深層学習に基づく音声合成における 2次統計量を用いたスペクトル特徴量のモデリングの検討*

◎松永悟行，大谷大和（(株）エーアイ），平原達也（富山県立大）

1　はじめに

近年の音声合成分野においては，Deep Neural Network（DNN）を利用した音声波形のモデル化 [1]や end-to-end の合成音声[2]などの研究が盛んにおこなわれている。一方で，音声合成システムとして DNN の計算量，パラメータの操作性，メンテナンス性を考慮すると，テキスト解析部，音声特徴量生成部，音声波形生成部という機能ごとに分離した構成を採用することも少なくない。この構成の場合, DNN はテキスト解析部で生成された言語特徴量を入力して音声特徴量を予測する部分に用いる。また，DNN の計算量の削減のために，構造が単純な Feed-Forward Neural Network（FFNN）を採用する。しかし，FFNN を用いる場合は，時系列である音声特徴量の隣接するフレームとの関係を無視して学習が行われる問題を解決する必要がある。

そこで本報告では，合成音声の品質のうち音質に関わるスペクトル特徴量を対象として，FFNN でも隣接するフレームの関係を学習できるようにするために，生成されたスペクトル特徴量から 2 次統計量を計算してその誤差を最小化する学習方法を提案する[3]。

2　2 次統計量を考慮した損失関数

$\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_T^\top]^\top$は言語特徴量系列, $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_t^\top, \cdots, \boldsymbol{y}_T^\top]^\top$は自然音声特徴量系列，$\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$は生成された音声特徴量系列とする。ここで，$t$と$T$はそれぞれフレームのインデックスと総数であり，$^\top$は転置記号である。$\boldsymbol{x}_t = [x_t^{(1)}, \cdots, x_t^{(i)}, \cdots, x_t^{(I)}]$ と $\boldsymbol{y}_t = [y_t^{(1)}, \cdots, y_t^{(d)}, \cdots, y_t^{(D)}]$はそれぞれフレーム t における言語特徴量ベクトルと音声特徴量ベクトルである。ここで，iとIおよびdとDはそれぞれ言語特徴量ベクトルおよび音声特徴量ベクトルの次元のインデックスと総数である。

提案する損失関数では，隣接するフレームとの関係性を考慮するため，$\boldsymbol{x}$と$\boldsymbol{y}$を短期の閉区間 $[t+L, t+R]$ で区切った系列$\mathbf{X}$ と $\boldsymbol{Y} = [\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_t, \cdots, \boldsymbol{Y}_T]$ を利用する。ここで，$\boldsymbol{Y}_t = [\boldsymbol{y}_{t+L}^\top, \cdots, \boldsymbol{y}_{t+\tau}^\top, \cdots, \boldsymbol{y}_{t+R}^\top]^\top$はフレーム$t$についての短期の系列であり，$L \leq 0$は後方参照するフレーム数，$R \geq 0$は前方参照するフレーム数であり，$\tau$ $(L \leq \tau \leq R)$は短期内の参照フレームインデックスである。FFNN では $\boldsymbol{x}_{t+\tau}$に対する $\hat{\boldsymbol{y}}_{t+\tau}$は隣接フレームとは関係なく独立して予測される。そこで，$\boldsymbol{Y}_t$に対して隣接するフレーム同士を関連付けるために局所内の分散（LV）[4]や局所内の分散共分散行列（LC）の損失関数を導入する。これらの損失関数の効果は$\boldsymbol{Y}_t$と$\boldsymbol{Y}_{t+\tau}$ がオーバーラップの関係となっているため，学習段階ですべてのフレームに波及する。また，提案する損失関数は明示的に定義した短期の関係が暗黙的に長期の関係に波及する設計となっているが，系列内の分散（GV）[5]や系列内の分散共分散行列（GC）の損失関数を導入することで長期の関係を明示的に定義することも可能である。さらに，多次元の音声特徴量については次元領域（DD）の損失関数を導入することで次元間の関係を考慮することが可能となる。提案する損失関数$\mathcal{L}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$はこれらの損失関数とベースラインとして一般的な$\boldsymbol{y}$と$\hat{\boldsymbol{y}}$の平均二乗誤差（BL）の損失関数の出力の重み付き和により定義される。

$$\mathcal{L}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_i \omega_i \mathcal{L}_i(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) \tag{1}$$

ここで，$i = \{\mathrm{BL}, \mathrm{LV}, \mathrm{LC}, \mathrm{GV}, \mathrm{GC}, \mathrm{DD}\}$は損失関数の識別子を表し，$\omega_i$は識別子$i$の損失に対する重みである。

2.1 局所内の分散

$\boldsymbol{Y}_{\mathrm{LV}} = [\boldsymbol{v}_1^\top, \cdots, \boldsymbol{v}_t^\top, \cdots, \boldsymbol{v}_T^\top]^\top$は閉区間$[t+L, t+R]$における分散ベクトルの系列であり，局所内の分散の損失関数$\mathcal{L}_{\mathrm{LV}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$は$\boldsymbol{Y}_{\mathrm{LV}}$と$\hat{\boldsymbol{Y}}_{\mathrm{LV}}$の平均絶対誤差で定義される。

$$\mathcal{L}_{\mathrm{LV}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{TD} \sum_{t=1}^{T} \sum_{d=1}^{D} \left| \boldsymbol{Y}_{\mathrm{LV}} - \hat{\boldsymbol{Y}}_{\mathrm{LV}} \right| \tag{2}$$

ここで，$\boldsymbol{v}_t = [v_t^{(1)}, \cdots, v_t^{(d)}, \cdots, v_t^{(D)}]$はフレーム$t$における$D$次元の分散ベクトルであり，次元$d$の分散$v_t^{(d)}$は以下により与えられる。

$$v_t^{(d)} = \frac{1}{-L+R+1} \sum_{\tau=L}^{R} \left(y_{t+\tau}^{(d)} - \bar{y}_t^{(d)} \right)^2 \tag{3}$$

ここで，$\bar{y}_t^{(d)}$は閉区間$[t+L, t+R]$における次元dの平均である。

$$\bar{y}_t^{(d)} = \frac{1}{-L+R+1}\sum_{\tau=L}^{R} y_{t+\tau}^{(d)} \tag{4}$$

この損失関数は文献[4]のように音素などの短区間における系列の変動を補償することを目的に設計した。

2.2 局所内の分散共分散行列

$\boldsymbol{Y}_{\mathrm{LC}} = [\boldsymbol{c}_1, \cdots, \boldsymbol{c}_t, \cdots, \boldsymbol{c}_T]$は閉区間$[t+L, t+R]$における分散共分散行列の系列であり，局所内の分散共分散行列の損失関数$\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$は$\boldsymbol{Y}_{\mathrm{LC}}$と$\hat{\boldsymbol{Y}}_{\mathrm{LC}}$の平均絶対誤差で定義される。

$$\mathcal{L}_{\mathrm{LC}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{TD^2}\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{d=1}^{D}\left|\boldsymbol{Y}_{\mathrm{LC}} - \hat{\boldsymbol{Y}}_{\mathrm{LC}}\right| \tag{5}$$

ここで，$\boldsymbol{c}_t$は$D \times D$の分散共分散行列である。

$$\boldsymbol{c}_t = \frac{1}{-L+R+1}(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}}_t)^{\top}(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}}_t) \tag{6}$$

$\bar{\boldsymbol{Y}}_t = [\bar{y}_t^{(1)}, \cdots, \bar{y}_t^{(d)}, \cdots, \bar{y}_t^{(D)}]$は閉区間$[t+L, t+R]$における平均ベクトルである。

この損失関数は文献[6, 7]のような出力系列の分布を補償する動作を音素などの短区間で模擬することを目的として設計した。

2.3 系列内の分散

$\boldsymbol{Y}_{\mathrm{GV}} = [V^{(1)}, \cdots, V^{(d)}, \cdots, V^{(D)}]$ は $\boldsymbol{y} = \boldsymbol{Y}|_{\tau=0}$ についての分散ベクトルであり，系列内の分散の損失関数$\mathcal{L}_{\mathrm{GV}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$は$\boldsymbol{Y}_{\mathrm{GV}}$と$\hat{\boldsymbol{Y}}_{\mathrm{GV}}$の平均絶対誤差で定義される。

$$\mathcal{L}_{\mathrm{GV}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{D}\left|\boldsymbol{Y}_{\mathrm{GV}} - \hat{\boldsymbol{Y}}_{\mathrm{GV}}\right| \tag{7}$$

ここで，$V^{(d)}$は次元dの分散であり以下で与えられる。

$$V^{(d)} = \frac{1}{T}\sum_{t=1}^{T}\left(y_t^{(d)} - \bar{y}^{(d)}\right)^2 \tag{8}$$

ここで，$\bar{y}^{(d)}$は次元dの平均であり以下で与えられる。

$$\bar{y}^{(d)} = \frac{1}{T}\sum_{t=1}^{T} y_t^{(d)} \tag{9}$$

この損失関数は文献[5]のように系列全体の変動を補償することを目的に設計した。

2.4 系列内の分散共分散行列

$\boldsymbol{Y}_{\mathrm{GC}}$は$\boldsymbol{y} = \boldsymbol{Y}|_{\tau=0}$についての$D \times D$の分散共分散行列であり，系列内の分散共分散行列の損失関数$\mathcal{L}_{\mathrm{GC}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$は$\boldsymbol{Y}_{\mathrm{GC}}$と$\hat{\boldsymbol{Y}}_{\mathrm{GC}}$の平均絶対誤差で定義される。

$$\mathcal{L}_{\mathrm{GC}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{D^2}\sum_{d=1}^{D}\sum_{d=1}^{D}\left|\boldsymbol{Y}_{\mathrm{GC}} - \hat{\boldsymbol{Y}}_{\mathrm{GC}}\right| \tag{10}$$

ここで，$\boldsymbol{Y}_{\mathrm{GC}}$は次式で与えられる。

$$\boldsymbol{Y}_{\mathrm{GC}} = \frac{1}{T}(\boldsymbol{y} - \bar{\boldsymbol{y}})^{\top}(\boldsymbol{y} - \bar{\boldsymbol{y}}) \tag{11}$$

ここで，$\bar{\boldsymbol{y}} = [\bar{y}^{(1)}, \cdots, \bar{y}^{(d)}, \cdots, \bar{y}^{(D)}]$は$D$次元の平均ベクトルである。

この損失関数は文献[6, 7]のような出力系列の分布を補償する動作を模擬することを目的に設計した。

2.5 次元領域制約

$\boldsymbol{Y}_{\mathrm{DD}} = \boldsymbol{y}\boldsymbol{W}$は次元間の関係を表す特徴量の系列であり，次元領域制約の損失関数$\mathcal{L}_{\mathrm{DD}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$は$\boldsymbol{Y}_{\mathrm{DD}}$と$\hat{\boldsymbol{Y}}_{\mathrm{DD}}$の平均二乗誤差で定義される。

$$\mathcal{L}_{\mathrm{DD}}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \frac{1}{TN}\sum_{t=1}^{T}\sum_{n=1}^{N}\left(\boldsymbol{Y}_{\mathrm{DD}} - \hat{\boldsymbol{Y}}_{\mathrm{DD}}\right)^2 \tag{12}$$

ここで，$\boldsymbol{W} = [\boldsymbol{W}_1^{\top}, \cdots, \boldsymbol{W}_n^{\top}, \cdots, \boldsymbol{W}_N^{\top}]$は次元間を関連付けるための係数行列であり，$\boldsymbol{W}_n = [w^{(1)}, \cdots, w^{(d)}, \cdots, w^{(D)}]$は$n$番目の係数ベクトルであり，$n$と$N$はそれぞれ係数ベクトルのインデックスと総数である。本報告では，スペクトル特徴量としてメルケプストラムを用いるため，メルケプストラムからケプストラムへ周波数変換するときの式を行列に展開したものを$\boldsymbol{W}$とする[8]。

3 実験条件

本報告では，東京方言のプロの女性話者一名の ATR503 文の文章を含む音声コーパスを使用した。コーパスは熟練したラベラーにより手動でラベリングした。音声は平静音声で，学習用には 2000 発話，評価用には学習用とは別に 100 発話を用意した。

言語特徴量は HTS に準拠した 527 次元のベクトル系列であり，外れ値が発生しないように発話内の正規化手法により正規化した[9]。スペクトル特徴量は 60 次元のメルケプストラム系列（$\alpha = 0.55$）である。メルケプストラムは 16 bit，48 kHz でサンプリングした収録音声から WORLD [10]により 5 ms のフレーム周期で抽出したスペクトルから求めた。また，無音区間は学習から除外し，学習セット全体から平均と分散を求めて標準化した。

DNNは，ノード数を512，ReLU [11]を活性化関数とする4層の隠れ層と，線形活性化関数の出力層で構成されるFFNNとした。最適化手法はAdam（学習率: 0.001，β1: 0.9，β2: 0.999，微小量: 1e-7，学習率減衰: 0.0）[12]とした。学習のエポックは20，バッチサイズは1発話単位として，ランダムに学習データを選択する手法を用いた。平均二乗誤差を従来の損失関数とし，提案法である2次統計量を考慮した損失関数と比較した。提案した損失関数の各パラメータは予備実験の結果より $L=-2, R=2, \omega_{\mathrm{BL}}=1, \omega_{\mathrm{GV}}=1, \omega_{\mathrm{GC}}=0, \omega_{\mathrm{LV}}=3,\ \omega_{\mathrm{LC}}=3, \omega_{\mathrm{DD}}=1$とした。

4　実験結果

Fig. 1〜3は評価セットから選んだ1発話の代表例である。Fig. 1は5次と10次のメルケプストラム系列，Fig. 2は5次と10次のメルケプストラムの散布図，Fig. 3は5次と10次のメルケプストラム系列の短時間フーリエ変換の平均パワースペクトル（変調スペクトル[13]）である。また，Fig. 4は評価セット全体についての目標と予測したメルケプストラムのフレームごとの絶対誤差，系列の標準偏差の絶対誤差，変調スペクトルの絶対誤差を箱ひげ図で描いたものである。

従来法の系列と目標の系列の代表例を比較すると，従来法の系列は微細構造が再現されておらず平滑化されており，系列の変動（振幅や分散）はやや小さかった（Fig. 1右）。また，系列の分布は十分な広がりがなく特定の範囲に集中していた（Fig. 2右）。さらに，変調スペクトルは30 Hz以上において10 dB低く，高周波成分を再現できていなかった（Fig. 3右）。一方で，提案法の系列と目標の系列の代表例を比較すると，提案法の系列は微細構造が再現されており，その変動もほぼ目標の系列と同じであった（Fig. 1中）。また，系列の分布は目標の分布と似ていた（Fig. 2中）。さらに，変調スペクトルは20〜80 Hzにおいて数dB低いが概ね同じであった（Fig. 3中）。

目標に対する従来法と提案法の誤差を比較すると，フレームごとの誤差は従来法の方が提案法よりも0.01小さかった（Fig. 4左）。一方で，系列の分布の広がりである標準偏差の誤差は提案法の方が従来法よりも0.03小さかった。（Fig. 4中）また，変調スペクトルの誤差も提案法の方が従来法よりも12 dB小さかった（Fig. 4右）。

5　各損失関数の効果

提案法における各損失関数の影響を調査したところ，以下のことが分かった。$\mathcal{L}_{\mathrm{GV}}$を利用すると，系列の振幅のスケールが目標に近づき、合成音声の籠りが緩和した。一方で，次元間の関係が無視されるため，局所においてパワーが急に大きくなる問題が発生した。

$\mathcal{L}_{\mathrm{LV}}$を利用すると，局所内での分散が補償されるため微細構造が現れ変調スペクトルが目標に近づいた。一方で，次元間の相関関係が無視されるため合成音声に不自然な揺らぎが現れた。

$\mathcal{L}_{\mathrm{GC}}$，$\mathcal{L}_{\mathrm{LC}}$，$\mathcal{L}_{\mathrm{DD}}$は単体では効果がなかったが，$\mathcal{L}_{\mathrm{GV}}$や$\mathcal{L}_{\mathrm{LV}}$と同時に利用することでそれらの制約として働いた。$\mathcal{L}_{\mathrm{GC}}$や$\mathcal{L}_{\mathrm{LC}}$を利用すると，$\mathcal{L}_{\mathrm{LV}}$で無視されていた相関関係が考慮されるため系列の分布の形状が目標に近づき，合成音声の不自然な揺らぎが緩和された。

$\mathcal{L}_{\mathrm{DD}}$を利用すると，$\mathcal{L}_{\mathrm{GV}}$を利用したときに発生する合成音声の局所における急なパワーの変化が抑制された。

提案法は敵対的学習やリカレント系や畳み込み系のネットワークが暗黙的に学習する系列の分布や隣接するフレームとの関係を明示的に設計しなければならない。一方で，提案法は基本周波数やメルケプストラムなど出力する音声特徴量ごとに適した2次統計量を取捨選択して学習できるので学習結果を操作しやすい利点がある。

6　おわりに

FFNNで時系列を考慮した学習をするために，2次統計量を用いた損失関数による学習方法を提案した。その結果，平均二乗誤差よりも目標値に迫るスペクトル特徴量系列を生成することができた。今後は，敵対的学習との比較評価および受聴実験を行う。

参考文献

[1] Oord *et al*, arXiv.1609.03499, 2016.
[2] Wang *et al*, Interspeech, 2238-2242, 2016.
[3] Matsunaga *et al*, SSW10, 2019.
[4] Nose *et al*, IEEE 8(2), 221-228, 2014.
[5] Toda *et al*, IEICE, E90-D(5), 816-824, 2007.
[6] Saito *et al*, IEEE/ACM, 26(1), 2018.

[7] 高道ら，音講論（秋），195-196, 2017.
[8] SPTK, http://sp-tk.sourceforge.net/
[9] 松永ら，音講論（春），1089-1090, 2019.
[10] Morise *et al*, IEICE, E99-D(7), 1877-1884, 2016.
[11] Nair *et al*, ICML, 807-814, 2010.
[12] Kingma *et al*, arXiv:1412.6980, 2014.
[13] 高道ら，音講論（秋），307-308, 2013.

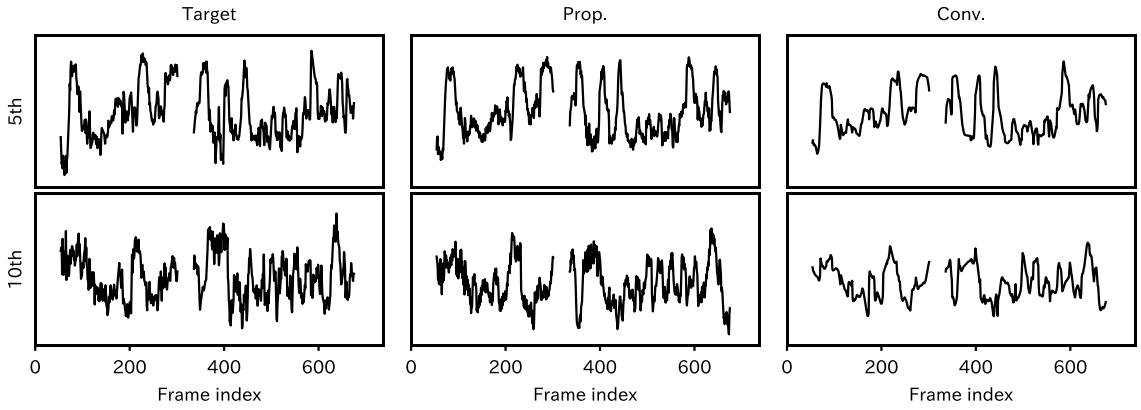

Fig. 1 Examples of 5th and 10th mel-cepstrum trajectories.

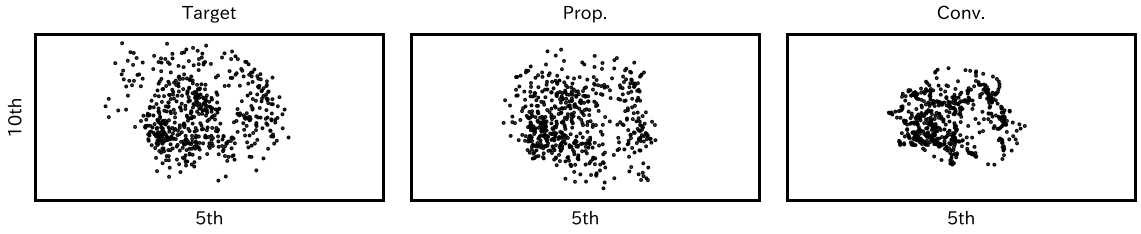

Fig. 2 Examples of scatter between 5th and 10th mel-cepstrums.

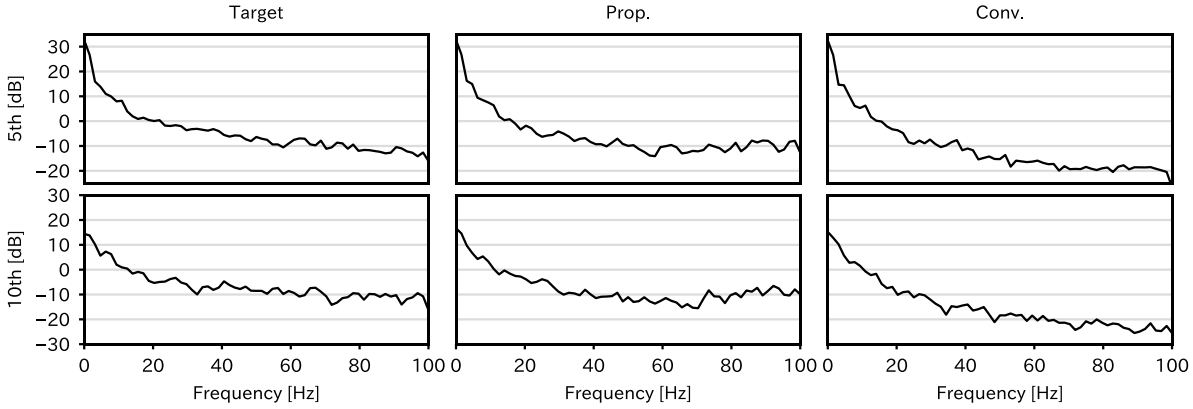

Fig. 3 Examples of modulation spectrum of 5th and 10th mel-cepstrum trajectories.

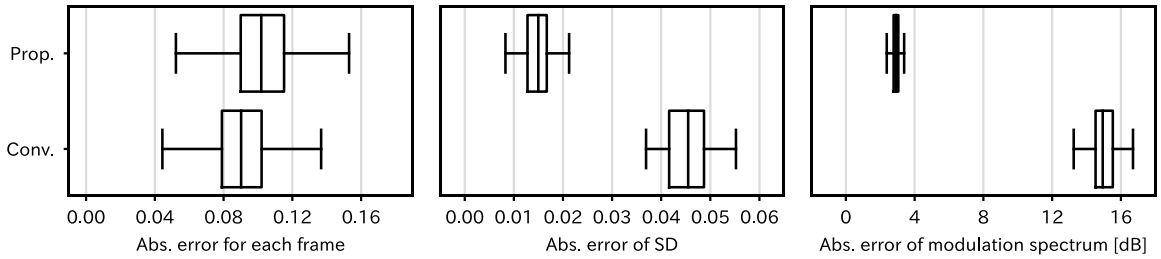

Fig. 4 Comparison between conventional loss function and proposed loss function.
Left: absolute errors of mel-cepstrum trajectories.
Middle: absolute errors of standard deviation of mel-cepstrum trajectories.
Right: absolute errors of modulation spectrum of mel-cepstrum trajectories.