コンテキスト事後確率の Sequence-to-Sequence 学習を用いた音声変換 *

☆三好 裕之（東大），齋藤 佑樹，高道 慎之介，猿渡 洋（東大院・情報理工）

1 はじめに

音声変換とは，入力音声の言語情報を保持したまま，パラ言語・非言語情報を変換する技術である．テキストと音声のパラレルデータを必要とするテキスト依存音声変換 [1]（音声認識と音声合成を用いた音声変換）は，複数話者音声のパラレルデータを必要とするテキスト非依存音声変換 [2] に比べ，学習データ収集のコストが小さい．コンテキスト事後確率の複写に基づく音声変換 [3, 4] は，テキスト依存音声変換から発展した手法であり，入力音声特徴量から推定されたコンテキスト事後確率から出力音声特徴量を生成する．これはテキスト依存音声変換と同様にデータ収集のコストを抑えつつ，明示的なテキスト推定を回避する手法である．一方で，従来の学習データである，テキストと音声のパラレルデータに加え，入出力話者の音声のパラレルデータが新たに与えられる場合，より高精度な音声変換が可能になると期待される．本稿では．入力音声のコンテキスト事後確率を出力音声のコンテキスト事後確率に変換する手法を提案する．コンテキスト事後確率の Sequence-to-Sequence (Seq2Seq) 学習を行うことで，コンテキスト情報の可変長変換を行う．提案法により，コンテキスト事後確率に含まれる話速や音韻性などの話者性を変換することが可能になる．実験的評価により，提案法の有効性を示す．

2 コンテキスト事後確率の複写に基づく音声変換

まず, テキストと音声のパラレルデータから音声認識の音響モデルである neural network を構築する．ここで，系列長 T_x の $\boldsymbol{x}$ と系列長 T_y の $\boldsymbol{y}$ をそれぞれ入出力話者の音声特徴量系列，$\boldsymbol{l}_x$，$\boldsymbol{l}_y$ をそれぞれ，$\boldsymbol{x}$ と $\boldsymbol{y}$ に対応するフレーム毎の離散コンテキスト系列とする．フレーム毎のコンテキスト事後確率を出力する neural network $\boldsymbol{R}(\cdot)$ は，$\boldsymbol{l}_x$ と $\boldsymbol{R}(\boldsymbol{x})$ 間の cross-entropy $L_{\mathrm{C}}(\boldsymbol{l}_x, \boldsymbol{R}(\boldsymbol{x}))$ を最小化するように学習される．この学習は $\boldsymbol{y}$ を含む多人数話者データを用いて行われ，最終的に，不特定話者モデルを構築する．

次に，音声認識により推定されたコンテキスト事後確率系列 $\hat{\boldsymbol{p}}_y = \boldsymbol{R}(\boldsymbol{y})$ と $\boldsymbol{y}$ を用いて，音声合成用の neural network $\boldsymbol{S}(\cdot)$ を構築する．この neural network は，$\boldsymbol{S}(\hat{\boldsymbol{p}}_y)$ と $\boldsymbol{y}$ 間の平均二乗誤差 $L_{\mathrm{G}}\left(\boldsymbol{y}, \boldsymbol{S}(\hat{\boldsymbol{p}}_y)\right)$ を最小化するように学習される．

音声変換時には，$\boldsymbol{x}$ に対応するコンテキスト事後確率系列 $\hat{\boldsymbol{p}}_x = \boldsymbol{R}(\boldsymbol{x})$ を推定し，合成音声特徴量系列 $\hat{\boldsymbol{y}}$ を次式で推定する．

$$\hat{\boldsymbol{y}} = \boldsymbol{S}(\hat{\boldsymbol{p}}_x) = \boldsymbol{S}(\boldsymbol{R}(\boldsymbol{x})) \tag{1}$$

Fig. 1 の上段と中段は，これらの手順の図示である．

この手法は入出力話者のパラレルデータを必要としないため，データ収集が容易である．しかし，入力音声のコンテキスト事後確率を複写するため，事後確率のもつ話者性（話速など）の変換は困難である．

3 コンテキスト事後確率の Seq2Seq 変換に基づく音声変換

2 節で記述した音声変換を構築する際に，新たに入出力話者のパラレルデータが入手可能な場合を想定し，入出力話者間のコンテキスト事後確率を変換するモデルを Seq2Seq 学習により構築する．

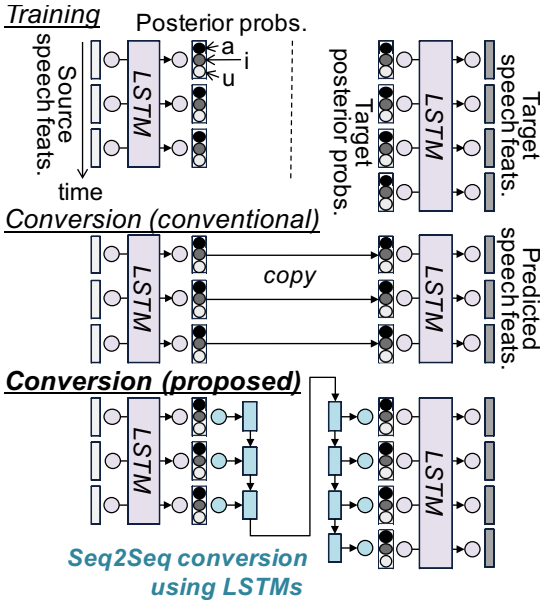

Fig. 1 従来法 (上段と中段) と提案法 (下段) の比較.

3.1 Seq2Seq 学習

一般的に，系列長の異なる入出力特徴量系列を変換する場合，動的時間伸縮 (DTW) などの強制アライメントにより，各特徴量系列を固定長系列に変換し，対応関係をとる．Neural network を用いた Seq2Seq 学習 [5] は，強制アライメントによる品質劣化を避け，系列単位で入出力系列の変換規則をモデル化する．

3.2 コンテキスト事後確率の Seq2Seq 変換

事後確率の変換手順を Fig. 1 の下段に示す．3.1 節の学習法をコンテキスト事後確率の Seq2Seq 変換に適用する．以降では，同一テキストに対する入出力音声特徴量系列 $\boldsymbol{x}$ 及び $\boldsymbol{y}$ を使用する．まず，2 節で構築した音声認識の音響モデルを用いて，$\boldsymbol{x}$ 及び $\boldsymbol{y}$ に対応するコンテキスト事後確率系列 $\hat{\boldsymbol{p}}_x$ 及び $\hat{\boldsymbol{p}}_y$ を取得する．次に，$\hat{\boldsymbol{p}}_x$ を $\hat{\boldsymbol{p}}_y$ に変換する neural network $\boldsymbol{C}(\cdot)$ を構築する．$\boldsymbol{C}(\cdot)$ は encoder-decoder model [5] で記述され，次式の損失関数 $L(\cdot)$ を最小化するように学習される．

$$L\left(\hat{\boldsymbol{p}}_y, \hat{\boldsymbol{p}}_x, \boldsymbol{l}_y\right) = L_{\mathrm{G}}\left(\boldsymbol{C}\left(\hat{\boldsymbol{p}}_x\right), \hat{\boldsymbol{p}}_y\right) + L_{\mathrm{C}}\left(\boldsymbol{C}\left(\hat{\boldsymbol{p}}_x\right), \boldsymbol{l}_y\right) \tag{2}$$

第一項は，$\hat{\boldsymbol{p}}_y$ をターゲットとした変換誤差に相当する．一方で第二項は，音声認識の学習時にターゲットとした $\boldsymbol{l}_y$ との cross-entropy を最小化する項であり，$\hat{\boldsymbol{p}}_y$ に含まれる音声認識エラーを緩和するものである．

音声変換時には，式 (1) に事後確率変換を追加した次式で $\hat{\boldsymbol{y}}$ を推定する．

$$\hat{\boldsymbol{y}} = \boldsymbol{S}(\boldsymbol{C}(\boldsymbol{R}(\boldsymbol{x}))) \tag{3}$$

---

* Voice Conversion Using Sequence-to-Sequence Learning of Context Posterior Probabilities, by MIYOSHI, Hiroyuki, SAITO, Yuki, TAKAMICHI, Shinnosuke, and SARUWATARI, Hiroshi (The University of Tokyo).

3.3 従来法との比較

テキスト依存音声変換は，入力音声特徴量セグメントを単一コンテキスト（例えば，音素，シラブル，単語単位）に強制アライメントし，コンテキスト系列から出力音声特徴量を生成する．この方法はコンテキスト系列の柔軟な変換（例えば，可変長変換）を可能にするが，音声特徴量セグメントから単一コンテキストへのマッピングに伴う時間量子化の影響は避けられない．DTW を用いるテキスト非依存音声変換 [6] は，入出力音声特徴量をフレーム単位で強制アライメントし変換する．フレームレベルの変換により時間量子化の影響を回避できるが，（暗黙的に考慮される）コンテキスト系列の変換は制限される．2 章の従来法 [3] もこのタイプに該当する．これは，入力音声のコンテキスト系列が直接的に出力音声の合成に使用されるためである．これらと比較して提案法は，フレームレベルかつ，強制アライメントを使用しない可変長の系列変換を行うため，時間量子化の影響を避け，コンテキスト系列の柔軟な変換を可能にする．

4 実験的評価

4.1 実験条件

実験に用いるデータとして，ATR 音素バランス 503 文 [7] を利用し，A–I セット 450 文を学習に J セット 53 文を評価に用いる．学習データのサンプリング周波数は 16 kHz，フレームシフトは 5 ms とする．STRAIGHT 分析 [8] により，0 次から 24 次のメルケプストラム，F_0，5 帯域の平均非周期性指標 [9] を抽出し，1 次以降のメルケプストラムの静的・動的特徴量を平均 0，分散 1 に正規化したものを音声特徴量とする．音声認識の音響モデルは，8 名の音声データを用いて学習される話者非依存 bi-directional LSTM (Long Short-Term Memory) である．この音声データには，女性入力話者 1 名と男性出力話者 1 名も含まれる．LSTM の隠れ素子数は 1024，出力層の活性化関数は softmax 関数である．離散コンテキストは quin-phone とし，学習時の損失関数として，quin-phone を先行・当該・後続音素などの 5 つのグループに分け，各グループに対しての cross-entropy の和を用いる．音声合成の音響モデルの構造は音声認識のモデルと同様だが，出力層の活性化関数は線形関数である．事後確率を変換する encoder-decoder model の構造には，encoder に bi-directional LSTM，decoder に LSTM を用いる．出力層の活性化関数は softmax 関数である．最適化アルゴリズムとして学習率 0.01 の AdaGrad を用いる．長い発話における Seq2Seq 学習の精度低下 [10] を回避するため，本稿では音素毎の事後確率変換を行う．学習データ及び評価データにおける入出力音声の音素継続長を既知とし，当該音素内の事後確率を変換する音素非依存 encoder-decoder model を構築する．合成音声の F_0 は，入力音声の F_0 に対して線形変換 [6] を施した後，時間伸縮により生成する．この時間伸縮の経路は，入力音声の事後確率系列と，encoder-decoder model で変換された合成音声の事後確率系列を用いた DTW により決定する．非周期性指標は，時間伸縮のみを適用する．

従来法 [3] と提案法を比較するため，客観評価と主観評価を行う．客観評価では，変換したメルケプストラムと出力話者のメルケプストラム間のメルケプストラム歪みを計算する．ただし，従来法により変換したメルケプストラムは，提案法の F_0 変換で使用した時間伸縮経路を用いて伸縮を行った後，歪みを計算する．これは，提案法の実験設定と同様に入出力音声の継続長を既知とした場合の従来法に対応する．主観評価では，変換音声の話者性に関する XAB テスト，音質に関する AB テストを実施する．評価者数は，各評価で 7 人である．

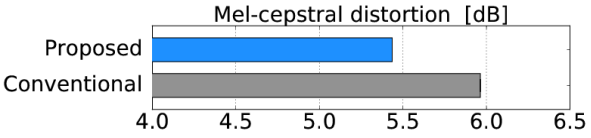

Fig. 2　客観評価結果

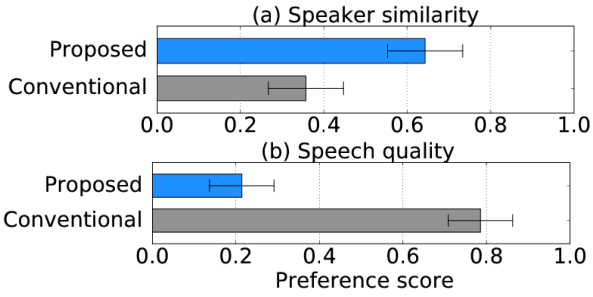

Fig. 3　主観評価結果（エラーバーは 95% 信頼区間）

4.2 客観評価結果

Fig. 3 にメルケプストラム歪みの結果を示す．従来法と比較して，提案法による大幅な改善がみられる．これは，提案法で DTW を避け Seq2Seq 変換を行ったためである．ただし，提案法の音声変換時に継続長を既知としているため，この評価値は提案法における理想値に相当することに注意する．

4.3 主観評価結果

話者性と音質に関する主観評価結果を Fig. 4 に示す．Fig. 4 (a) より話者性の改善がみられる．一方で Fig. 4 (b) に関しては従来法を下回ることが分かる．この原因に関して，いくつかの評価データにおいて，Seq2Seq 変換の誤りにより音韻性が失われる現象を観測した．今後は，この音韻性の消失について検討する必要がある．

5 まとめ

本稿では，コンテキスト事後確率の Seq2Seq 変換による音声変換を提案した．実験的評価により，提案法は，変換音声のメルケプストラム歪みと話者性に関して，従来の事後確率複写による音声変換を上回ることを明らかにした．今後は，事後確率変換による音韻性消失を緩和するため，認識・変換・合成の同時学習を検討する．

謝辞　本研究は，総合科学技術・イノベーション会議による革新的研究推進プログラム（ImPACT），セコム科学技術振興財団，及び JSPS 科研費 16H06681 の支援を受けた．

参考文献

[1] A. Kain et al., *Proc. ICASSP*, pp. 285–288, 1998.
[2] Y. Stylianou et al., *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1988.
[3] L. Sun et al., *Proc. IEEE ICME*, 2016.
[4] L. Sun et al., *Proc. INTERSPEECH*, pp. 322–326, 2016.
[5] S. Ilya et al., *Proc. NIPS*, pp. 1–6, 2014.
[6] T. Toda et al., *IEEE Trans. on ASLP*, vol. 15, no. 8, pp. 2222–2235, 2007.
[7] 阿部 他, ATR テクニカルレポート, TR–I–0166, 1990.
[8] Kawahara et al., *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.
[9] Ohtani et al., *Proc. INTERSPEECH*, pp. 2266–2269, 2006.
[10] W. Wang et al., *Proc. INTERSPEECH*, pp. 2243–2247, 2016.