features:

$$f \approx f^{'} = \sum_{i=1}^{r} \sigma_i \begin{pmatrix} | \\ u_i \\ | \end{pmatrix} \begin{pmatrix} - & v_i & - \end{pmatrix} \tag{4}$$

where $r \in \mathcal{N} \cap [1, k]$, and $\frac{r}{k} \in [\frac{1}{k}, 1]$ denotes the compression rate for acoustic spectrogram features. In contrast to the original method where we needed to store U, S, V to recover f, now we only need to save $U^{'} \in \mathcal{R}^{m\times r}, S^{'} \in \mathcal{R}^r, V^{'} \in \mathcal{R}^{r\times n}$ to recover $f^{'}$, resulting in a saved matrix size equal to $\frac{r}{k}$ of the original.

Subsequently, we compress the Librispeech [35] and CommonVoice [36] datasets at various compression rates and assess the Whisper system's performance in recognizing the compressed acoustic spectrogram features. As a common metric of the performance of a speech recognition or machine translation system, word error rate (WER) is employed to evaluate the performance of whisper on both datasets and can be caculated by the following formulation:

$$WER = \frac{S + D + I}{S + D + C} \tag{5}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and C is the number of correct words. The results depicted in the Fig. 1 demonstrate that when the compression rate exceeds 10%, the Whisper system exhibits commendable speech recognition performance even for compressed speech.

Although edge devices may lack the computational power for large-scale models, extracting Fbank features and compressing them for server transmission is feasible. Compared to direct audio file transmission, sending compressed spectrograms reduces bandwidth usage and communication time. Previous studies show that SVD-based compression at 12.5% for spectrograms (6.25% for audio files) minimally impacts ASR performance. AMSER will further verify that this compression rate maintains accuracy in speech sentiment analysis.

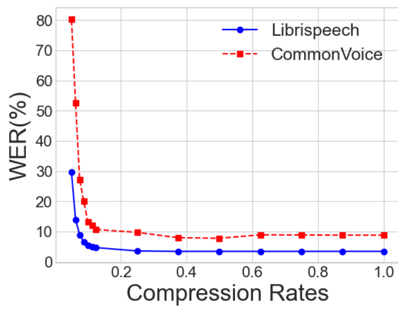

Fig. 1. Impact of Compression Rate for Whisper.

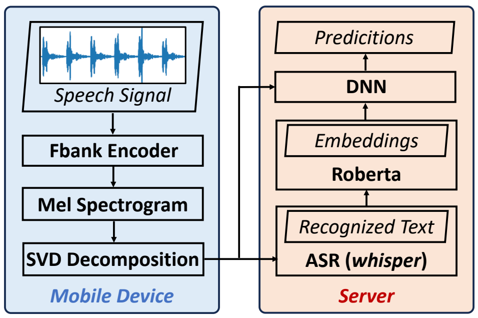

Fig. 2. The system architecture of AMSER.

IV. SYSTEM

We present AMSER, a real-time speech emotion recognition system. It consists of two parts: the mobile device acquires speech and extracts features using an Fbank encoder to output a Mel Spectrogram, which is then compressed to reduce storage. The compressed features retain text and emotion information, and the server performs text and emotion recognition using Whisper and RoBERTa embeddings. The system architecture is shown in Fig. 2.

A. Signal Preprocess

1) Feature Extraction: The mobile device extracts Fbank features from the user's speech through pre-emphasis, frame splitting, windowing, short-time Fourier transform, and Mel filtering. Pre-emphasis enhances high-frequency signals, while frame overlap prevents abrupt changes. A Hamming window smooths the signal, and FFT converts it to the frequency domain. Mel filtering then aligns the features with human auditory perception.

2) Signal Compression: In addition to computing power, the limitation of storage space is also not negligible for mobile devices. The system uses SVD described in detail in Sec III to compress the speech features, de-noising, and retaining the textual and emotional information in the features as much as possible.

B. Emotion Recognition

1) Modality Input: The server performs emotion recognition on the compressed features sent from mobile devices (see Fig. 3). First, the compressed features, derived from the Mel spectrogram (via STFT), capture signal energy changes over time, aligning with auditory perception. After SVD decomposition, the features retain text information, which is converted into text using ASR. Pre-trained RoBERTa [37] is then applied to enrich the features with external knowledge.

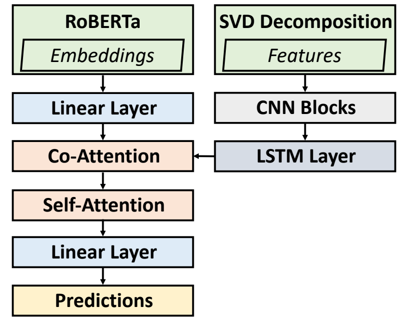

Fig. 3. The architecture of deep neural network.

2) Multi-modal fusion: For the features after SVD decomposition, the system uses Conv-BatchNorm-ReLU structure to further extract the features in time and frequency dimensions of the speech signal, and extracts the deeper features in time dimension by LSTM layer. In addition, the feature extracted from RoBERTa is a 1024-dimensional vector, which have good temporal structure and contain rich information. The system uses a linear layer for dimensionality adjustment for subsequent multimodal fusion and information compression.

Then the system fuses the compressed audio features and the RoBERTa coded features, introducing external knowledge from the outside world into the knowledge within the dataset with the help of a pre-trained model. The fusion process is divided into two phases: extracting features from one modality using the knowledge of the other, and subsequently, merging these additional extracted features into a single representation.

Specifically, in the first stage the system uses a Co-Attention module to achieve cross-modal feature extraction (Fig. 4), and the module employs an encoder-decoder structure stacking multiple layers of attention modules [38]. In this, the first modality uses Self-Attention to extract deep information about itself. Subsequently, the second modality performs a Self-Attention operation, during which a Guided-Attention is performed to extract more information, considering both modalities simultaneously. Both Self-Attention and Guided-Attention are based on the attention mechanism [39]. The attention module helps to construct a holistic view of the entire time span of the speech process. The attention consists of a query q, a key k and a value v:

$$Attention(q, k, v) = softmax(\frac{qk^T}{\sqrt{k}})v \tag{6}$$

Unlike simply using the self-attentive output of another modality as the input depth for guided attention, utilising the final output of the self-attentive layer provides richer information and more accurate guidance. The features of the two modalities are fused through the concatenation method, as both lack a unified temporal structure. The concatenation method retains more information and facilitates the fusion of knowledge from the external world with knowledge from within the dataset.

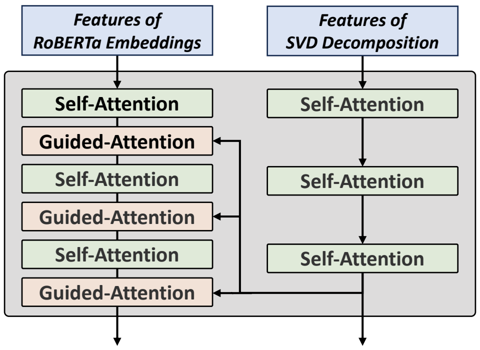

Fig. 4. The architecture of co-attention module.

V. EVALUATION

A. Dataset

We use IEMOCAP [40] to evaluate our AMSER system and train and test our model based on this dataset. IEMOCAP, a Multimodal Emotion Recognition dataset, comprises 151 recorded dialogue videos. Each segment in it is annotated for the presence of various common emotions (angry, happy, neutral, and sad), along with valence, arousal, and dominance. The recordings span 5 sessions involving 5 pairs of speakers.

B. Experimental Setup

1) Device: **Sever.** We utilize a server equipped with 188 GB of RAM and a $48.0GB$ VRAM's NVIDIA A40 as our evaluation system for model training and testing. **Client.** Redmi Note 12 Pro equipped with 8 GB of RAM and mediatek dimensity 1080 is used as a system client for audio file processing and compression.

2) Augmentation: We enhance the audio signals through three methods: introducing noise based on SNR, applying pitch shifts, and employing time stretching.

3) Model training: The model was trained for 100 steps with a batch size of 256. The optimizer used is Adam with a learning rate of $1e-5$ and a weight decay of 0. Meanwhile, we employ the Cross-Entropy loss to optimize the model.

4) Evaluation Metrics: We use the following two metrics to evaluate the effectiveness of our model on the speech emotion recognition task.

Accuracy. Speech emotion recognition, being a classification task, relies on accuracy as its fundamental evaluation metric. We employ accuracy to evaluate the core classification performance of the model.

F1 Score. We utilize the F1-score as an additional metric to ensure a more balanced evaluation of the model's performance. The F1-score is of the form: $F1 = \frac{2\cdot(precision\cdot recall)}{precision+recall}$

By taking both precision and recall into account, the F1-score can capture the bias in model predictions, indicating whether a model achieves high accuracy by correctly predicting the majority classes.

C. Micro Benchmark

1) Model Comparison: The experiments in this section validate the emotion recognition accuracy comparing different deep neural networks including UMONS [33], Xu [30] and Yoon [31], [32]. In addition, in order to verify the effect of signal compression on speech emotion recognition, the experiments evaluate the recognition accuracy under different compression rates. Firstly, compared with other networks, the proposed deep neural network introduces the knowledge of the external world, and the emotion performance accuracy is significantly higher than other networks, with an accuracy of 0.70126. Moreover, with the increase of compression rate, the emotion recognition accuracy only weakly decreases from 0.70126 to 0.69833, which is still significantly higher than that of other networks (Tab. I). The system measures the accuracy and recall of the model using F1-score as shown in Tab. II, demonstrating that the proposed deep neural network outperforms existing emotion recognition methods.

Compress Rate	Ours	UMONS	Xu	Yoon
12.50%	**0.69833**	0.67840	0.63343	0.55523
18.75%	**0.69540**	0.67644	0.63636	0.55914
25.00%	**0.69735**	0.67644	0.63742	0.56207
50.00%	**0.69840**	0.67742	0.63832	0.56891
100.00%	**0.70126**	0.67644	0.64321	0.58260

TABLE I
ACCURACY COMPARISON OF DIFFERENT MODELS

Compress Rate	Ours	UMONS	Xu	Yoon
12.50%	**0.69786**	0.67713	0.62987	0.54849
18.75%	**0.69486**	0.67539	0.63329	0.55306
25.00%	**0.69696**	0.67560	0.63298	0.55639
50.00%	**0.69688**	0.67654	0.63487	0.56381
100.00%	**0.70089**	0.67540	0.63968	0.57749

TABLE II
F1 SCORE COMPARISON OF DIFFERENT MODELS

Model	Raw	AMSER
transmission time	406.58s	180.75s
transmission energy overhead	0.0056kWh	0.0025 kWh

TABLE III
TRANSMISSION TIME AND ENERGY CONSUMPTION

2) Energy: In this section, the experiment verifies the effect of signal compression on power consumption. We utilize the compression rate of 6.25% for the 22,366 files transferred. Compared to translate the raw audio files, AMSER achieves a 2.24 times improvement in transfer rates and reduces energy overhead by 55.35%.

VI. CONCLUSION

We propose AMSER, a real-time speech emotion recognition framework for mobile devices. The system offloads deep neural network computations to a server, reducing mobile device load. Speech signals are compressed using Fbank features and SVD, minimizing storage requirements. By leveraging a pre-trained RoBERTa model, the system enhances emotion recognition accuracy. Extensive experiments validate its feasibility for mobile speech emotion recognition.

ACKNOWLEDGMENTS

This work is supported in part by National Natural Science Foundation of China (No. 61936015), Natural Science Foundation of Shanghai (No. 24ZR1430600) and Shanghai Key Laboratory of Trusted Data Circulation and Governance, and Web3.

References

[1] fortunebusinessinsights. speech-and-voice-recognition-market-101382. https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382, 2023.

[2] Ashish B Ingale and DS Chaudhari. Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(1):235–238, 2012.

[3] Ke Wu and Donald G Childers. Gender recognition from speech. part i: Coarse analysis. *The journal of the Acoustical society of America*, 90(4):1828–1840, 1991.

[4] Donald G Childers and Ke Wu. Gender recognition from speech. part ii: Fine analysis. *The Journal of the Acoustical society of America*, 90(4):1841–1856, 1991.

[5] Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion recognition using wireless signals. In *Proceedings of the 22nd annual international conference on mobile computing and networking*, pages 95–108, 2016.

[6] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[8] Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*, 2021.

[9] Wuyang Zhang, Zhezhi He, Luyang Liu, Zhenhua Jia, Yunxin Liu, Marco Gruteser, Dipankar Raychaudhuri, and Yanyong Zhang. Elf: accelerate high-resolution mobile deep vision with content-aware parallel offloading. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 201–214, 2021.

[10] Yu Lu, Dian Ding, Hao Pan, Yongjian Fu, Liyun Zhang, Feitong Tan, Ran Wang, Yi-Chao Chen, Guangtao Xue, and Ju Ren. M3cam: Extreme super-resolution via multi-modal optical flow for mobile cameras. In *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, pages 744–756, 2024.

[11] Yu Lu, Hao Pan, Feitong Tan, Yi-Chao Chen, Jiadi Yu, Jinghai He, and Guangtao Xue. Effectively learning moiré qr code decryption from simulated data. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2023.

[12] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

[13] Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression. *arXiv preprint arXiv:1702.04008*, 2017.

[14] Kristian Dokic, Marko Martinovic, and Dubravka Mandusic. Inference speed and quantisation of neural networks with tensorflow lite for microcontrollers framework. In *2020 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–6. IEEE, 2020.

[15] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019.

[16] Dian Ding, Lanqing Yang, Yi-Chao Chen, and Guangtao Xue. Leakage or identification: Behavior-irrelevant user identification leveraging leakage current on laptops. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(4), December 2022.

[17] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

[18] John Michalakes and Manish Vachharajani. Gpu acceleration of numerical weather prediction. In *2008 IEEE International Symposium on Parallel and Distributed Processing*, pages 1–7. IEEE, 2008.

[19] Sicheng Li, Chunpeng Wu, Hai Li, Boxun Li, Yu Wang, and Qinru Qiu. Fpga acceleration of recurrent neural network based language model. In *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*, pages 111–118. IEEE, 2015.

[20] Eriko Nurvitadhi, Jaewoong Sim, David Sheffield, Asit Mishra, Srivatsan Krishnan, and Debbie Marr. Accelerating recurrent neural networks in analytics servers: Comparison of fpga, cpu, gpu, and asic. In *2016 26th International Conference on Field Programmable Logic and Applications (FPL)*, pages 1–4. IEEE, 2016.

[21] Dingming Wu, Ang Chen, TS Eugene Ng, Guohui Wang, and Haiyong Wang. Accelerated service chaining on a single switch asic. In *Proceedings of the 18th ACM Workshop on Hot Topics in Networks*, pages 141–149, 2019.

[22] Stephen B Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[23] Kirk Baker. Singular value decomposition tutorial. *The Ohio State University*, 24:511, 2005.

[24] Tan Zhang, Aakanksha Chowdhery, Paramvir Bahl, Kyle Jamieson, and Suman Banerjee. The design and implementation of a wireless video surveillance system. In *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, pages 426–438, 2015.

[25] Yuanqi Li, Arthi Padmanabhan, Pengzhan Zhao, Yufei Wang, Guoqing Harry Xu, and Ravi Netravali. Reducto: On-camera filtering for resource-efficient real-time video analytics. In *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, pages 359–376, 2020.

[26] Christopher Canel, Thomas Kim, Giulio Zhou, Conglong Li, Hyeontaek Lim, David G Andersen, Michael Kaminsky, and Subramanya Dulloor. Scaling video analytics on constrained edge nodes. *Proceedings of Machine Learning and Systems*, 1:406–417, 2019.

[27] Saman Naderiparizi, Pengyu Zhang, Matthai Philipose, Bodhi Priyantha, Jie Liu, and Deepak Ganesan. Glimpse: A programmable early-discard camera architecture for continuous mobile vision. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 292–305, 2017.

[28] John Emmons, Sadjad Fouladi, Ganesh Ananthanarayanan, Shivaram Venkataraman, Silvio Savarese, and Keith Winstein. Cracking open the dnn black-box: Video analytics with dnns across the camera-cloud boundary. In *Proceedings of the 2019 workshop on hot topics in video analytics and intelligent edges*, pages 27–32, 2019.

[29] Shiqi Jiang, Zhiqi Lin, Yuanchun Li, Yuanchao Shu, and Yunxin Liu. Flexible high-resolution object detection on edge devices with tunable latency. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 559–572, 2021.

[30] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. Learning alignment for multimodal emotion recognition from speech. *Proc. Interspeech 2019*, pages 3569–3573, 2019.

[31] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. Multimodal speech emotion recognition using audio and text. In *2018 SLT*, pages 112–118. IEEE, 2018.

[32] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. Speech emotion recognition using multi-hop attention mechanism. In *ICASSP 2019*, pages 2822–2826. IEEE, 2019.

[33] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *ACL 2020*, page 1, 2020.

[34] Achiya Dax. The eckart-young theorem and ky fan's maximum principle: Two sides of the same coin. In *Householder Symposium XVIII on Numerical Linear Algebra*, page 49. Citeseer, 2011.

[35] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[36] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

[37] Jochen Hartmann. Emotion english distilroberta-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.

[38] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6281–6290, 2019.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurlPS*, 30, 2017.

[40] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.