LOSS FUNCTION WEIGHTING BASED ON SOURCE DOMINANCE FOR MONAURAL SOURCE SEPARATION USING RECURRENT NEURAL NETWORKS

Seungtae Kang and Gil-Jin Jang

School of Electronics Engineering, Kyungpook National University
80 Daehak-ro, Daegu 41566, South Korea
cdef3456@naver.com, gjang@knu.ac.kr

ABSTRACT

In this paper, we propose a weighted loss function for monaural source separation using recurrent neural networks where appropriate training data for the original sources are available. The weight varies for each time-frequency instance according to the mutual dominance of the binaural source signals. The mutual dominance is computed by the multiplications of the inverse source-to-mixture ratios of the ground truth signals of the two sources, and the weights are obtained by appropriate scaling of the mutual dominance. The squared error between the target and the estimated becomes more important as the difference becomes larger. The proposed weighting is applied to one of the conventional monaural source separation techniques that exploits recurrent neural networks, and showed improved performances over the same dataset.

1. MONAURAL SOURCE SEPARATION USING RECURRENT NEURAL NETWORKS

Monaural source separation has been regarded as one of the most difficult tasks among many different kinds of signal processing applications. However, with the help of recent advancement in recurrent neural networks (RNNs), many promising results have been made to the problem [1, 2]. The problem that we are dealing with is not blind, because some amount of labeled training data that reflects the characteristics of target source is required to train the RNN. We propose a weighted loss function that varies the weight on each time-frequency instance according to the difference in power spectral density. The baseline source separation network architecture adopted in this paper is based on [2].

2. LOSS FUNCTION WEIGHTING

One of the most general loss function between the given ground truth and the approximate signals is the mean-squared error (MSE) over all the time and frequency units in the power spectral domain. More specifically, sum of the error between $\mathbf{y}_1$ and $\tilde{\mathbf{y}}_1$, and between $\mathbf{y}_2$ and $\tilde{\mathbf{y}}_2$, where $\mathbf{y}_1$ and $\mathbf{y}_2$ are the short-time Fourier transforms (STFTs) of the time domain input, which can be computed as following formula:

$$J_{MSE} = \frac{1}{TF} \sum_{t=1}^{T} \sum_{f=1}^{F} \Delta(t,f)\,, \tag{1}$$

where T and F are the numbers of time frames and frequency bins, respectively, and $\Delta(t,f)$ is the sum of the difference between ground truth and network outputs given as:

$$\Delta(t,f) = |y_{1,t}(f) - \tilde{y}_{1,t}(f)|^2 + |y_{2,t}(f) - \tilde{y}_{2,t}(f)|^2\,. \tag{2}$$

The loss function in Equation 1 applies the same amount of importance for all of the time-frequency units. However, imposing different weight to each unit based on their importance in describing individual source signals can lead to better learning of the networks.

Our assumption is that if one signal is dominant to the other when they are mixed, it is more efficient in describing the corresponding source. The dominance of source 1 is defined by the ratio of source 1 to the sum of source 1 and 2 in the power spectral domain as follows:

$$\gamma_i(t,f) = \frac{|y_{i,t}(f)|}{|y_{1,t}(f) + y_{2,t}(f)|}\,, \quad i = \{1,2\}\,. \tag{3}$$

The relationship between the dominance for source 1 and source 2 is, $\gamma_2 \simeq 1 - \gamma_1$. If we multiply the two dominance factors to consider both sources, $\gamma_1\gamma_2 \simeq \gamma_1(1-\gamma_1)$, is maximum at $\gamma_1 = \gamma_2 = 0.5$. However, our assumption is that the importance should be proportional to the dominance, so we define a mutual dominance factor by the multiplication of the inverse of the individual dominance values:

$$\begin{aligned}\gamma_{mutual}(t,f) &= \frac{1}{\gamma_1(t,f)} \frac{1}{\gamma_2(t,f)} \\ &= \frac{|y_{1,t}(f) + y_{2,t}(f)|^2}{|y_{1,t}(f)||y_{2,t}(f)|}\,.\end{aligned} \tag{4}$$

The mutual dominance factor approaches ∞ very rapidly as either source goes to 0, that is, either γ_1 or γ_2 becomes 0. To make use of the mutual dominance as weight factors to the loss function in Equation 1, we apply $\log$ function to

Corresponding Author is Gil-Jin Jang. This work was supported by the National Research Foundation of Korea (NRF) grant (No. NRF-2017M3C1B6071400, 50%) and by Institute for Information and communications Technology Promotion (IITP) grant (No.2017-0-00053, A Technology Development of Artificial Intelligence Doctors for Cardiovascular Disease, 50%). Both grants are funded by the Korea government (MSIP).

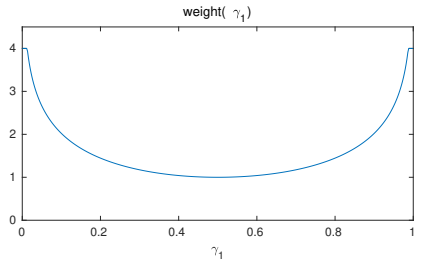

Figure 1: Loss function weight with respect to γ_1. The behavior near 0 and 1 becomes less steep when compared to Equation 4, and it is linearly scaled so that minimum and maximum weights are 1.0 and 4.0, respectively.

γ_{mutual} in Equation 4 to make it not too steep near 0 and 1, and scale it appropriatly as follows:

$$w(t,f) = g\left(\log \gamma_{mutual}(t,f)\right) , \tag{5}$$

where the scale function $g(\cdot)$ rectifies the input at 90 percentile and scales the value linearly in $[w_{\min}\ w_{\max}]$. The statistics of $\log \gamma_{mutual}(t,f)$ is obtained by the training data of the source separation RNN.

Figure 1 shows the weight values with respect to source 1 dominance γ_1. The weight becomes roughly two times when $\gamma_1 = 0.9$ or 0.1, and maximum when $\gamma_1 > 0.99$ or $\gamma_1 < 0.01$. We apply the weight in Equation 5 to the loss function Equation 1 to obtain a new weighted loss function:

$$J_{wMSE} = \frac{1}{TF}\sum_{t=1}^{T}\sum_{f=1}^{F} w(t,f)\Delta(t,f)\,. \tag{6}$$

3. EXPERIMENTAL RESULTS

To show the effectiveness of the proposed method, we carried out source separation experiments on MIR-1K dataset [3]. It is composed of 1,000 sound clips from 110 songs by 19 Chinese amateur singers. The sound files are stereo recordings, with left channels music accompaniment and right channels singing voice. We used one male and one female ('abjones' and 'amy') for training and development set, and recordings of the remaining 17 singers as test set. The performance was evaluated by GNSDR (global normalized source to distortion ratio), GSIR (global source to interference ratio), and GSAR (global source to artifact ratio) measures in BSS-EVAL 3.0 metrics [4]. To construct input to the RNNs, we generated sum of each pair of stereo sound files, $y_{1+2}[n] = y_1[n] + y_2[n]$, applied 1024-point STFT with 25% overlap, extracted magnitude spectrum $\mathbf{y}_t$, and concatenated the previous and the next frames:

$$\mathbf{x}_t = [\mathbf{y}_{t-1}\ \mathbf{y}_t\ \mathbf{y}_{t+1}]\ . \tag{7}$$

The separation network is a stacked RNN with 3 layers, and the number of hidden nodes in each layer was 1024.

Metric	baseline	$w_{\max}$ 5.0	10.0	15.0	20.0	25.0
$GNSDR_m$	5.7	5.8	5.9	5.9	5.8	5.7
$GSIR_m$	12.6	12.4	12.2	12.4	11.8	12.2
$GSAR_m$	7.3	7.5	7.6	7.6	7.7	7.5
$GNSDR_v$	5.5	5.6	5.7	5.7	5.6	5.6
$GSIR_v$	12.1	12.4	12.8	12.6	12.8	12.5
$GSAR_v$	7.4	7.4	7.4	7.4	7.3	7.3
Average	8.43	8.52	8.60	8.60	8.50	8.47

Table 1: Separation performance comparison in terms of BSS-EVAL 3.0 metrics.

Training parameters are: batch size 128, Adam optimizer with learning late 10^{-4}, and the number of learning steps 30,000. The evaluation results are shown in Table 1. First column lists the used metrics. $GNSDR_m$ and $GNSDR_v$ are the measured GNSDRs for music and voice sources, respectively. The other two metrics, GSIR and GSAR, are computed for both music and voice sources as well. 'baseline' is the RNN with the unweighted loss function in Equation 1, and the last 5 columns are from RNNs with the proposed weighted loss function in Equation 6, with the maximum weights ($w_{\max}$) vary from 5.0 to 25.0. The minimum weights ($w_{\min}$) are all fixed to 1.0. The proposed method outperformed the baseline except $GSIR_m$, by 0.04 to 0.17 improvements on the average over the baseline method.

4. REFERENCES

[1] P.-S. Huang, M. Kim, and M. Hasegawa-Johnson, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 2136–2147, 12 2015.

[2] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 1562–1566, 2014.

[3] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2010.

[4] E. Vincent and C. F. Rémi Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462–1469, 12 2006.