SpeechMix - Augmenting Deep Sound Recognition using Hidden Space Interpolations

Amit Jindal, Narayanan Elavathur Ranganatha, Aniket Didolkar, Arijit Ghosh Chowdhury, Di Jin, Ramit Sawhney, Rajiv Ratn Shah

INTERSPEECH 2020

1. INTRODUCTION

- Problem
  - Deep neural networks tend to contain millions to billions of parameters.
  - Thus prone to overfitting, due to a lack of sufficient train data

- Solutions
  - Data augmentation
  - Regularization and so on … (generalization)

2. Related Work

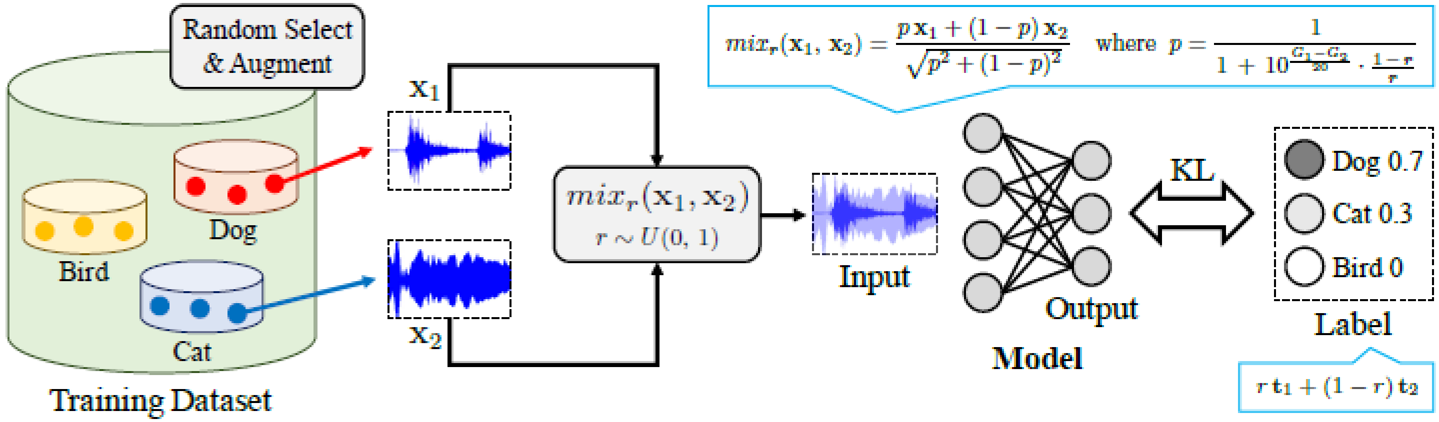
- Deep speech recognition (models)
  - [Piczak, 2015] : apply CNNs to the log-mel features extracted from raw waveforms
  - [Aytar, 2016] : 1D convolutional and pooling layers named *SoundNet*
  - [Harada, 2017] : 1D and 2D convolutional and pooling layer named *EnvNet*
  - [Tokozume, 2017] : *EnvNet-v2* → a higher number of layers and a higher sampling rate

- Data augmentation for speech
  - Cropping
  - Time stretching, pitch shifting, dynamic range compression, and adding background noise chosen from an external dataset
  - [Park, 2019] : SpecAugment → warping features, masking blocks of time steps and frequency channels
  - [Peddinti, 2015] Audio signal speed alteration
  - [Peddiniti, 2015] Artificial reverberation into the records

→ Quality, the noise than phonetic or acoustic imformation ..

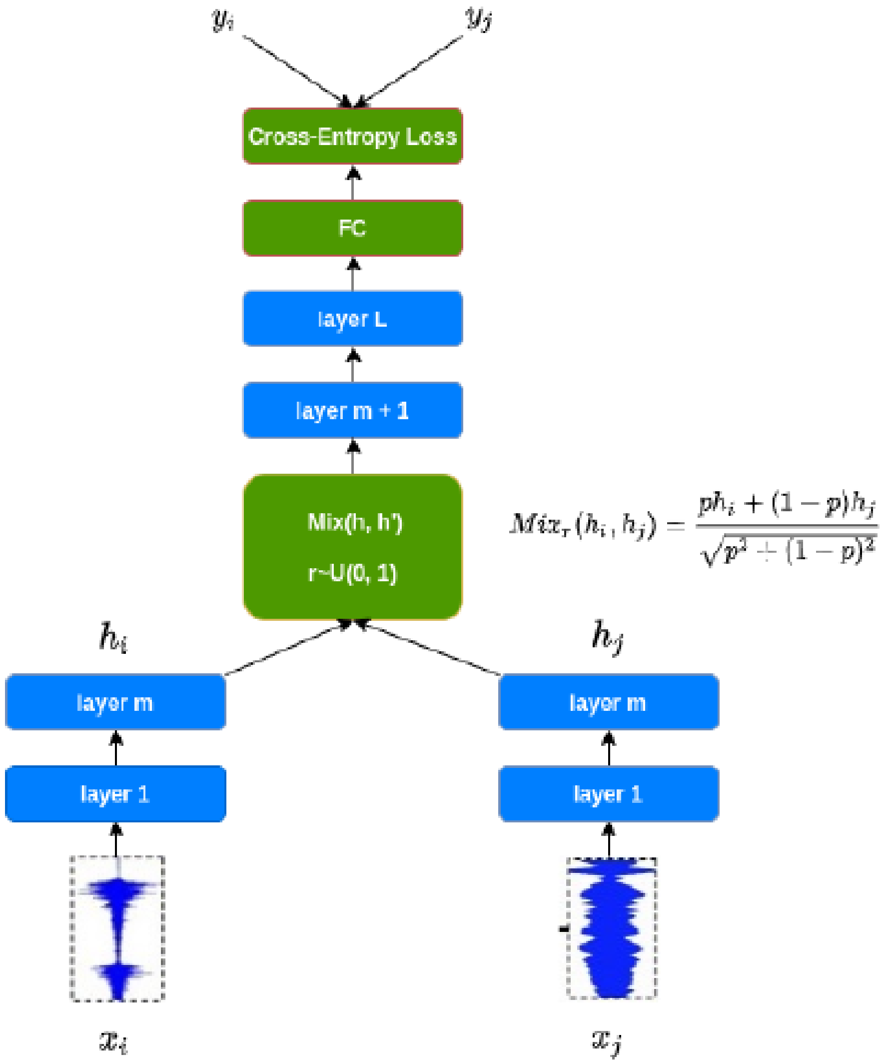
2. Related Work

- Interpolation based regularizers
  - [Zhang, 2017] **Mixup**
    - Data-agnostic augmentation technique that constructs virtual training examples by interpolating pairs of training samples from its vicinal distribution
  - [Tokozume, 2017] **Between-Class Learning (BC learning)**
    - Mix the input signals by taking auditory perception of sound into account to generate virtual samples
    - Using the *EnvNet-v2*



2. Methodology

- **SpeechMix**
  - In BC learning, mixup occurs in training examples before they are sent as input to the model
  - <u>SpeechMix augments BC learning by employing mixup of hidden states</u>



3. Methodology

- Preliminary
  - Mixup algorithm → creates virtual training samples by linear interpolation
    - Two data points, one hot representation of the label, mix ratio

$$\tilde{x} = mix(x_i, x_j) = rx_i + (1-r)x_j \qquad (1)$$

$$\tilde{y} = mix(y_i, y_j) = ry_i + (1-r)y_j \qquad (2)$$

- Replacement → since it takes into account the relationship between energy and amplitude

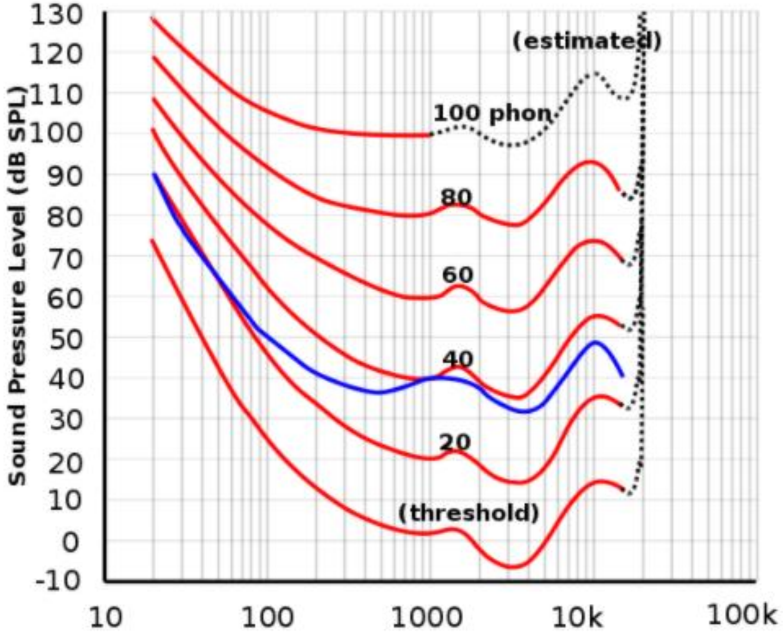
$$\tilde{x} = mix(x_i, x_j) = \frac{rx_i + (1-r)x_j}{\sqrt{r^2 + (1-r)^2}} \qquad (3)$$

3. Methodology

- Preliminary
  - BC learning algorithm → the Mixup formula used in BC learning is derived by taking auditory perceptions of sound into account
    - Transformation of mixing ratio
    - G scales are the sound pressure level of input [dB] → A-weighting (based on equal loudness contours)

$$mix(x_i, x_j) = \frac{px_i + (1-p)x_j}{\sqrt{p^2 + (1-p)^2}} \tag{4}$$

$$where\ p = \frac{1}{1 + 10^{\frac{G_i - G_j}{20} \cdot \frac{1-r}{r}}}$$



3. Methodology

- SpeechMix
  - Where the neural network is trained on interpolations of the hidden states
    - Hidden representations

$$h_l^i = g_l(h_{l-1}^i, \theta), l \in [1, m] \qquad (5)$$

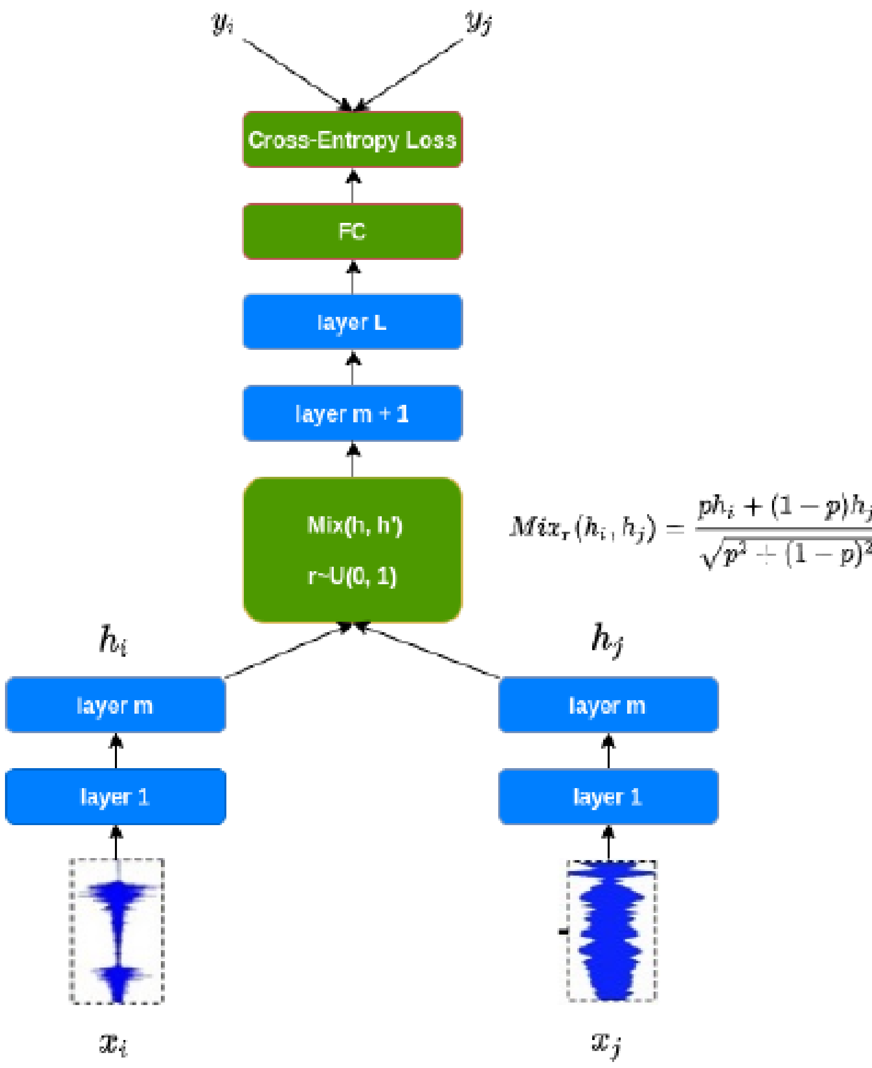
$$h_l^j = g_l(h_{l-1}^j, \theta), l \in [1, m] \qquad (6)$$

    - Mixed representation

$$\tilde{h}_m = \frac{p h_m^i + (1-p) h_m^j}{\sqrt{p^2 + (1-p)^2}} \qquad (7)$$

    - The continued forward pass after mixed hidden representation

$$\tilde{h}_l = g_l(\tilde{h}_{l-1}, \theta), l \in [m+1, M] \qquad (8)$$



3. Methodology

- Optimization
  - *n* : the number of samples in a mini-batch
  - *r* : mixing ratio
  - *m* : the layer at which miup occurs and S as the set of layers

→ For each mini-batch, *m* is sampled randomly from *S*

  - Minimize the KL-divergence between the mixed label and softmax of the generated outputs

$$L = \frac{1}{n}\sum_{i=0}^{n} D_{KL}(\tilde{y}^i || softmax(\tilde{h}_M^i)) \qquad (9)$$

$$where$$

$$D_{KL}(\tilde{y}^i || softmax(\tilde{h}_M^i)) = \sum_{k=0}^{c} \tilde{y}_k^i log \frac{\tilde{y}_k^i}{\{softmax(\tilde{h}_M^i)\}_k}$$

4. Experiments

- Dataset and Preprocessing
  - Sound event dataset

Dataset	Classes	Samples	Duration
UrbanSound8k	10	8732	9.7 hours
ESC-50	50	2000	2.8 hours
ESC-10	10	400	33 min

Table 2: *Statistics of sound classification datasets.*

  - Preprocessing
    - Padding → T/2 seconds of zero on each side,
    - Cropping → T second section is randomly cropped from padded sound * 10 → 10 crops
    - Regularization → a range of -1 to 1, dividing by 32,768

4. Experiments

- Experimental settings
  - Nestrov's accelerated gradient using momentum of 0.9
  - Weight decay of 0.0005
  - Mini-batch size of 64
  - 5-fold cross validation on ESC-10 and ESC-50
  - Scale augmentation [0.8, 1.25] → before zero padding
  - Gain augmentation [-6dB, +6dB] → before inputting to the network

- Result → In the paper !