Efficient Speech Emotion Recognition Using Gating Functions

Esha Maheshwari, Michael Strouse, Soumyadip Nandi

Abstract

Speech emotion recognition (SER) presents challenges when efficiency is a primary concern, such as on edge devices for always-on listening. We developed a deep learning model with trainable binary gates to avoid unnecessary computation without impacting performance. This approach has many practical use cases, such as on smart home devices, and can also be deployed in privacy enhanced frameworks. Our method also demonstrates the utility of a few recent innovations in audio machine learning, such as the Conformer model and a dynamic hop-length calculation.

1 Introduction

Always-on data collection and analysis are becoming increasingly important through a wide range of domains. Internet-of-things, manufacturing sensors, healthcare monitoring, audio and visual surveillance, among others all provide consistent streams of data that can be applicable to machine learning techniques. However, applying deep learning models on edge devices for these use cases is both computational and memory intensive, making it difficult, slow, or even impossible. Furthermore, developing fully tuned-in models for always-on data invokes serious privacy concerns. Existing models that consider efficiency-driven dynamic deep-learning approaches are not applicable to all modalities and often neglect privacy considerations. For example, SkipNet was introduced in 2017 to increase efficiency in image classification using gating functions (Wang et al., 2017), but similar techniques were not implemented for audio tasks such as automatic speech recognition (ASR) until 2023 (Peng et al., 2023). We applied this approach to SER. Although other research on computational efficient SER has been done, this gating strategy that skips modules offers a novel approach to increase efficiency without losing accuracy.

Other efficiency SER research has often focused on designing lightweight models from scratch or from upstream tasks such as ASR. As an example, Ren et al. (2022) used self-distillation on wav2vec2 and Akinpelu et al. (2023) designed a compact version of the popular VGGNet. However, more dynamic strategies to avoid computation like gating functions have not been as well researched. For ASR, Peng et al. (2023) introduced trainable binary gates into a Transformer model, and Bittar et al. (2024) built off that research by implementing gates on a Conformer model for keyword spotting. Our overarching idea was inspired by these latter papers, as well as our many of our data preprocessing and model choices.

Our approach involves adding learnable binary gating functionality to a Conformer model in order to avoid unnecessary computation by skipping layers. The model is designed to efficiently classify emotions from statically labeled utterances and avoid additional computation on non-speech audio. This model was designed from scratch and takes as input MFB and MFB Delta features from emotion utterances. Our gated Conformer model achieved 65.0%, 64.7%, and 56.4% accuracy on the EMODB, RAVDESS, and CREMA-D datasets respectively. This was a 0.3% average increase in accuracy from our baseline non-gated Conformer. The gated Conformer achieved this while skipping on average 38.8% of the modules on speech input. Furthermore, on non-speech input from the MS-SNSD dataset, our model skipped on average 36.3% more of the modules to avoid computation. This makes our approach particularly applicable to computational restrictive use cases such as an always-on smart home device.

Our model takes an “always listening but ignore if not important” approach. The Conformer

encoder outputs detected features in a latent dimensionality. These features can be used for classification without decoding them into text data, while a larger ASR model can be invoked only on localized audio sections. Thus, although not directly enhancing privacy, our model encourages the use of privacy enhanced frameworks.

In summary, our efficient SER Conformer model with binary gating functionality effectively avoids unnecessary computation without losing performance and can be used in privacy-focused frameworks.

The paper is organized as follows: section 2 reviews related work, section 3 overviews the methods, section 5 discusses the results, section 6 highlights ethical considerations, and in section 7 we present our conclusions.

2 Related Work

The use cases for SER have increased as mobile and edge devices continue to benefit from extracting emotional context alongside textual data. This has led to an influx of novel efficiency approaches and lightweight models for SER. Ren et al.'s (2022) self-distillation approach simultaneously fine-tunes a pretrained upstream model while training shallower versions of itself to create a small yet effective SER classifier. Both Akinpelu et al. (2023) and Tursunov et al.'s (2020) approach minimizes the number of layers and parameters needed based on popular CNN models. Additionally, Zhong et al.'s (2020) research utilized architectural innovations such as depthwise separable convolution and inverted residuals to limit computation. However, the more limited amount of research on dynamic efficiency approaches for SER inspired our topic.

Overall, there has been extensive work on dynamic approaches to improve efficiency and performance in machine learning (Han et al., 2021). Examples include dynamic depth early exiting, dynamic width neuron skipping, gating functions, among numerous others. As this paper illustrates, it is of considerable interest to test the effectiveness of these strategies on more niche machine learning areas after they prove successful in relatable and upstream tasks.

Peng et al. (2023) and Bittar et al.'s (2024) research on learnable binary gates, described as 'input-dependent dynamic depth', offered strong accuracy-efficiency trade-offs. Furthermore, the 'plug-and-play' property of gating functions makes their research applicable to a wide range of model architectures (Han et al., 2021). In essence, any modular/layered design could potentially benefit from gates. Our work demonstrates this for a SER Conformer.

We were incentivized to choose a Conformer by Bittar et al.'s (2024) remark of the model as a "state-of-the-art" choice for ASR and Seo and Lee's (2022) description of its "remarkable performance" on SER.

Before discussing our data in section 3, it's important to note the limitations faced for developing SER models based on available datasets. Capturing audio data with emotion labels that represents a variety of demographics and added noise has been challenging. Jahangir et al.'s (2021) research discussed the clear lack of generalizability for SER classifiers due to audio conditions and environments. This is why we utilized multiple datasets in training and evaluation.

3 Methods

The model, training, and data preprocessing files can all be found at our group's public GitHub: https://github.com/mstrouse16/GatedConformer

3.1 Data and Input Processing

Our emotion data comes from the EMODB, RAVDESS, and CREMA-D datasets. We chose categorical emotion labels rather than continuous valence/activation labels as there were more usable datasets, allowing for data diversity. For non-speech audio, we used air conditioning audio from MS-SNSD, designating it with a categorical catch-all label of 'background'. The MS-SNSD dataset offered a variety of 'noisy' audio, but we found that including a diverse range of non-speech input made it more challenging to classify, and thus more challenging to skip. We also felt that including only a single background noise type was more realistic for potential use cases. As an example, a smart home device will likely only be exposed to a single background noise depending on where it is placed, such as a home's air conditioning unit. Air conditioning was ultimately chosen from MS-SNSD over other sounds, such as distant babble, based on its large amount of available data.

We split the emotion datasets into train, test-seen, and test-unseen groups. Test-seen contained seen speakers but unseen utterances and test

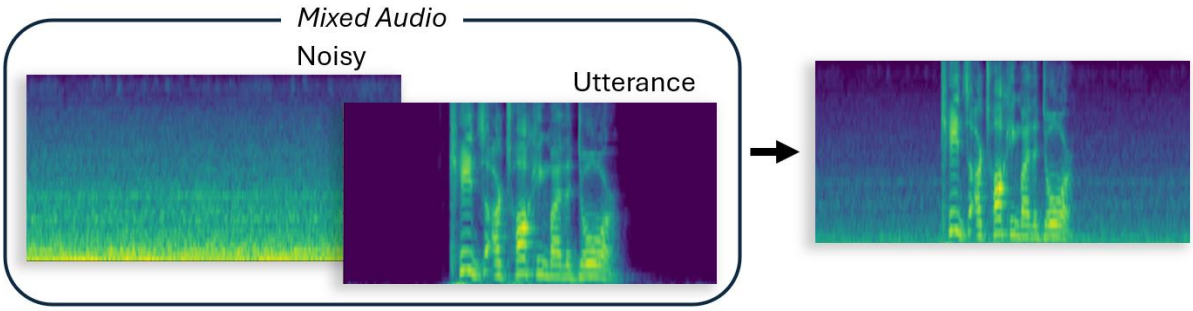


Figure 1: Illustration of audio mixing process

unseen contained completely unseen speakers. The air conditioning MS-SNSD audio was split into train and test. In order to expose our model to a more natural environment and more non-speech data, we overlayed every emotional utterance onto the air conditioning audio with a signal to noise ratio of 20-30dB. In practical terms, this kept the volume of the utterance the same, but varied the volume of the background air conditioning noise. The underlying air conditioning clips were chosen at random and from the separated train and test sets to ensure fairness in evaluation. The code to mix the data was provided through the MS-SNSD GitHub (https://github.com/microsoft/MS-SNSD). Additionally, we had air conditioning only audio that represented 20% of our overall data to test the model's ability to skip more computation on non-speech input. The emotion classes of neutral, happy, fear, disgust, anger, and sad had relatively equal weights. The mixing process is illustrated through spectrograms in Figure 1. The data distribution before the train/test split is shown in Figure 2.

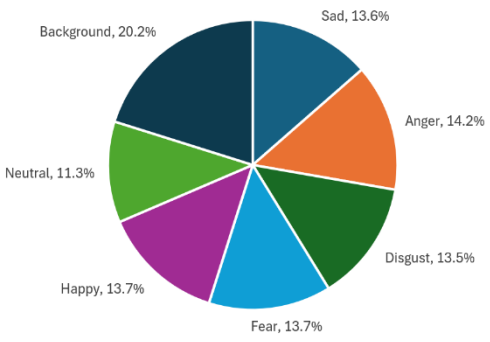


Figure 2: Distribution of data

From the audio files, we extracted MFB features with 40 filter banks and Delta. We chose MFBs over MFCCs because the Conformer's modules, especially the convolutional layer, can learn from the spatial relationship present among correlated MFB frequency components.

Lastly, we chose to use fixed length data to avoid masking and pooling complexity throughout our Conformer modules. In order to create same-sized data, we utilized a novel dynamic hop-length technique developed by Lin and Busso (2021). With a fixed window size and number of windows of 25ms and 360, the average hop length became ~10ms. In aggregate, there were 22,424 3.615 second clips, or 22.52 hours of data.

3.2 Conformer Encoder

Both Bittar et al.'s (2024) gated Conformer research and Gulati et al.'s (2020) original Conformer design discussed a variety of design and parameter choices, which is the primary reason we chose this architecture. Furthermore, as discussed in related work, Conformers have been quite successful on audio tasks.

The Conformer begins with additional preprocessing layers to reduce the time dimension for efficiency and expand the feature dimension to a desired latent size used throughout the model. Our input data was an 80 (40 filter banks + MFB Delta) x 360 (windows) matrix, and our latent model dimension was 120 (hidden features) x 90 (subsampled windows). Next, the data is sent through N Conformer blocks, each of which are identical. In each Conformer block, there are feedforward, multi-headed self-attention, convolution, and feedforward modules. The local characteristics that can be learned through the model's convolutional layer along with the global context captured in the attention layer are what allow this structure to be so effective. Some of the key Conformer parameters and our choices included the number of blocks (16), the kernel size in the convolutional layer (31), and the number of attention heads (4). The input dependent binary gates are locally placed before each module in each

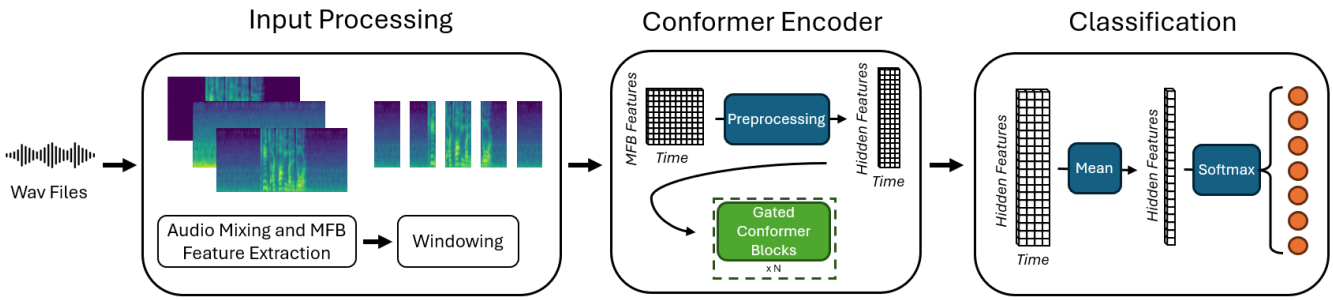


Figure 3: Speech emotion recognition gated Conformer pipeline

Conformer block. Importantly, the input and output dimensionality of the data for each module is the same, allowing residual connections to be utilized when skipping. A detailed diagram of the encoder is shown in Appendix 1.

We designed the Conformer from scratch, but we utilized a few public GitHubs to guide our design (https://github.com/jreremy/conformer, https://github.com/msalhab96/Conformer/blob/main/model.py).

3.3 Learnable Binary Gates

The primary issue with standard binary gates is that they are not differentiable. Bittar et al. (2024) utilized the Gumbel-Softmax trick to develop differentiable, and thus learnable, gates. We utilized Peng et al.'s (2023) implementation description and Jang et al.'s (2017) original research to develop gates specific to our model.

To start, the input feature matrix is averaged across the time dimension and sent through a linear layer. This layer outputs two logits that represent executing or skipping the current module. This provides a categorical distribution, and the Gumbel-Softmax trick creates a continuous representation of this categorical distribution in order to make it differentiable. Instead of directly sampling a softmax of the outputted logits, "noise" is sampled from a fixed Gumbel distribution that allows for the simulation of discrete choices during training (execute or skip). A temperature parameter, λ, is utilized to determine how extreme the final values are.

Bittar et al. (2024) mentioned they produced better results training the model without gates for a few epochs before enabling them, and we noticed a similar pattern. We defined a hyperparameter, *gate activation*, that enabled the gates after a certain percentage of epochs (we found the most success at 50% for 15 epochs).

During inference, a softmax is applied to the logits and the first probability, representing $P_{execute}$, is simply compared against a threshold (0.5) to determine if the module is executed or skipped.

A detailed diagram of the gate functionality during training is shown in Appendix 2.

3.4 Classification

The final layer averages across the time dimension and runs the resulting vector through a linear layer. A softmax then outputs the final classification probabilities which represent the probabilities of each categorical emotion plus 'background'.

3.5 Loss Function

We used a custom loss function with cross entropy loss as the base and a second term that averages $P_{execute}$ across every gate to encourage the model to skip modules. We found the most success applying a 0.1 weight to this second term.

Our overall SER pipeline is illustrated in Figure 3.

4 Results and Discussion

We trained our model using every dataset, and we used a 5-fold cross validation with grid search to optimize our parameter choices. To gauge performance, we focused on evaluating accuracy, as it provided a holistic score that was widely utilized in related work. In order to establish a baseline comparison for accuracy, we trained and evaluated our Conformer model without any gates and no altered architectural or parameter choices. Both our baseline and gated Conformer had ~5.6 million parameters and were trained on 15 epochs. After 15 epochs, the model would consistently overfit. We also compared our results against current state-of-the-art approaches.

To determine the effectiveness of our gating functionality, we created a *skip increase* metric.

Dataset	EMODB	RAVDESS	CREMA-D
VGG-optiVMD (Rudd et al., 2023) (SOTA)	96.1%		
VQ-MAE-S-12 (Frame) + Query2Emo (Sadok et a., 2023) (SOTA)		84.1%	
ViT with coordinates (Kim and Lee, 2023) (SOTA)			83.0%
Baseline Conformer (Test Seen, Test Unseen)	67.2%, 63.5%	64.8%, 59.8%	57.4%, 58.2%
Gated Conformer (Test Seen, Test Unseen)	70.9%, 59.1%	69.4%, 59.9%	57.8%, 55.0%

Table 1: Accuracy results on the EMODB, RAVDESS, and CREMA-D datasets.

This metric was the percentage increase of modules skipped on only non-speech input data compared to the average percentage of modules skipped across utterances with emotion labels. A higher *skip increase* means the model is able to avoid more computation on non-speech input. We then evaluated our gated Conformer on an aggregate set that contained test utterances from each emotion database as well as non-speech only input from the MS-SNSD set. As a reminder, every input had an air conditioning (non-speech) background, but around 80% of the inputs also had an emotional speech utterance overlayed on that background noise.

Table 1 displays a few state-of-the-art accuracies as well as our performance results. Our baseline Conformer did not meet state-of-the-art results, but it still achieved accuracies between 57 – 67%.

EMODB has very high state-of-the-art (SOTA) accuracies, and many architectures now perform above 90%, such as Rudd et al.'s (2023) VGG-optiVMD 96.1% that we include in the table. It was our highest performing dataset of the three at a 65.4% average between test seen and unseen. Similar to the SOTA results, our baseline performed worse on RAVDESS and CREMA-D. For those datasets, we achieved 62.3% and 57.8% (average between test seen and unseen) respectively.

Our original hypothesis was that the accuracy of the gated Conformer would decrease within 5-10%, but the results illustrate that the gated Conformer performed even better than our baseline on a few test sets. On average between test seen and unseen, we achieved 65.0%, 64.7%, 56.4% accuracy for the gated Conformer on EMODB, RAVDESS, and CREMA-D respectively. On average, this was a 0.3% increase in performance. This demonstrates that having local gates before each module can aid the model in determining which are most important. Furthermore, it can prevent overfitting as skipping a module helps reduce complexity and computation.

Input Type	% Skips
Sad	40.1%
Anger	39.1%
Disgust	37.5%
Fear	39.9%
Happy	38.8%
Neutral	37.4%
Avg. Across Emotions	38.8%
MS-SNSD A/C only	52.9%
Skip Increase	*36.3%*

Table 2: Percentage of modules skipped in gated Conformer on each input type.

Table 2 displays the percentage of modules skipped in our gated Conformer on each input type as well as our *skip increase* metric. In the gated Conformer, there were 16 blocks which each had 4 modules, which means there were 64 modules that could potentially be skipped at inference. For the emotional utterances, the model learned to consistently skip around 39% of these modules with little variation. When the input contained no emotional utterance and only background noise, it learned to skip 52.9% of the modules, or a 36.3% *skip increase*. Note that for this aggregate test set, the model had an additional categorical label of 'background' to classify, and it achieved an accuracy consistent with those on the individual datasets of 68.3% with a 0.7% increase from the baseline Conformer.

Bittar et al.'s (2024) research at Apple demonstrated the effectiveness of learnable binary gates on keyword detection. Their design skipped 42% of modules on keywords surrounded by background noise and 97% of modules on background noise only, representing a 223% skip increase. The success of their results was likely

driven by the fact that keywords are shorter utterances and have less variability per class. Our work demonstrates the effectiveness of gating functionality on a more complex speech classification task. With further tuning and a larger model, our skip increase would likely have continued to improve.

Our model's ability to avoid additional computation on non-speech input demonstrates the utility of this dynamic efficiency approach for many use cases. For example, edge and IoT devices that take in consistent audio streams could use this method under more computational restrictive environments. Furthermore, as described earlier, the 'plug-and-play' property of this technique makes it applicable to a wide range of model architectures and modalities.

5 Ethical Implications

The one ethical implication that is always present with machine learning technologies is bias and discrimination. SER systems trained on limited or biased datasets may not perform equally well across diverse demographics, potentially leading to the model skipping relevant information in test data. For example, although small, there is a slight difference in the number of female to male participants in the CREMA-D dataset. The EMODB dataset consists of only German speakers and considering the variations in speech and linguistics across languages, this could also be a point of bias in the model. To mitigate these concerns about bias and discrimination, it is critical to use datasets that represent a wide and fair range of demographics. Continuous evaluation and updating of the model with new data can also help reduce bias. Moreover, transparency about the limitations of the model and the datasets used for training, as well as efforts to improve inclusivity in the training data can help mitigate discriminatory outcomes.

As mentioned in our introduction, privacy is an important concern for always-on data collection, and despite our model's "always listening but ignoring if not important" approach, the fact remains that our model is still *always listening*. The constant analysis of vocal inputs for emotional states could lead to unwanted surveillance and misuse of private emotional data. To mitigate this ethical concern, the use of our model must require explicit user consent protocols, ensuring that users are fully aware of when and how their data is being recorded and used. Clear communication and transparent policies must be put in place so that users are aware of what data is being collected, how it is being used, for what purpose, and who has access to it. Consent should be a continuous process, allowing users to opt in or out at any stage.

Depending on the future use cases of our model, if our model is used to make or inform decisions that affect humans (e.g. hiring, law enforcement, customer service, etc.), it is crucial to ensure accountability for those decisions. It is possible that our model skips over important emotional information and results in inaccurate results. Decisions based on these inaccurately assessed emotions may lead to negative outcomes. To address this ethical issue, clear guidelines and regulations must be put in place that outline the responsibilities and accountability standards for when our model is part of the decision-making process. A log should be kept for decisions influenced by our model to help trace back any mistakes or biases.

6 Conclusion

Adding trainable binary gates into a Conformer model for SER successfully allowed the model to classify emotion more efficiently in speech, while avoiding additional computation on non-speech input. The SER conformer with the gating functionality performed, on average, 0.3% better than the implemented baseline conformer. Most importantly, the gating mechanism allowed our model to substantially skip computationally intensive portions when processing non-speech inputs.

Overall, the developed model not only demonstrates an increase in accuracy over non-gated baselines but also shows the practical usage of such dynamic efficiency strategies in environments such as IoT devices. By leveraging learnable binary gates, the benefits of the model extend to efficiency-driven approaches of audio processing tasks, where processing capabilities are limited.

Lastly, although our model may be intended for use in 'always-on' settings, we believe it can encourage privacy enhanced frameworks. The computational savings from our model allow it to be used upfront, and a more invasive and data collective model such as ASR can then be utilized on only localized sections of the total input.

References

Akinpelu, S., Viriri, S., & Adegun, A. (2023). Lightweight Deep Learning Framework for Speech Emotion Recognition. IEEE.

Bittar, A., Dixon, P., Samragh, M., Nishu, K., & Naik, D. (2024). Improving Vision-Inspired Keyword Spotting Using Dynamic Module Skipping in Streaming Conformer Encoder. Apple.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. arXiv.

Han, Y., Huang, G., Song, S., Yang, L., Wang, H., & Wang, Y. (2021). Dynamic Neural Networks: A Survey. arXiv.

Jahangir, R., Teh, Y. W., Hanif, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: state of the art and research challenges. Springer Link.

Jang, E., Gu, S., & Poole, B. (2017). Categorical Reparameterization with Gumbel-Softmax. arXiv.

Kim, J.-Y., & Lee, S.-H. (2023). CoordViT: A Novel Method to Improve Vision Transformer-Based Speech Emotion Recognition using Coordinate Information Concatenation. IEEE Xplore.

Lin, W.C., & Busso, C. (2021). Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling. IEEE Xplore.

Peng, Y., Lee, J., & Watanabe, S. (2023). I3D: Transformer Architectures with Input-Dependent Dynamic Depth for Speech Recognition. arXiv.

Ren, Z., Nguyen, T. T., Chang, Y., & Schuller, B. W. (2022). Fast Yet Effective Speech Emotion Recognition with Self-Distillation. arXiv.

Rudd, D. H., Huo, H., & Xu, G. (2023). An Extended Variational Mode Decomposition Algorithm Developed Speech Emotion Recognition Performance. arXiv.

Sadok, S., Leglaive, S., & Seguier, R. (2023). A Vector Quantized Masked Autoencoder for Speech Emotion Recognition. arXiv.

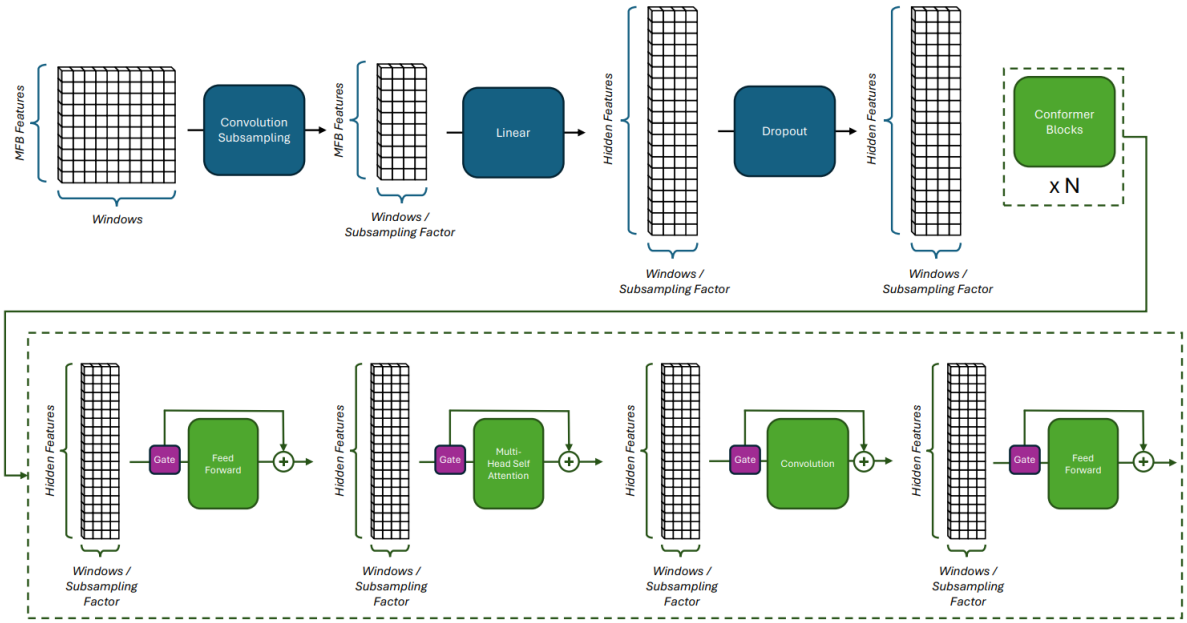
Seo, J., & Lee, B. (2022). Multi-Task Conformer with Multi-Feature Combination for Speech Emotion Recognition. MDPI.

Tursunov, A., Mustaqeem, & Kwon, S. (2020). Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features. MDPI.

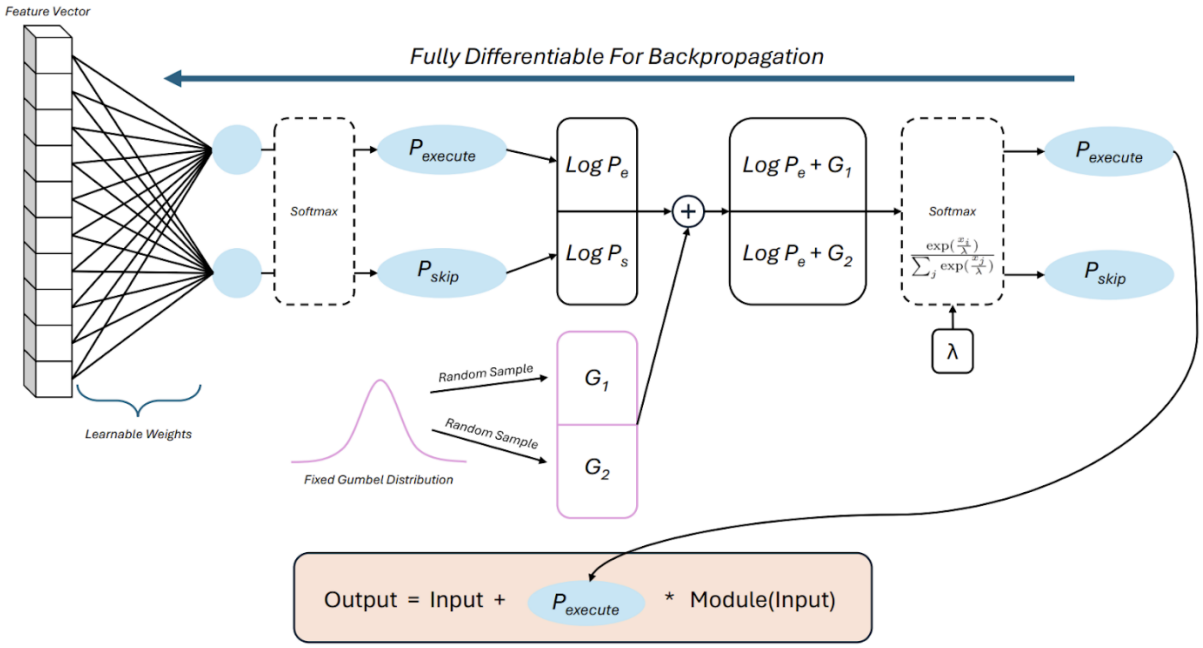
Wang, X., Yu, F., Dou, Z.-Y., Darrell, T., & Gonzalez, J. E. (2017). SkipNet: Learning Dynamic Routing in Convolux`tional Networks. arXiv.

Zhong, Y., Hu, Y., Huang, H., & Silamu, W. (2020). A Lightweight Model Based on Separable Convolution for Speech Emotion Recognition. arXiv.

A Appendices



Appendix 1: Conformer Encoder Architecture



Appendix 2: Input Dependent Gates – The Gumbel-Softmax Trick