深層学習に基づく日本語音声合成の基本周波数のための言語特徴量の正規化手法の検討*

○松永悟行（(株)エーアイ，富山県立大），
大谷大和（(株)エーアイ），平原達也（富山県立大）

1　はじめに

統計的音声合成は，テキストと音声の関係を音響モデルと呼ばれる統計モデルによって表現される．近年では統計モデルとして Deep Neural Network（DNN）が用いられる[1]．DNN 音声合成の入力である言語特徴量は，音素や品詞などの二値で表現される属性や，アクセント型やモーラ数などの整数値で表現される属性や，word2vec などの分散表現された属性で構成される[2]．これらの属性のうち整数値で表現される属性は，学習データから求めた最小値と最大値による正規化が適用される．

音声合成では自由文章が入力であり，正規化の範囲を超える外れ値が言語特徴量に含まれる場合がある．また，DNN の外挿能力が十分でないため[3]，外れ値の対策が必要となる．一般的には，学習データを多くしてカバーする範囲を広くする対策を採る．しかし，この対策ではすべての入力のパターンをカバーすることはできないため，完全に外れ値の発生を防ぐことはできない．

そこで本報告では日本語音声合成の品質を決定する重要な要素のひとつである基本周波数を対象に，自由文章において外れ値を発生させない正規化手法を提案し，従来手法と比較することでその有効性を示す．

2　従来の言語特徴量の正規化手法

一発話の言語特徴量は時刻tにおけるd次元の属性を要素に持つ言語特徴量ベクトルの系列であり，

$$\boldsymbol{L} = [\boldsymbol{L}_1, \cdots, \boldsymbol{L}_t, \cdots, \boldsymbol{L}_T] \quad (1)$$

$$\boldsymbol{L}_t = [L_{t1}, \cdots, L_{td}, \cdots, L_{tD}] \quad (2)$$

と表せる．学習に用いる言語特徴量の集合を$\boldsymbol{L} \in \boldsymbol{U}$とすると，言語特徴量の$d$次元の属性を正規化するための最小値と最大値は

$$k_d^{min} = \min_{t,\boldsymbol{U}} L_{td} \quad (3)$$

$$k_d^{max} = \max_{t,\boldsymbol{U}} L_{td} \quad (4)$$

となり，正規化された言語特徴量は

$$\boldsymbol{M} = [\boldsymbol{M}_1, \cdots, \boldsymbol{M}_t, \cdots, \boldsymbol{M}_T] \quad (5)$$

$$\boldsymbol{M}_t = [M_{t1}, \cdots, M_{td}, \cdots, M_{tD}] \quad (6)$$

$$M_{td} = \frac{L_{td} - k_d^{min}}{k_d^{max} - k_d^{min}} \quad (7)$$

となる．M_{td}はL_{td}のスケールが変化しただけであり，値が持つ意味は変化せず，それぞれの値は一意になる．M_{td}が外れ値となるかは$k_d^{\min}$と$k_d^{\max}$によって決まるため，この正規化手法は学習データへの依存性が高い．

3　提案する言語特徴量の正規化手法

提案する手法は，外れ値を発生させないために 1 発話内の値のみを用いて，言語的に関連する属性の比をとることにより正規化する．正規化後の言語特徴量は

$$\boldsymbol{N} = [\boldsymbol{N}_1, \cdots, \boldsymbol{N}_t, \cdots, \boldsymbol{N}_T] \quad (8)$$

$$\boldsymbol{N}_t = [N_{t1}, \cdots, N_{td}, \cdots, N_{tD}] \quad (9)$$

$$N_{td} = \frac{L_{td}}{L_{t\delta}}, \qquad d \neq \delta, L_{td} < L_{t\delta} \quad (10)$$

となる．この手法では，N_{td}はL_{td}とは異なる表現になり，値が持つ意味が変化する．N_{td}は 1 発話内の閉じた条件で正規化されるため外れ値が発生することはない．さらに，$k_d^{\min}$と$k_d^{\max}$の影響がないため，この手法は学習データへの依存性が低い．

4　実験条件

本実験では，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，100 から 2000 発話までの 7 つの学習セットを用意した（$\boldsymbol{U}_{\{m|m=100,200,300,400,500,1000,2000\}}$）．言語特徴量は HTS に準拠するものであり[4]，従来手法と提案手法のベクトルの次元数はそ

* Normalized method of linguistic feature for fundamental frequency of Japanese speech synthesis using deep neural network, by MATSUNAGA, Noriyuki, OHTANI, Yamato and HIRAHARA, Tatsuya.

れぞれ 522 次元と 527 次元である．

基本周波数は，16 bit，48 kHz でサンプリングした収録音声から WORLD により 5 ms のフレーム周期で抽出した[5]．無音および無声区間を補間した後に対数化して，2 次までの動的特徴量を付加した．また，無音区間は学習から除外し，学習セットごとに平均と分散を求め標準化した．さらに，予測値に対しては動的特徴量を考慮したパラメータ生成法により平滑化した[6]．継続長については音素境界の条件を一致させるため学習および予測においてラベリングされた値をそのまま使用した．

DNN は，ノード数を 512，ReLU 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順伝播型のネットワークとした．損失関数は平均二乗誤差であり，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）とした[7]．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．

5　実験結果

評価用のデータは学習に使用した 100 発話（$\boldsymbol{U}_{\mathrm{closed}} = \boldsymbol{U}_{100}$）と外れ値を含む学習に使用していない 100 発話（$\boldsymbol{U}_{\mathrm{open}}$）の合計で 200 発話を使用した．Fig. 1 は学習データと同様の前処理を適用した基本周波数を正解値としたときの予測値のフレームごとの絶対誤差を箱ひげ図で表現したものである．また，この実験を 5 回行った結果，外れ値はモデルの収束した最適解によりばらつきがあったため参考として扱い，最適解に大きく影響しなかった中央値と四分位範囲にて評価する．

$\boldsymbol{U}_{\mathrm{closed}}$については，いずれの学習セットにおいても両手法ともにほぼ同じ誤差分布であり，学習データが増加するごとに誤差は微増した．一方で，$\boldsymbol{U}_{\mathrm{open}}$については，いずれの学習データセットにおいても提案手法の方が従来手法よりも誤差は小さく，両手法とも学習データが増加するごとに誤差は減少した．$\boldsymbol{U}_{100}$の提案手法の分布と$\boldsymbol{U}_{2000}$の従来手法の分布はほぼ同じであり，提案手法の方が少ない学習データでも安定した予測ができていた．

6　考察

提案手法の方が少量の学習データでも良好な予測結果だった要因としては，提案手法により外れ値が発生しなくなったことに加え，正規化後の値が基本周波数に適したものとなったことにあると考える．例えば，4 モーラ 2 型と 6 モーラ 3 型のアクセント句では，どちらも基本周波数は句の中間あたりで下降する．これらのアクセント型の値は，従来手法ではそれぞれ別々の値であり，DNN はそれぞれの値に対する基本周波数を学習することになる．一方，提案手法ではどちらの値も 0.5 となり，基本周波数が句の中間で下降するという特徴を捉えた値となる．これにより，アクセント型と基本周波数のパターンが明確化され，従来手法よりも効率的に学習できたと考える．

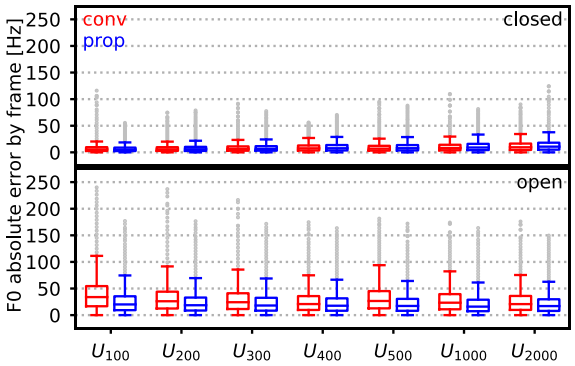

Fig. 1　Relationship between amount of training data and prediction error.

また、少量のデータでも良好な学習結果が得られることは，学習済みの DNN に対して少量のデータで fine tuning する場合においても良好な適応結果が得られることを示唆している．

7　おわりに

提案した 1 発話内の値のみを利用した言語特徴量の正規化手法は，従来の学習データから求めた最小値と最大値による正規化手法よりも少ない学習データ量で安定した基本周波数の予測を可能にした．今後は，主観評価を行い，提案手法の総合的な評価を行う．

参考文献

[1] H. Zen *et al.*, ICASSP, 7962-7966, 2013.
[2] T. Mikolob *et al.*, arXiv:1301.3781, 2013.
[3] D. Yu *et al.*, ICML, vol.32, II-1188-II-1196, 2014.
[4] HTS, http://hts.sp.nitech.ac.jp/
[5] M. Morise *et al.*, ICASSP, 3933-3936, 2008.
[6] 徳田 他，日本音響学会誌，vol.53，no.3，192-200，1997.
[7] D. P. Kingma *et al.*, arXiv:1412.6980, 2014.