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 in neural network training
- Motivation of Small Energy Masking (SEM)
- Small energy masking algorithm
  - Algorithm overview
  - Structure of Small Energy Masking (SEM) algorithm
  - Spectrograms with different energy thresholds
- Experimental results
  - Word Error Rate (WER) dependence on the bounds of energy threshold
  - WER comparison between SEM, fixed threshold masking and random input dropout
  - Recognition result with a modified shallow fusion
- Conclusions
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 augmentation 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 frequently 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

- \( \eta \): The ratio of filterbank energy \( e[m, c] \) to the peak filterbank energy \( e_{peak} \) in dB:

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

- The Probability Density Function (PDF) of \( \eta \) 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.
Motivation of Small Energy Masking - Cumulative density function and energy portion below the threshold

- The cumulative function $\eta$ is shown on the right-hand side.

- We define $r_e(\eta_{th})$ as the portion of energy below the threshold $\eta_{th}$ 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 call $\eta_{th}$) for each utterance uniformly from the following interval.

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

- $\mathcal{U}$: Uniform distribution
- $\eta_a$: The lower bound. We use the value of -80 dB.
- $\eta_b$: 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.
- As a baseline system, we use the following **power-mel feature** pipeline.
  [C. Kim et. al., ASRU 2019, C. Kim et. al. INTERSPEECH 2019]
As a baseline system, we use the following **power-mel feature** pipeline.  
[C. Kim et. al., ASRU 2019, C. Kim et. al. INTERSPEECH 2019]

\[
e[m, c] = \sum_{k=0}^{K/2} |X[m, e^{jk\omega}]|^2 M_c[e^{jk\omega}]
\]
As a baseline system, we use the following **power-mel feature** pipeline.  
[C. Kim et. al., ASRU 2019, C. Kim et. al. INTERSPEECH 2019]
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**
  \[ p[m, c] = e[m, c]^{15} \]

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

- **Global Mean and Variance Normalization**
- **Power-Law Nonlinearity**
- **Filterbank Energy Calculation**

**Small Energy Masing Algorithm - Conventional Pipeline**
Small Energy Masing Algorithm - Masking Application

input feature $\tilde{x}[m] = \{x[m,c] \mid 0 \leq c \leq C - 1\}$
Small Energy Masing Algorithm - Masking Application

input feature $\tilde{x}[m] = \{x[m,c] \mid 0 \leq c \leq C - 1\}$

$e_{peak} := \text{The 95-th percentile of } e[m,c]$
Small Energy Masing Algorithm - Masking Application

\[ \eta_{th} \sim U(\eta_a, \eta_b) \]

Input feature: \( \bar{x}[m] = \{x[m,c] \mid 0 \leq c \leq C - 1 \} \)

Scaled feature: \( r \)

Global Mean and Variance Normalization

Power-Law Nonlinearity

Filterbank Energy Calculation

Peak Filterbank Energy Calculation

Filterbank Energy Calculation

Uniform Random Generation

Filterbank Energy Threshold Calculation

Binary Mask Construction

Scaling Up to Keep the Sum

\( \eta_{th} \)

\( \varepsilon_{th} \)

\( \varepsilon_{peak} \)
Small Energy Masing Algorithm - Masking Application

\[ e_{th} = e_{peak} 10^{\eta_{th}} \]

- Uniform Random Generation
- Energy Threshold Calculation
- Peak Filterbank Energy Calculation
- Power-Law Nonlinearity
- Filterbank Energy Calculation

Input feature: \( \bar{x}[m] = \{x[m,c] | 0 \leq c \leq C - 1 \} \)

Masked feature: \( r \)

Global Mean and Variance Normalization:

\[ \mu[m,c] \]
Small Energy Masing Algorithm - Masking Application

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

input feature \( \bar{x}[m] = \{x[m,c] \mid 0 \leq c \leq C - 1 \} \)

masked feature \( r \)

Global Mean and Variance Normalization

Power-Law Nonlinearity

Filterbank Energy Calculation

Peak Filterbank Energy Calculation

Filterbank Energy Threshold Calculation

Binary Mask Construction

Scaling Up to Keep the Sum

Uniform Random Generation

\( \eta_{th} \)

\( e_{th} \)

\( \tilde{e}[m] \)

speech
Small Energy Masing Algorithm - Masking Application

- Uniform Random Generation
- Filterbank Energy Threshold Calculation
- Peak Filterbank Energy Calculation
- Global Mean and Variance Normalization
- Power-Law Nonlinearity
- Filterbank Energy Calculation
- Scaling Up to Keep the Sum
- Binary Mask Construction

$x_{\mu}[m, c] = x[m, c]\mu[m, c]$

input feature $\bar{x}[m] = \{x[m, c] \mid 0 \leq c \leq C - 1\}$

masked feature
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

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

input feature: \( \bar{x}[m] = \{x[m, c] \mid 0 \leq c \leq C - 1 \} \)
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 $\eta_{th}$ 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 $\eta_{th}$ 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 $\eta_{th}$ of 0 dB

The ratio of the number of masked time-frequency bins: 95 %
The ratio of the masked energy: 88 %

---

Samsung Research
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 $\eta_b$

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

- In this experiment, $\eta_a$ is fixed to at -80 dB.
- Dependence on $\eta_b$ is tested.
- If $\eta_b$ becomes larger than 20 dB, performance starts degrading.

<table>
<thead>
<tr>
<th>$\eta_b$</th>
<th>-60 dB</th>
<th>-40 dB</th>
<th>-20 dB</th>
<th>0 dB</th>
<th>baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>test-clean</td>
<td>4.03%</td>
<td>4.05%</td>
<td>3.89%</td>
<td>3.72%</td>
<td>4.19%</td>
</tr>
<tr>
<td>test-other</td>
<td>13.64%</td>
<td>13.69%</td>
<td>12.74%</td>
<td>11.65%</td>
<td>13.47%</td>
</tr>
<tr>
<td>average</td>
<td>8.84%</td>
<td>8.87%</td>
<td>8.32%</td>
<td>7.69%</td>
<td>8.83%</td>
</tr>
</tbody>
</table>
Experimental Results - Small Energy Masking: dependence on $\eta_a$

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

- In this experiment, $\eta_b$ is fixed to at 0 dB.
- Dependence on $\eta_a$ is tested.

<table>
<thead>
<tr>
<th>$\eta_a$</th>
<th>-20 dB</th>
<th>-40 dB</th>
<th>-60 dB</th>
<th>-80 dB</th>
<th>baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>test-clean</td>
<td>45.15 %</td>
<td>6.57 %</td>
<td>4.07 %</td>
<td>3.72 %</td>
<td>4.19 %</td>
</tr>
<tr>
<td>test-other</td>
<td>77.71 %</td>
<td>20.43 %</td>
<td>12.73 %</td>
<td>11.65 %</td>
<td>13.47 %</td>
</tr>
<tr>
<td>average</td>
<td>61.43 %</td>
<td>13.5 %</td>
<td>8.40 %</td>
<td>7.69 %</td>
<td>8.83 %</td>
</tr>
</tbody>
</table>
Experimental Results - Small Energy Masking: selection of $\eta_a$ and $\eta_b$

- From the previous experiments, we observe that $\eta_a = -80$ dB and $\eta_b = 0$ dB are good choices.
- From the following probability density function of $\eta$, this distribution covers the entire range.

![Probability density function graph](image)

- 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 ($\eta_{th}$) rather than a random threshold?

<table>
<thead>
<tr>
<th>$\eta_{th}$</th>
<th>baseline $-\infty$ dB</th>
<th>-80 dB</th>
<th>-70 dB</th>
<th>-60 dB</th>
<th>-50 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>test-clean</td>
<td>4.19 %</td>
<td>4.27 %</td>
<td>4.26 %</td>
<td>4.31 %</td>
<td>4.52 %</td>
</tr>
<tr>
<td>test-other</td>
<td>13.47 %</td>
<td>13.92 %</td>
<td>13.93 %</td>
<td>14.09 %</td>
<td>15.67 %</td>
</tr>
<tr>
<td>average</td>
<td>8.83 %</td>
<td>9.10 %</td>
<td>9.10 %</td>
<td>9.20 %</td>
<td>10.10 %</td>
</tr>
</tbody>
</table>

- 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 $r$.  

<table>
<thead>
<tr>
<th></th>
<th>baseline $r = 0$</th>
<th>$r = 0.1$</th>
<th>$r = 0.2$</th>
<th>$r = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>test-clean</td>
<td>4.19 %</td>
<td><strong>4.03 %</strong></td>
<td>4.29 %</td>
<td>4.27 %</td>
</tr>
<tr>
<td>test-other</td>
<td>13.47 %</td>
<td><strong>13.18 %</strong></td>
<td>13.77 %</td>
<td>14.59 %</td>
</tr>
<tr>
<td>average</td>
<td>8.83 %</td>
<td><strong>8.61 %</strong></td>
<td>9.03 %</td>
<td>9.43 %</td>
</tr>
</tbody>
</table>

- 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].

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

<table>
<thead>
<tr>
<th>$\lambda_p$</th>
<th>$\lambda_{lm}$</th>
<th>test-clean</th>
<th>test-other</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.003</td>
<td>0.36</td>
<td>2.52 %</td>
<td>7.93 %</td>
<td>5.23 %</td>
</tr>
<tr>
<td>0.003</td>
<td>0.40</td>
<td>2.62 %</td>
<td>7.87 %</td>
<td>5.25 %</td>
</tr>
<tr>
<td>0.003</td>
<td>0.44</td>
<td>2.62 %</td>
<td>7.87 %</td>
<td>5.25 %</td>
</tr>
<tr>
<td>0.003</td>
<td>0.48</td>
<td>2.66 %</td>
<td>8.33 %</td>
<td>5.50 %</td>
</tr>
</tbody>
</table>

- When $\lambda_p = 0.003$ and $\lambda_{lm} = 0.4$ 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


References


