Samsung Research

SMALL ENERGY MASKING FOR IMPROVED NEURAL NETWORK TRAINING FOR END-TO-END SPEECH RECOGNITION

Chanwoo Kim, Kwangyoun Kim, and Sathish Reddy Indurthi
{chanw.com, ky85.kim, s.indurthi}@samsung.com

Contents

Introduction-Regularization and Data Augmentation

- Regularization
  - L1/L2 regularization
  - Dropout [N. Srivastava, et. al, JMLR, 2014]

- Data Augmentation
  - SpecAugment [D. S. Park, et. al., INTERSPEECH 2019]
  - Acoustic Simulator [C. Kim, et. al., INTERSPEECH 2017]
  - Vocal Tract Length Perturbation, Speed Perturbation, etc.

- Data augmetnation itself can be considered as a way of applying regularization.

Motivation of Small Energy Masking

- Regularization is important for training large-size neural network models.
- In the conventional input-dropout, masking is applied completely randomly to the input features.
- In speech features, time-frequency bins with small energy may be more adversely affected by distortion or noise [C. Kim and R. M. Stern, ASRU 2009].
- → Applies masking more frequenty to time-frequency bins with smaller energy.

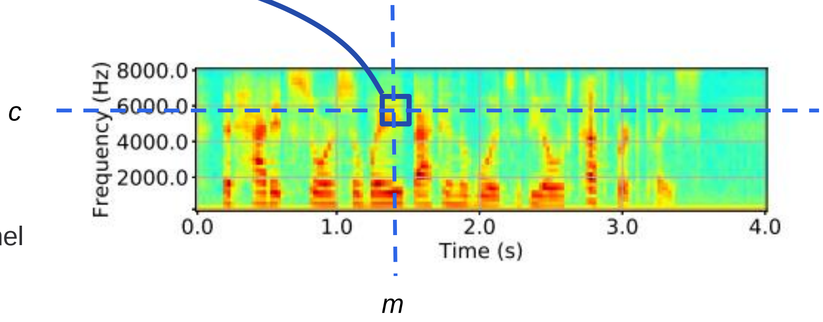
Motivation of Small Energy Masking- Filterbank Energy *e*[*m*, *c*] and Peak Filter Bank Energy e_{peak}

- *The filter bank energy* ***e[m, c]*** *in each time-frequency bin* is defined by:

$$e[m,c] = \sum_{k=0}^{K/2} |X[m,e^{j\omega_k}]|^2 M_c[e^{j\omega_k}]$$

Where

- *m* : Frame index
- *c* : Filterbank channel index
- $X[m, e^{j\omega_k}]$: Short-time Fourier Transform of the speech signal
- $M_c[e^{j\omega_k}]$: The frequency response of the c-th Filterbank channel



- *The peak filterbank energy* e_{peak} is defined to be the 95-percentile value of *e*[*m*, *c*] for each utterance. [C. Kim and R. M. Stern, ASRU 2009]

Motivation of Small Energy Masking-Distribution of Filterbank Energy

- η : The ratio of filterbank energy *e*[*m*, *c*] to the peak filterbank energy e_{peak} in dB:

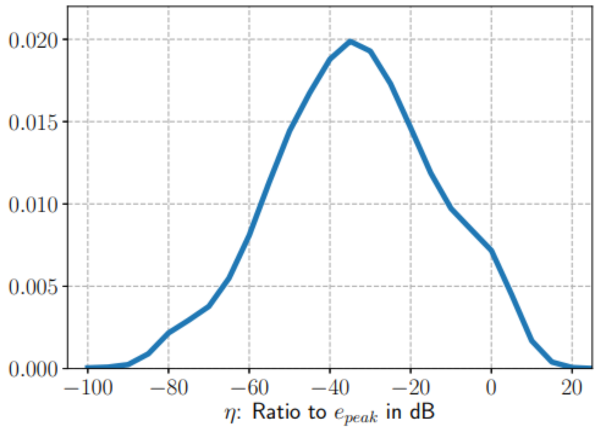
$$\eta = f\left(e[m,c]\right) := 10\log_{10}\left(\frac{e[m,c]}{e_{\text{peak}}}\right)$$

- The Probability Density Function (PDF) of η is shown on the right-hand side:

To calculate the statistical information shown in this slide, and in the next slide, we randomly selected 1,000 utterances from the LibriSpeech training set.

- The distribution mainly exists from -100 dB up to 20 dB.

Probability Density Function (PDF) of η



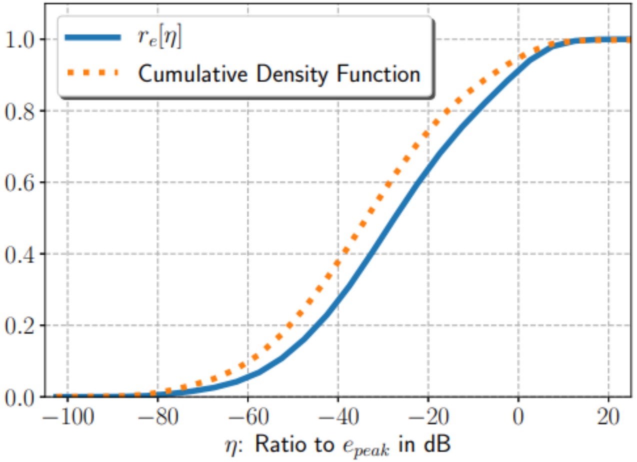
Motivation of Small Energy Masking - Cumulative density function and energy portion below the threshold

- The cumulative function η is shown on the right-hand side.

- We define $r_e(\eta_{th})$ as the portion of energy below the threshold η_{th} as shown blow:

$$r_e(\eta_{th}) = \frac{\sum_{f(e[m,c])<\eta_{th}} e[m,c]}{\sum e[m,c]}$$

- From this figure, if we select time-frequency bins whose energy is 20 dB below from e_{peak}, they comprise roughly 70 percent of all the bins, and 60 percent of the energy.



Small Energy Masing Algorithm - Algorithm Overview

- Selects a random energy ratio threshold (let's ca η_{th}) for for each utterance uniformly from the following interval.

$$\eta_{th} \sim \mathcal{U}(\eta_a, \eta_b)$$

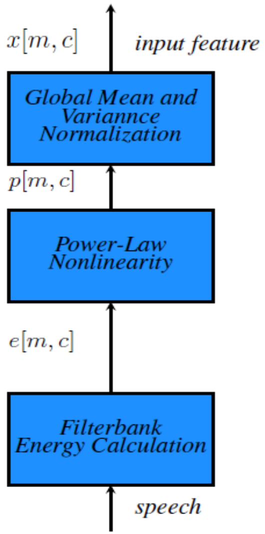
  - $\mathcal{U}$: Uniform distribution
  - η_a : The lower bound. We use the value of -80 dB.
  - η_b : The upper bound. We use the value of 0 dB.

- All the feature values below this ratio threshold is masked to have zero values.

- The unmasked feature values are scaled so that the sum is maintained.

Small Energy Masing Algorithm - Conventional Pipeline

- As a baseline system, we use the following power-mel feature pipeline.
  [C. Kim et. al., ASRU 2019, C. Kim et. al. INTERSPEECH 2019]

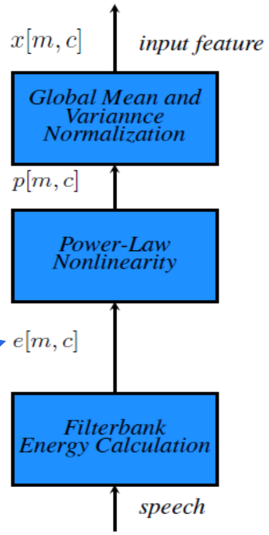


Small Energy Masing Algorithm - Conventional Pipeline

- As a baseline system, we use the following power-mel feature pipeline.
  [C. Kim et. al., ASRU 2019, C. Kim et. al. INTERSPEECH 2019]

Filterbank Energy

$$e[m,c] = \sum_{k=0}^{K/2} |X[m, e^{j\omega_k}]|^2 M_c[e^{j\omega_k}]$$



Small Energy Masing Algorithm - Conventional Pipeline

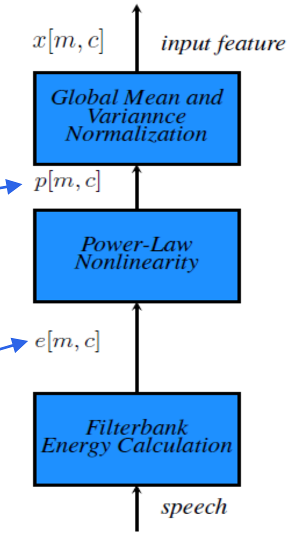
- As a baseline system, we use the following power-mel feature pipeline.
  [C. Kim et. al., ASRU 2019, C. Kim et. al. INTERSPEECH 2019]

Power-Mel Feature

$$p[m, c] = e[m, c]^{\frac{1}{15}}$$

Filterbank Energy

$$e[m, c] = \sum_{k=0}^{K/2} |X[m, e^{j\omega_k}]|^2 M_c[e^{j\omega_k}]$$



Small Energy Masing Algorithm - Conventional Pipeline

- As a baseline system, we use the following power-mel feature pipeline.
  [C. Kim et. al., ASRU 2019, C. Kim et. al. INTERSPEECH 2019]

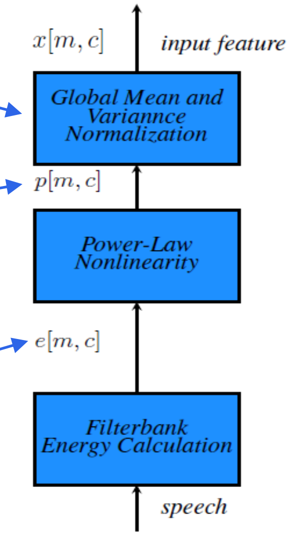
Just performs scaling and shifting of the feature globally.
Not an utterance-by-utterance mean variance normalization

Power-Mel Feature

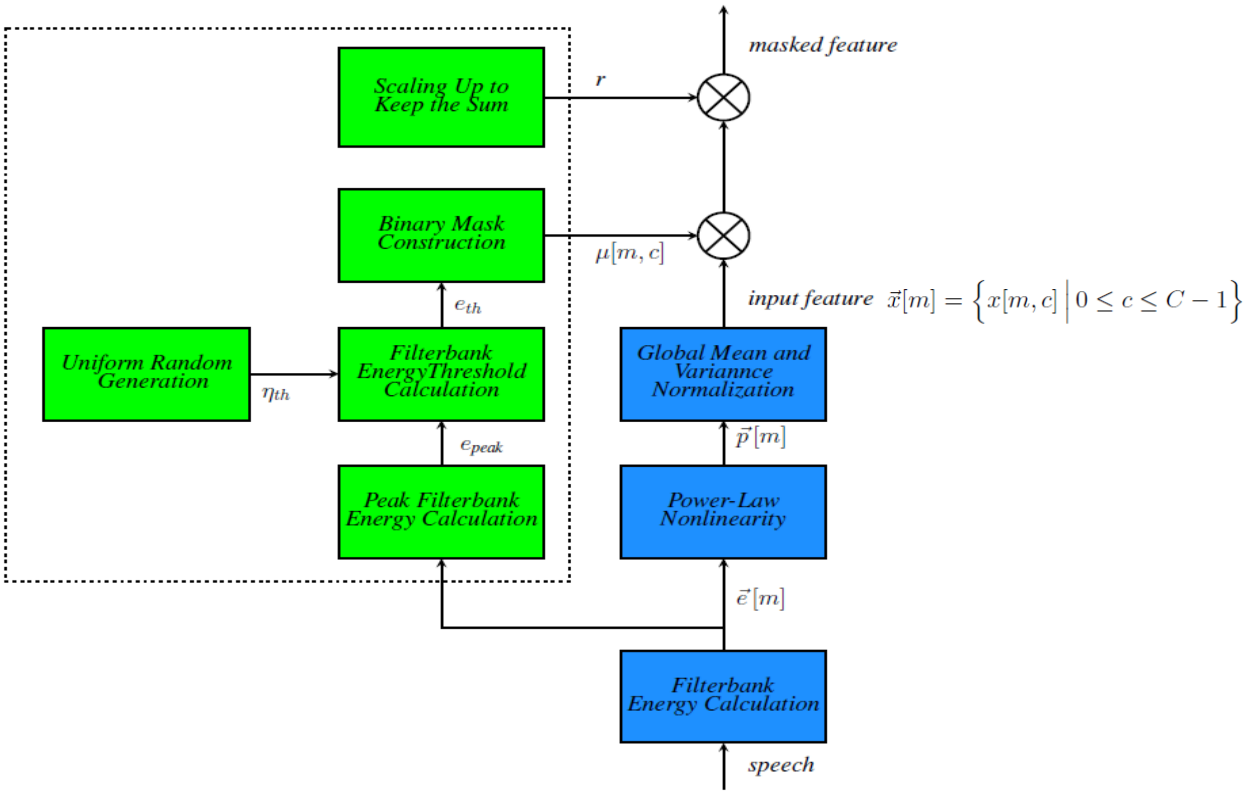
$$p[m,c] = e[m,c]^{\frac{1}{15}}$$

Filterbank Energy

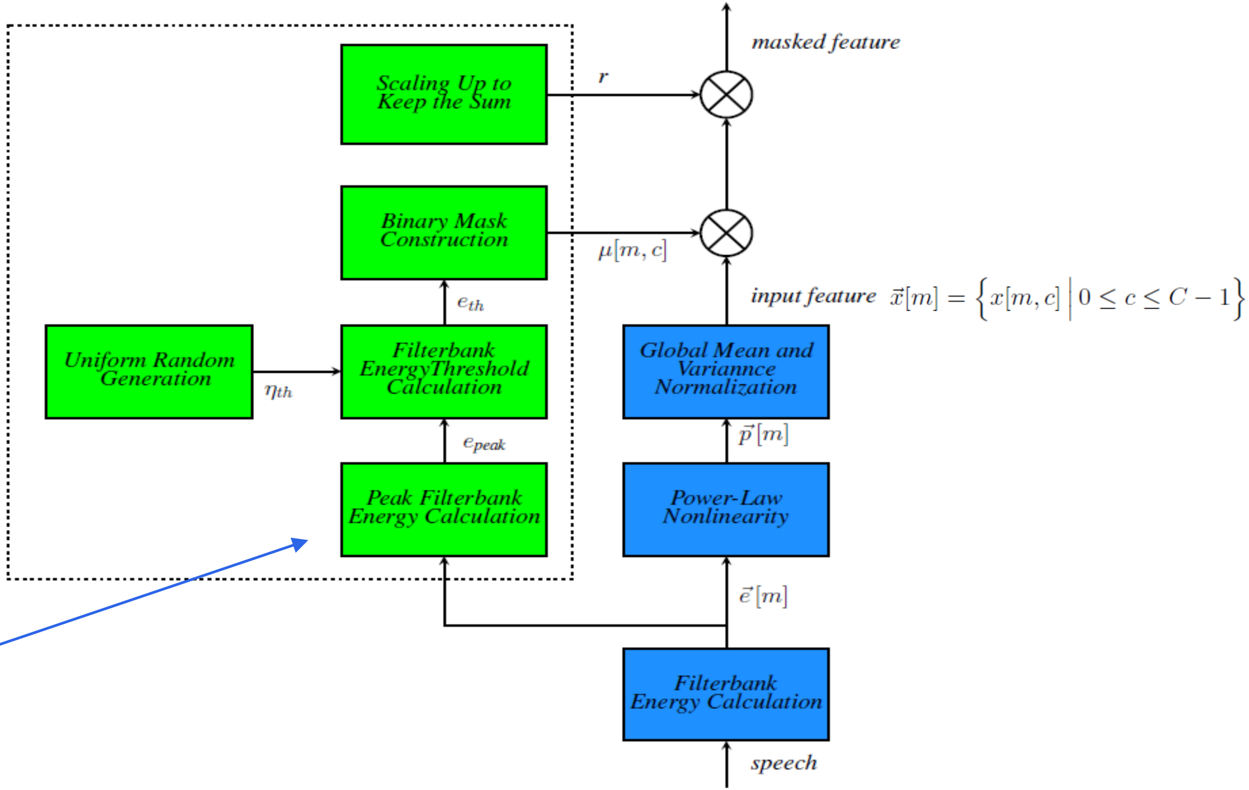
$$e[m,c] = \sum_{k=0}^{K/2} |X[m, e^{j\omega_k}]|^2 M_c[e^{j\omega_k}]$$



Small Energy Masing Algorithm - Masking Application



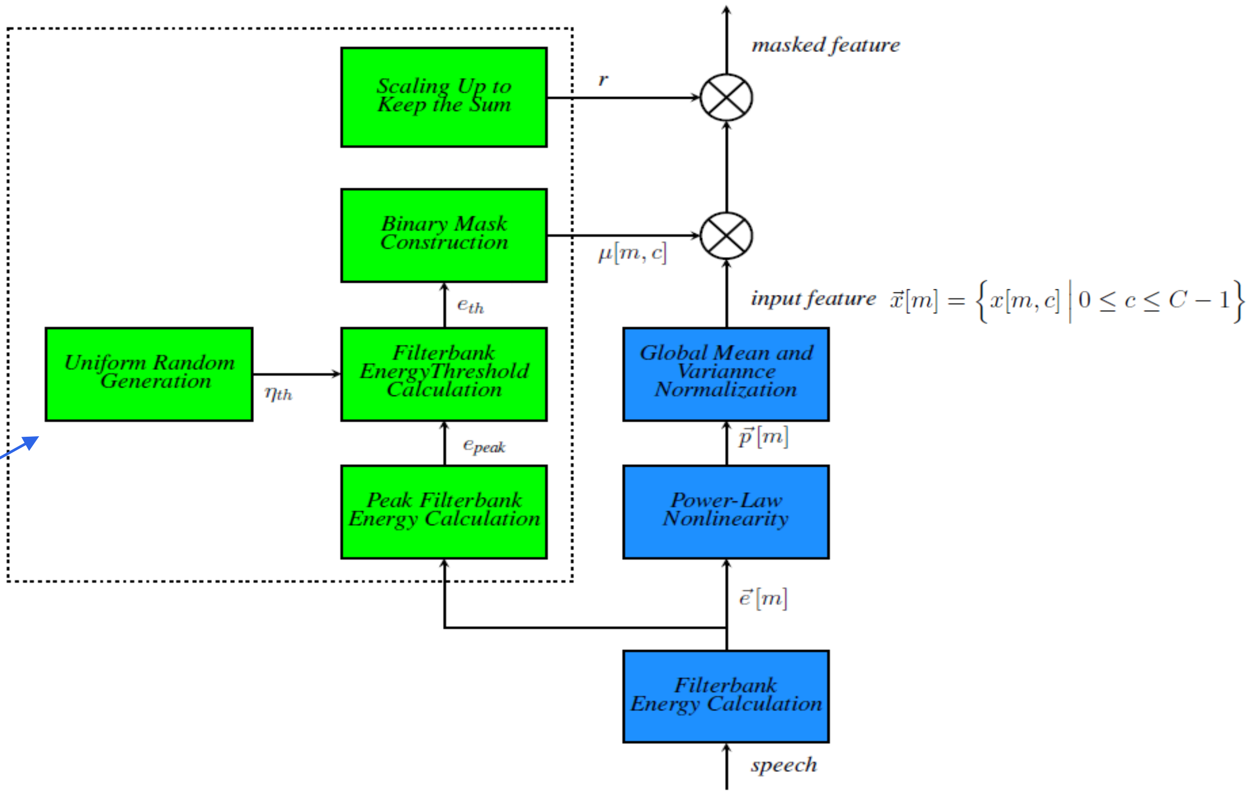
Small Energy Masing Algorithm - Masking Application



e_{peak} := The 95-th percentile of $e[m, c]$

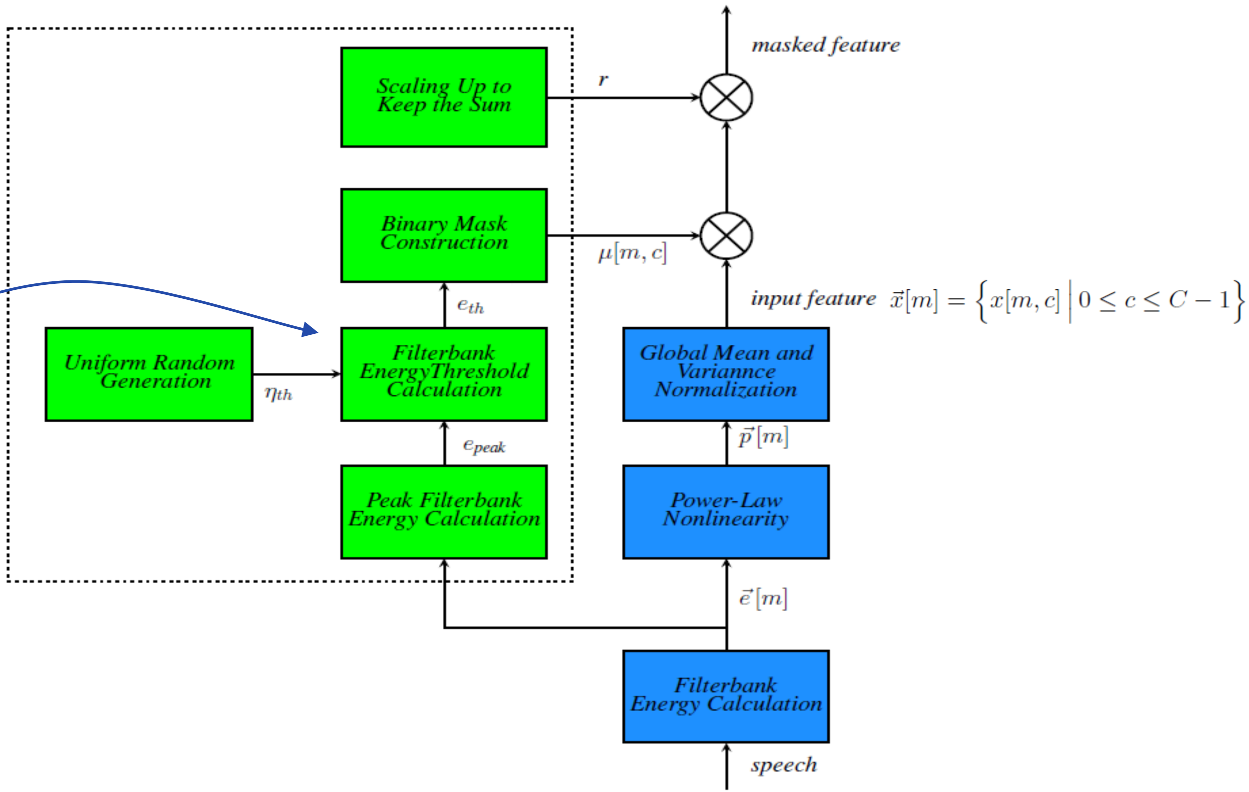
Small Energy Masing Algorithm - Masking Application

$$\eta_{th} \sim \mathcal{U}(\eta_a, \eta_b).$$



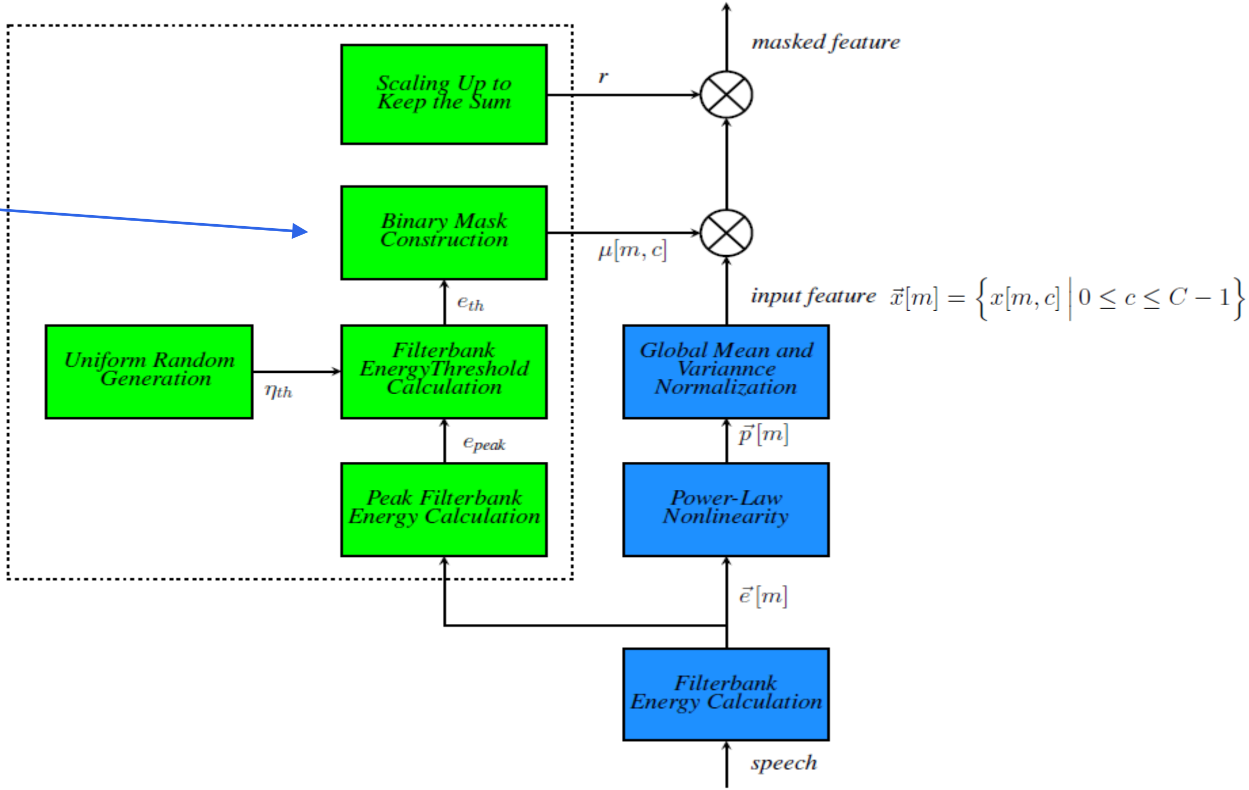
Small Energy Masing Algorithm - Masking Application

$$e_{th} = e_{\text{peak}} 10^{\frac{\eta_{th}}{10}}$$

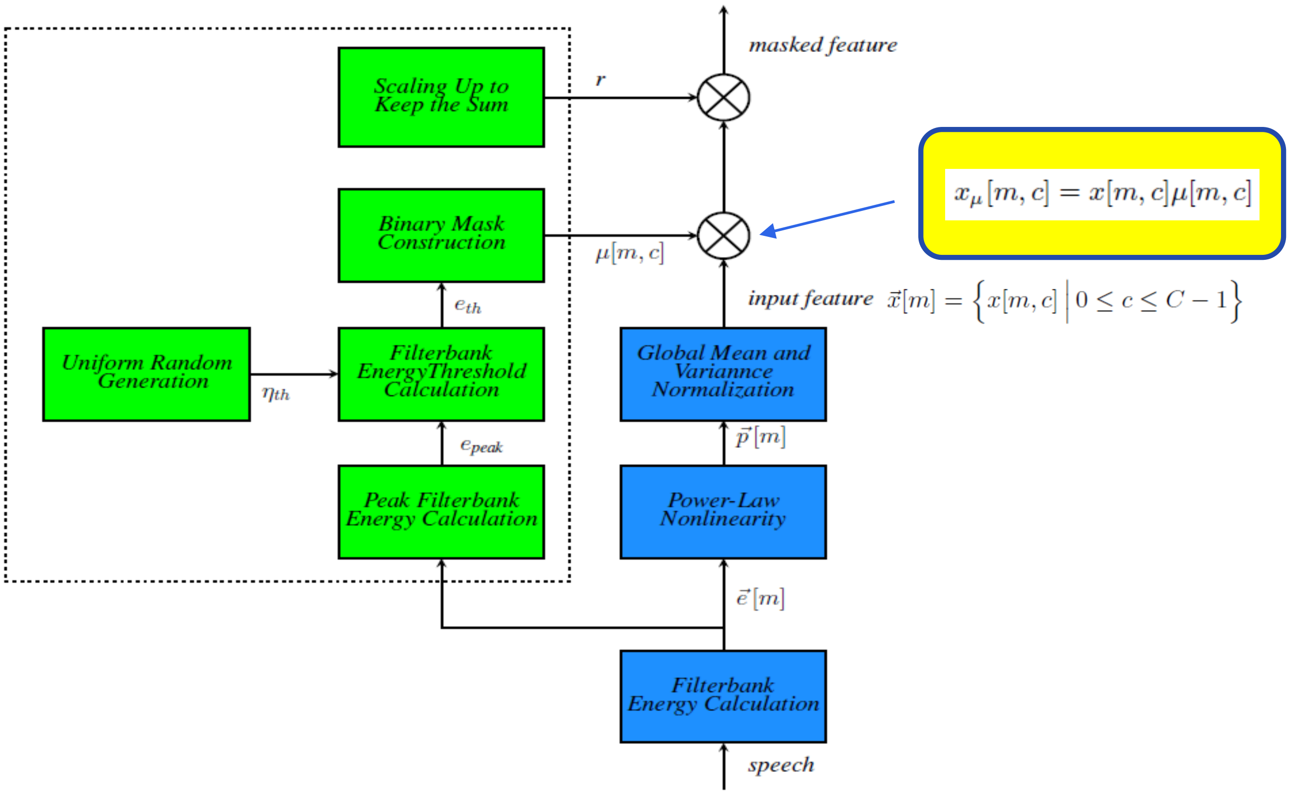


Small Energy Masing Algorithm - Masking Application

$$\mu[m,c] = \begin{cases} 1, & e[m,c] \geq e_{th}, \\ 0, & e[m,c] < e_{th}. \end{cases}$$

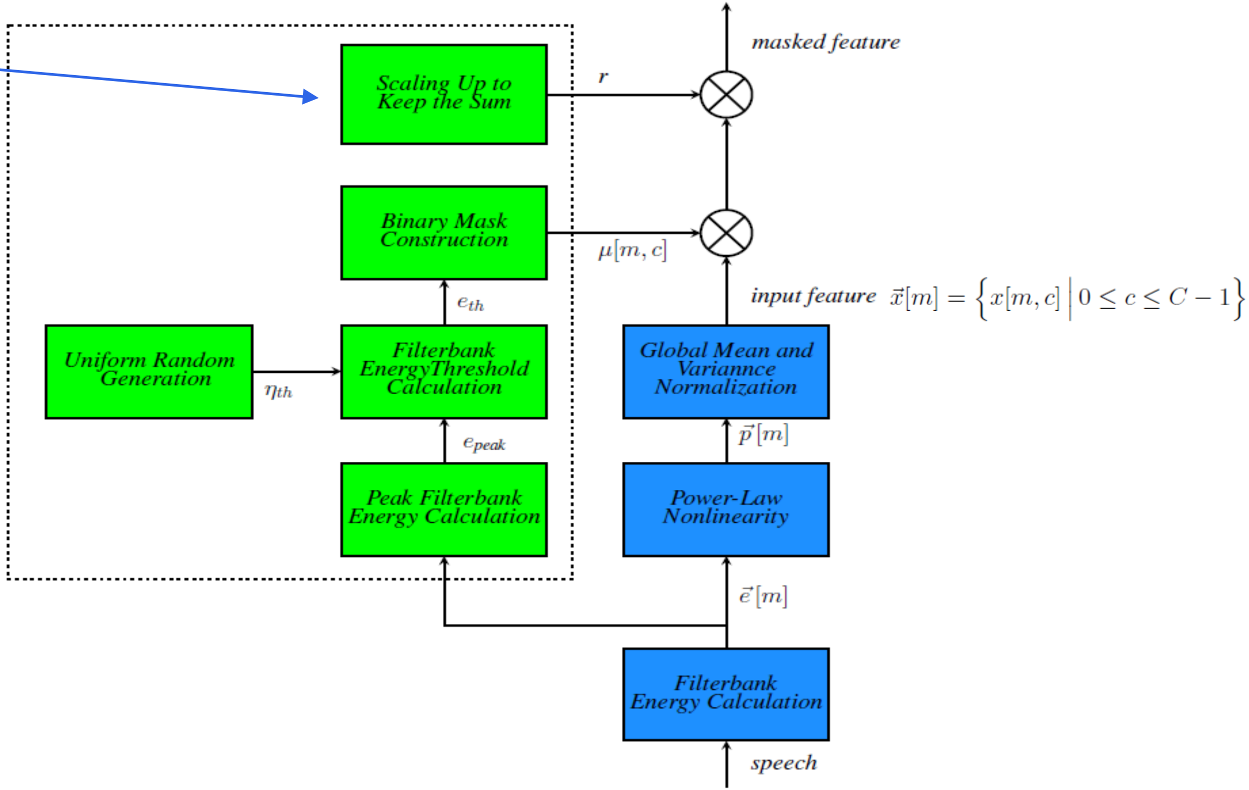


Small Energy Masing Algorithm - Masking Application

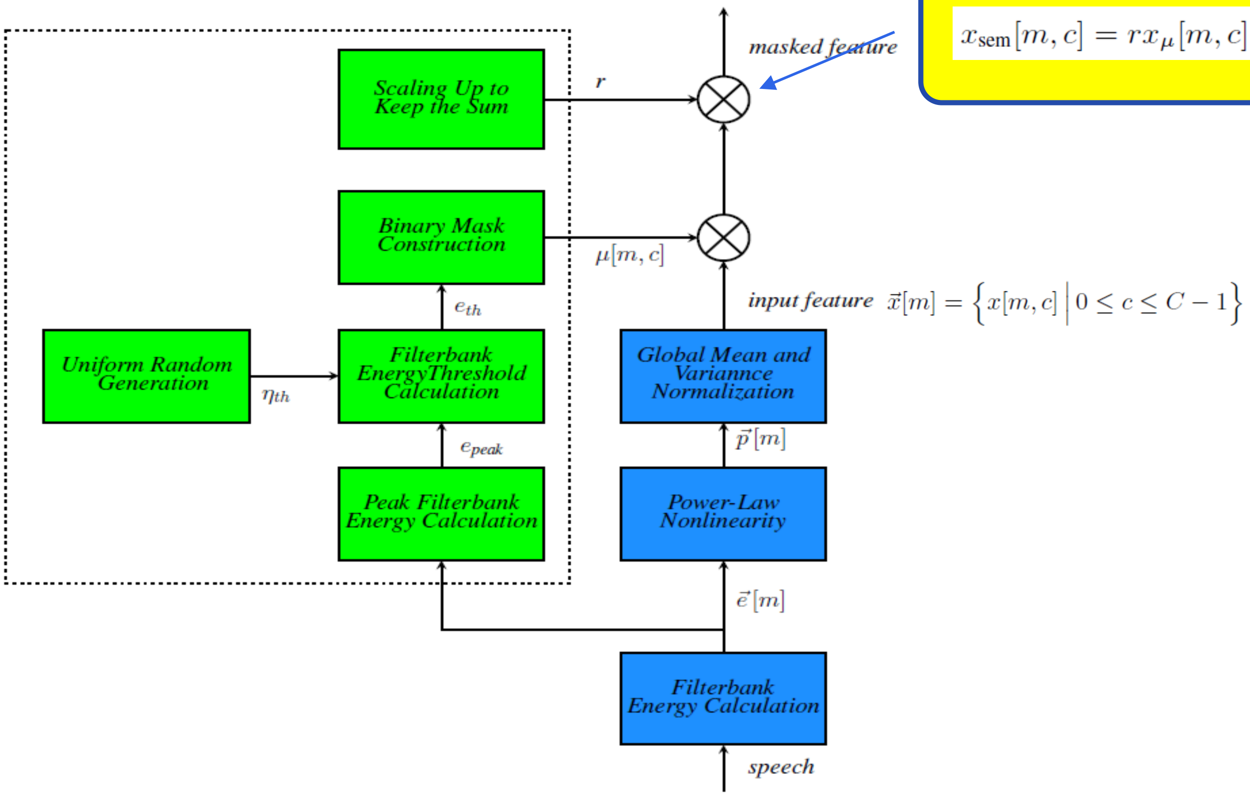


Small Energy Masing Algorithm - Masking Application

$$r = \frac{\sum_{\text{for each utt.}} x[m,c]}{\sum_{\text{for each utt.}} x_\mu[m,c]}$$

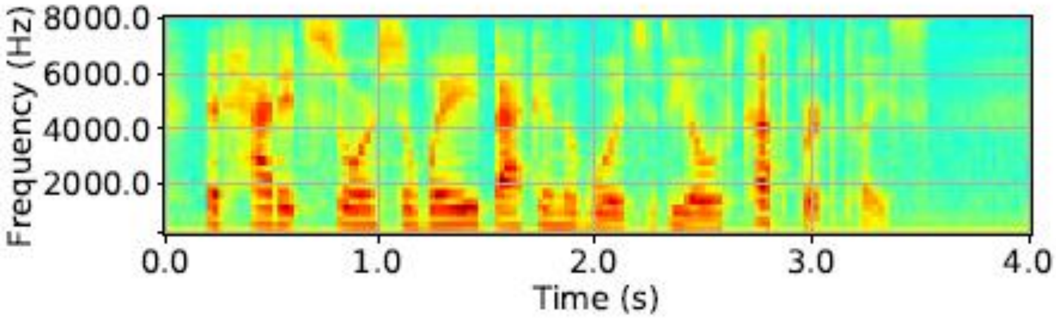


Small Energy Masing Algorithm - Masking Application



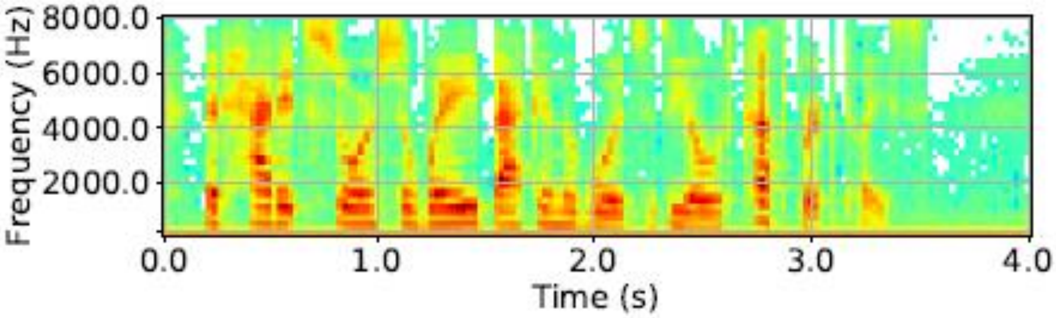
$$x_{\text{sem}}[m, c] = r x_{\mu}[m, c]$$

Small Energy Masing Algorithm - Original Spectrogram



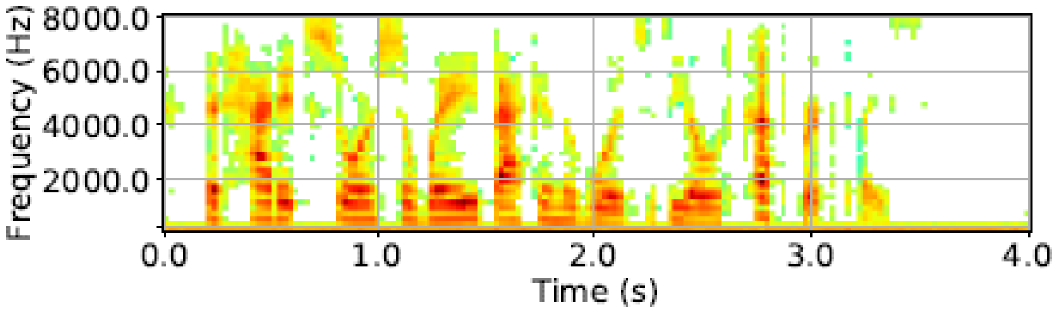
The ratio of the number of masked time-frequency : 0 %
The ratio of the masked energy : 0 %

Small Energy Masing Algorithm – Spectrogram with η_{th} of -40 dB



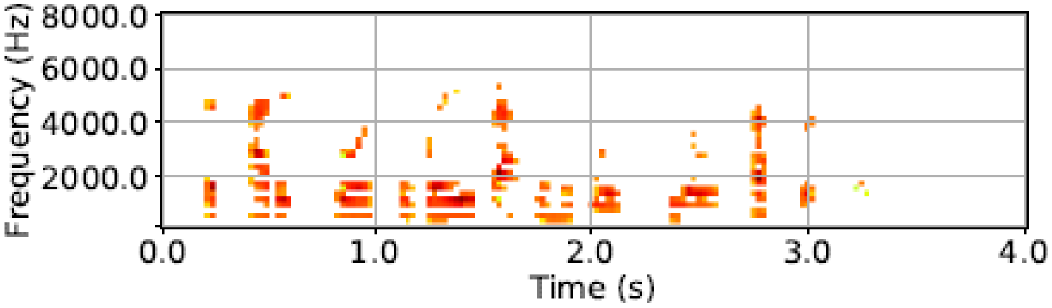
The ratio of the number of masked time-frequency bins : 38 %
The ratio of the masked energy : 25 %

Small Energy Masing Algorithm – Spectrogram with η_{th} of -20 dB



The ratio of the number of masked time-frequency bins : 75 %
The ratio of the masked energy : 62 %

Small Energy Masing Algorithm – Spectrogram with η_{th} of 0 dB

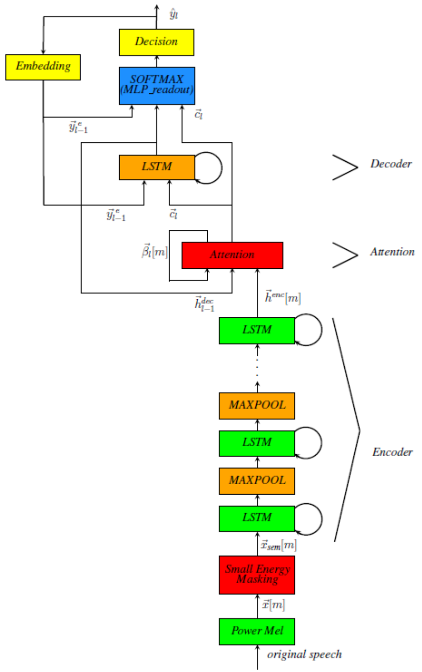


The ratio of the number of masked time-frequency bins : 95 %

The ratio of the masked energy : 88 %

Experimental Results - Speech Recognition System Structure

- The speech recognition system is based on the attention-based encoder-decoder model, modified from our previous system[C. Kim et. al., ASRU 2019].
- 6 LSTM layers in the encoder, and 1 LSTM layer in the decoder are used. The unit size is 1024.
- Pre-training strategy is employed [A. Zeyer et. al. INTERSPEECH 2018].
- Power-mel feature is employed [C. Kim et. al. INTERSPEECH 2019].



Experimental Results - Small Energy Masking: Word Error Rate (WER) dependence on η_b

$$\eta_{th} \sim \mathcal{U}(\eta_a, \eta_b)$$

- In this experiment, η_a is fixed to at -80 dB.
- Dependence on η_b is tested.
- If η_b becomes larger than 20 dB, performance starts degrading.

η_b	-60 dB	-40 dB	-20 dB	0 dB	**baseline**
test-clean	4.03 %	4.05 %	3.89 %	**3.72 %**	4.19 %
test-other	13.64 %	13.69 %	12.74 %	**11.65 %**	13.47 %
average	8.84 %	8.87 %	8.32 %	**7.69 %**	8.83 %

Experimental Results - Small Energy Masking: dependence on η_a

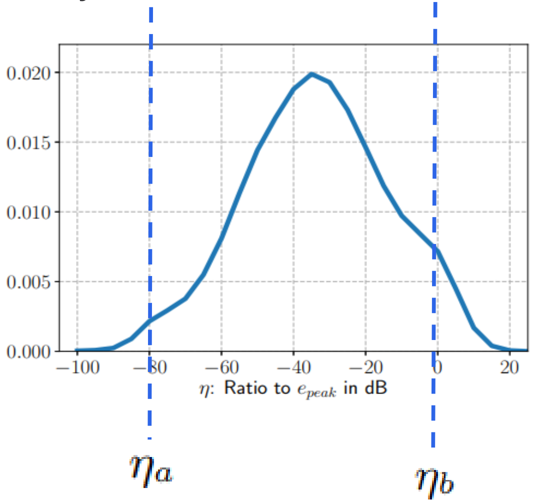
$$\eta_{th} \sim \mathcal{U}(\eta_a, \eta_b)$$

- In this experiment, η_b is fixed to at 0 dB.
- Dependence on η_a is tested.

η_a	-20 dB	-40 dB	-60 dB	-80 dB	**baseline**
test-clean	45.15 %	6.57 %	4.07 %	**3.72 %**	4.19 %
test-other	77.71 %	20.43 %	12.73 %	**11.65 %**	13.47 %
average	61.43 %	13.5 %	8.40 %	**7.69 %**	8.83 %

Experimental Results - Small Energy Masking: selection of η_a and η_b

- From the previous experiments, we observe that η_a = -80 dB η_b = 0 dB are good choices.
- From the following probability density function of η , this distribution covers the entire range.



- The relative performance improvement over the baseline is 11.2 % and 13.5 % on LibriSpeech test-clean and test-other respectively.

Experimental Results – Fixed Threshold Masking

- What happens if we use a fixed threshold (η_{th}) rather than a random threshold?

η_{th}	baseline $-\infty$ dB	-80 dB	-70 dB	-60 dB	-50 dB
test-clean	**4.19 %**	4.27 %	4.26 %	4.31 %	4.52 %
test-other	**13.47 %**	13.92 %	13.93 %	14.09 %	15.67 %
average	**8.83 %**	9.10 %	9.10 %	9.20 %	10.10 %

- As shown in the above table, fixed threshold masking always results in performance degradation.
- From this result, we may observe that the randomization of the threshold level plays a critically important role in obtaining good performance.

Experimental Results – Random input dropout

- We applied a conventional input dropout approach to the input layer with a different drop out rate *r*.

	baseline $r = 0$	$r = 0.1$	$r = 0.2$	$r = 0.3$
test-clean	4.19 %	**4.03 %**	4.29 %	4.27 %
test-other	13.47 %	**13.18 %**	13.77 %	14.59 %
average	8.83 %	**8.61 %**	9.03 %	9.43 %

- The best performance was obtained when r = 0.1. However, SEM shows 7.7 % and 11.6 % Relative WER (WERR) improvements over this random input dropout for the test-clean and test-other respectively.

Experimental Results – Modified shallow fusion with a Transformer LM.

- We used the modified shallow fusion [C. Kim, et. al., INTERSPEECH 2019] with a Transformer LM [A. Vaswani, et. al., NIPS 2017].

$$y^*_{0:L} = \arg\max_{y_{0:L}} \sum_{l=0}^{L-1} \Big[\log P\left(y_l | \boldsymbol{x}[0:M], y_{0:l}\right) - \lambda_p \log P(y_l) + \lambda_{\text{lm}} \log P\left(y_l | y_{0:l}\right) \Big]$$

λ_p λ_{lm}	0.003 0.36	0.003 0.4	0.003 0.44	0.003 0.48
test-clean	**2.52** %	2.62 %	2.62 %	2.66 %
test-other	7.93 %	**7.87** %	**7.87** %	8.33 %
average	5.23 %	**5.25** %	**5.25** %	5.50 %

- When λ_p = 0.003 and λ_{lm} = 0.4 or 0.44, 2.62 % and 7.87 % WERs are obtained for LibriSpeech test-clean and test-other sets.

Conclusions

- Motivation:
  - Regularization is important for training the neural network model.
  - Time frequency-bins with small energy may be more adversely affected by distortion or noise.
- Small Energy Masking (SEM) algorithm:
  - A random energy threshold is generated from the uniform distribution.
  - All the feature values below that threshold is masked to zero.
  - The unmasked feature values are scaled so that the sum is maintained.
- Experimental Results:
  - SEM shows 11.2 % and 13.5 % Relative WER (WERR) improvements on the standard LibriSpeech test-clean and test-other sets over the baseline.
  - SEM shows 7.7 % and 11.6 % Relative WER (WERR) improvements on the same LibriSpeech test-clean and test-other sets over the random input dropout.
  - With a modified shallow fusion with a Transformer-baesd LM, we achieved 2.62 % and 7.87 % WERs on the LibriSpeech test-clean and test-other sets.

References

[1] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014. [Online]. Available: http://jmlr.org/papers/v15/srivastava14a.html

[2] 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," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2680

[3] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. N. 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 *Proc. Interspeech 2017*, 2017, pp. 379–383. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1510

[4] C. Kim, K. Kumar and R. M. Stern, "Robust speech recognition using small power boosting algorithm," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2009, pp. 243–248.

References

[5] C. Kim, M. Kumar, K. Kim, and D. Gowda, "Power-law nonlinearity with maximally uniform distribution criterion for improved neural network training in automatic speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019, pp. 988–995.

[6] C. Kim, S. Kim, K. Kim, M. Kumar, J. Kim, K. Lee, C. Han, A. Garg, E. Kim, M. Shin, S. Singh, L. Heck, and D. Gowda, "End-to-end training of a large vocabulary end-to-end speech recognition system," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019, pp. 562–569.

[7] A. Zeyer, K. Irie, R. Schlüter, and H. Ney, "Improved training of end-to-end attention models for speech recognition," in *INTERSPEECH-2018*, 2018, pp. 7–11. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1616

[8] C. Kim, M. Shin, A. Garg, and D. Gowda, "Improved vocal tract length perturbation for a state-of-the-art end-to-end speech recognition system," in *INTERSPEECH-2019*, Graz, Austria, Sept. 2019, pp. 739–743. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3227