社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

コンテキスト事後確率の Sequence-to-Sequence 学習を用いた音声変換と Dual Learning の評価

三好　裕之† 　齋藤　佑樹† 　高道慎之介† 　猿渡　洋†

† 東京大学大学院 情報理工学系研究科　〒 113–8656 東京都文京区本郷 7-3-1

あらまし　本稿では，コンテキスト事後確率の sequence-to-sequence 学習を用いた音声変換を提案する．従来のコンテキスト事後確率の複写に基づく音声変換は，パラレルデータが不要であるという利点があるが，入力音声特徴量に対するコンテキスト事後確率を複写するため，事後確率に含まれる話速や音韻性の変換が困難であった．提案手法では，学習データに部分的なパラレルデータが含まれていると仮定し，入出力話者のコンテキスト事後確率の可変長変換を行うことで，事後確率に含まれる話速や音韻性の変換を可能にする．さらに，音声認識・事後確率変換・音声合成の全てのモデルを同時に最適化するための dual learning も導入する．実験的評価では，sequence-to-sequence 学習及び dual learning の有効性を客観評価及び主観評価により検証する．

キーワード　音声変換，コンテキスト事後確率，sequence-to-sequence 学習，dual learning

Voice Conversion Using Sequence-to-Sequence Learning of Context Posterior Probabilities and Evaluation of the Dual Learning

Hiroyuki MIYOSHI†, Yuki SAITO†, Shinnosuke TAKAMICHI†, and Hiroshi SARUWATARI†

† Graduate School of Information Science and Technology, The University of Tokyo, Hongo 7–3–1, Bunkyo-ku, Tokyo, 113–8656 Japan

Abstract　Voice conversion (VC) using sequence-to-sequence learning of context posterior probabilities is proposed. Conventional VC using shared context posterior probabilities predicts target speech parameters from the context posterior probabilities estimated from the source speech parameters. Although conventional VC can be built from non-parallel data, it is difficult to convert speaker individuality such as phonetic property and speaking rate contained in the posterior probabilities because the source posterior probabilities are directly used for predicting target speech parameters. In this work, we assume that the training data partly include parallel speech data and propose sequence-to-sequence learning between the source and target posterior probabilities. The conversion models perform non-linear and variable-length transformation from the source probability sequence to the target one. Further, we propose a joint training algorithm for the modules. In contrast to conventional VC, which separately trains the speech recognition that estimates posterior probabilities and the speech synthesis that predicts target speech parameters, our proposed method jointly trains these modules along with the proposed probability conversion modules. Experimental results demonstrate that our approach outperforms the conventional VC.

Key words　Voice conversion, context posterior probabilities, sequence-to-sequence learning, dual learning

1. はじめに

音声変換とは，入力音声に含まれる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り，音声強調 [1,2] や非母語話者の言語教育 [3] などに応用されている．音声変換の方式は，テキスト非依存音声変換と，テキスト依存音声変換の 2 つに大別される．テキスト非依存音声変換では，入力音声特徴量と出力音声特徴量の組（パラレルデータ）を用いて音響モデルを学習する．音響モデルとして Gaussian mixture model [4,5] や deep neural network (DNN) [6] が用いられ，入力音声特徴量から出力音声特徴量を直接的に予測する．この手法は，高品質な音声変換を実現可能だが，パラレルデータの収集を必要とする．一方で，テキスト依存音声変換 [7,8] は，音声の言語情報を活用して音声特徴量を変換する．音声変換部は，入力音声

からその言語情報を推定する音声認識モデルと，推定された言語情報を用いて音声特徴量を予測する音声合成モデルの 2 つで構成される．この手法では，音声変換部の構築にパラレルデータが不要であるため，学習データを容易に収集できる．しかし，音素，単語などの単位で音声特徴量を変換するため，フレーム単位で音声特徴量を変換するテキスト非依存音声変換と比較して品質が劣化する．

フレーム単位でのテキスト依存音声変換を実現する手法として，コンテキスト事後確率に基づく音声変換 [9] が提案されている．この手法では，入力音声特徴量からコンテキスト事後確率をフレーム毎に推定し，その推定結果を用いて出力音声特徴量を予測する．この手法はソフトなテキスト依存音声変換方式と解釈でき，クロスリンガルテキスト音声合成 [10, 11] に拡張できる．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，コンテキスト事後確率に含まれる話速や音韻性等の話者性の変換が困難である．

上記の問題点を踏まえ，本稿では，コンテキスト事後確率の sequence-to-sequence 学習を用いた音声変換を提案する．ここでは，学習データに部分的なパラレルデータ（例えば部分的なフレーズの一致など）が含まれると仮定し，可変長の系列を変換する encoder-decoder モデル [12] を用いて入力音声のコンテキスト事後確率を出力音声の事後確率に変換する．事後確率変換のモデルは，従来の音声認識モデルと音声合成モデルの間に挿入される．事後確率変換モデルを構築しない場合は，従来のコンテキスト依存音声変換 [9] を利用できる．さらに，本稿では，音声認識・事後確率変換・音声合成の 3 つのモデルを同時に学習する dual learning も導入する．従来のコンテキスト依存音声変換 [9] が音声認識・合成のモデルを独立に学習していたのに対し，提案手法では入力音声特徴量を出力音声特徴量に変換するための一連のモデルを同時に学習する．実験的評価により，(1) コンテキスト事後確率の sequence-to-sequence 学習が合成音声の話者性の改善に有効であること，(2) 音声認識・合成の dual learning が合成音声の品質及び話者性の両方の改善に有効であることを示す．

2. コンテキスト事後確率の複写に基づく音声変換

コンテキスト事後確率の複写に基づく音声変換 [9] では，音声認識と音声合成の 2 つのモデルを用いる．これらのモデルは個別に学習され，変換時には 2 つのモデルを連結する．図 1 にコンテキスト事後確率の例を示す．入力音声のコンテキスト事後確率は音声認識モデルにより推定され（図 2 上），推定結果を複写したものを音声合成モデルに入力することで合成音声特徴量を予測する（図 2 中央）．

2.1 音声認識・合成モデルの学習部

入力音声特徴量系列を $\boldsymbol{x} = [\boldsymbol{x}_1^\top, \cdots, \boldsymbol{x}_{T_x}^\top]^\top$ とし，出力音声特徴量系列を $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_{T_y}^\top]^\top$ とする．ここで，$\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ はフレーム t における音声パラメータであり，T_x と T_y は特徴量の系列長である．また，$\boldsymbol{x}$ と $\boldsymbol{y}$ に対応するコンテキストラベル系列を $\boldsymbol{l}^{(x)} = [\boldsymbol{l}_1^{(x)}, \cdots, \boldsymbol{l}_{T_x}^{(x)}]^\top$，$\boldsymbol{l}^{(y)} = [\boldsymbol{l}_1^{(y)}, \cdots, \boldsymbol{l}_{T_y}^{(y)}]^\top$ とする．話者非依存の音声認識モデルの DNN $\boldsymbol{R}(\cdot)$ は，これらの音声特徴量を用いて学習される．音声認識モデルは，正解ラベル系列 $\boldsymbol{l}^{(x)}$ と認識結果 $\boldsymbol{R}(\boldsymbol{x})$ の cross-entropy $L_C(\boldsymbol{l}^{(x)}, \boldsymbol{R}(\boldsymbol{x}))$ を最小化するように学習される．

音声合成モデルは，音声認識モデルの推定結果として得られるコンテキスト事後確率 $\hat{\boldsymbol{p}}_y = \boldsymbol{R}(\boldsymbol{y})$ から，出力音声特徴量系列 $\boldsymbol{y}$ を予測する．出力話者依存の音声合成モデルの DNN $\boldsymbol{G}(\cdot)$ は，目的話者の音声特徴量 $\boldsymbol{y}$ と予測結果 $\boldsymbol{G}(\hat{\boldsymbol{p}}_y)$．の二乗誤差 $L_G(\boldsymbol{y}, \boldsymbol{G}(\hat{\boldsymbol{p}}_y))$ を最小化するように学習される．

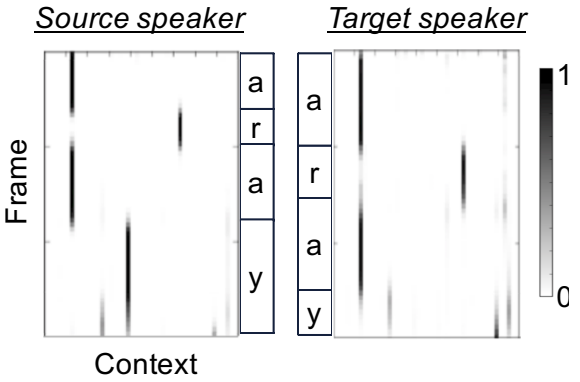

図 1　コンテキスト事後確率の例（左: 入力音声，右: 出力音声）
Fig. 1　An example of source (left) and target (right) context posterior probabilities.

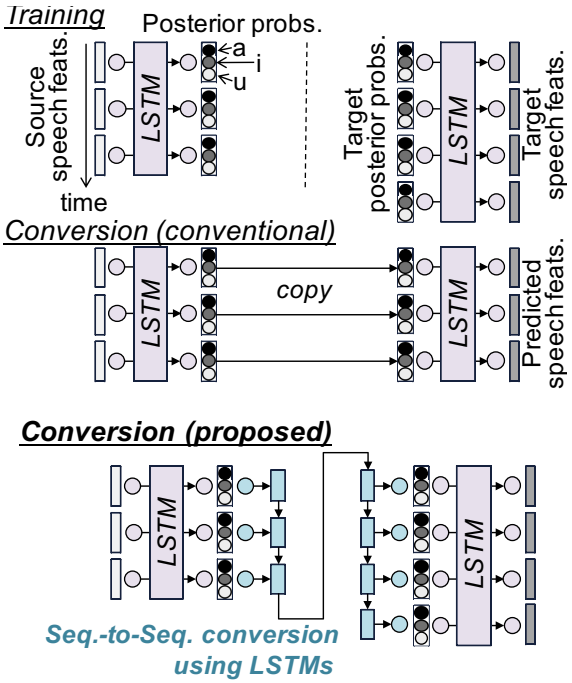

図 2　従来手法及び提案手法における音声変換の手順．提案手法では，入力音声のコンテキスト事後確率が出力音声のものに変換される．
Fig. 2　Training and conversion procedures of conventional and proposed VC. In the proposed VC, the source context posterior probabilities are transformed into the target posterior probabilities.

2.2 音声変換

音声変換時には，出力音声特徴量系列の予測結果 $\hat{\boldsymbol{y}}$ は，個別に学習された音声認識・合成モデルを連結することで得られる．すなわち，入力音声特徴量系列 $\boldsymbol{x}$ から推定されるコンテキスト事後確率を $\hat{\boldsymbol{p}}_x$ とすると，$\hat{\boldsymbol{y}} = \boldsymbol{G}(\hat{\boldsymbol{p}}_x) = \boldsymbol{G}(\boldsymbol{R}(\boldsymbol{x}))$ として出力音声特徴量系列を予測する．ここで，$\boldsymbol{x}, \hat{\boldsymbol{p}}_x, \hat{\boldsymbol{y}}$ の系列長は全て T_x となる．

2.3 従来手法の問題点

従来手法では，音声認識モデルの推定結果として得られるコンテキスト事後確率を音声合成モデルに直接利用するため，図 1 に示すような事後確率に含まれる話速（系列長）や音韻性といった話者性の変換が困難である．また，音声認識モデルの精度を改善させることが，必ずしも合成音声の品質を改善させるとは限らない．

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

従来手法の制限を越える手法として，入力音声のコンテキスト事後確率を出力音声の事後確率に変換する sequence-to-sequence 学習を導入する．

3.1 Sequence-to-Sequence 学習

Sequence-to-sequence 学習は，recurent neural network (RNN) [13] を用いて可変長な系列の変換を実現するための手法である．本稿で採用する encoder-decoder モデルは，入力特徴量系列を固定次元のベクトルに圧縮し，それを異なる長さの出力特徴量系列に変換する．各フレームにおいて，入力側の RNN (encoder) と出力側の RNN (decoder) は，次のフレームの特徴量を予測する．本稿では，encoder-decoder モデルを用いた事後確率変換により，コンテキスト事後確率の可変長変換を実現する．提案手法の枠組みを図 2 下に示す．

3.2 事後確率変換モデルの学習

事後確率変換モデルの学習法として，(1) 事後確率変換モデルの個別学習，(2) 事後確率変換モデルを介した音声認識・合成モデルの dual learning [14] の 2 つを提案する．音声変換時には，学習済みの音声認識・事後確率変換・音声合成の 3 つのモデルを連結させることで入力音声特徴量を変換する．

3.2.1 事後確率変換モデルの個別学習

入出力音声のコンテキスト事後確率のパラレルデータを用いて，事後確率変換の encoder-decoder モデル $\boldsymbol{C}(\cdot)$ を学習する．学習時の損失関数は，次式で与えられる．

$$L(\boldsymbol{l}^{(y)}, \hat{\boldsymbol{p}}_x, \hat{\boldsymbol{p}}_y) = L_G(\hat{\boldsymbol{p}}_y, \boldsymbol{C}(\hat{\boldsymbol{p}}_x)) + L_C(\boldsymbol{l}^{(y)}, \boldsymbol{C}(\hat{\boldsymbol{p}}_x)) \quad (1)$$

ここで，第一項は事後確率変換時の誤差を最小化するための損失である．第二項は $\boldsymbol{l}^{(y)}$ を用いて計算される cross-entropy であり，$\boldsymbol{l}^{(y)}$ に含まれる推定誤差の影響を軽減する効果を持つ．予備実験より，第一項のみを用いた場合と比較して，この定式化により変換精度が向上することを確認している．

Sequence-to-sequence 学習では，長期間の依存性を学習させる際に誤差が蓄積する [15]．そこで，本稿では，入出力音声の音素境界フレームが既知であると仮定し，音素単位での sequence-to-sequence 学習を行う．

3.2.2 事後確率変換モデルを介した音声認識・合成モデルの dual learning

本手法の最終的な目的は音声合成時の誤差最小化であるため，その前処理である音声認識・事後確率変換モデルも，この誤差を最小化するように学習されるすべきである．そこで，本稿では，認識誤差だけでなく，音声合成の誤差も最小化するように音声認識モデル $\boldsymbol{R}(\cdot)$ を学習する．この学習に用いる損失関数は $L_C(\boldsymbol{l}^{(x)}, \boldsymbol{R}(\boldsymbol{x})) + L_G(\boldsymbol{x}, \boldsymbol{G}(\boldsymbol{R}(\boldsymbol{x})))$ である．さらに，事後確率変換モデル $\boldsymbol{C}(\cdot)$ の学習時も同様に音声合成の誤差を考慮し，式 (1) と $L_G(\boldsymbol{G}(\boldsymbol{y}, \boldsymbol{C}(\hat{\boldsymbol{p}}(\boldsymbol{x}))))$ の和を最小化する．音声合成モデルの学習は，従来手法と同様に，二乗誤差最小化基準を用いて行う．

3.3 考　　察

テキスト依存音声変換では，入力音声を単一のコンテキスト（音素，音節，単語など）にアライメントし，そのコンテキストを用いて音声特徴量を生成する．そのため，可変長の系列変換が可能だが，アライメント時に時間の量子化が生じる．一方で，dynamic time warping (DTW) [4] を用いたテキスト非依存音声変換では，入出力音声特徴量をフレーム単位でアライメントする．しかし，アライメントにより系列長が固定されるため，音声特徴量に内包されるコンテキストの変換が困難である．従来のコンテキスト依存音声変換 [9,10] は入力音声のコンテキスト事後確率を直接的に用いて音声合成を行うため，後者に対応する．提案手法ではアライメントを行わずにフレーム単位での変換を行うため，時間の量子化による影響を回避し，コンテキストの柔軟な変換が可能である．

3.2.2 節で提案する学習法は，所望のクラスラベルを用いた auto-encoder [16] とみなすことができる．そのため，クラスラベル（コンテキスト）だけでなく，潜在変数 [17,18] を有する variational auto-encoder [19] の教師あり学習に拡張できる．

4. 実験的評価

4.1 実 験 条 件

従来手法 [9] と提案手法ではそれぞれ，ノンパラレルデータ及び部分的なパラレルデータを用いることが可能だが，本稿では全ての実験においてパラレルデータを用いる．実験的評価に用いるデータとして，ATR 音素バランス 503 文 [20] を利用する．話者非依存の音声認識モデルは，評価に用いる入出力話者を含む計 8 名の音声データで構築する．話者依存の事後確率変換・音声合成モデルは，評価に用いる男性話者 1 名及び女性話者 1 名の音声データのみを用いて構築する．学習には A–I セット 450 文を利用し，評価には J セット 53 文を利用する．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．スペクトル特徴量として STRAIGHT 分析 [21] による 0 次から 24 次のメルケプストラム係数，音源特徴量として F_0, 5 周波数帯域における平均非周期成分 [22,23] を用いる．スペクトル特徴量に対する前処理として，50 Hz のカットオフ変調周波数による trajectory smoothing [24] を利用する．DNN 学習時には，スペクトル特徴量を平均 0，分散 1 に正規化

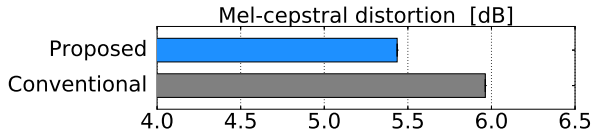

図 3　従来手法と提案手法のメルケプストラム歪み

Fig. 3　Mel-cepstral distortion of conventional VC and proposed VC integrating posterior probability conversion.

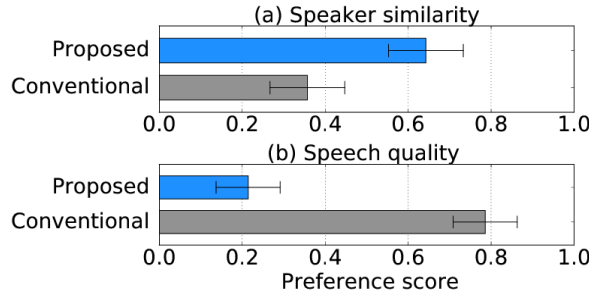

図 4　従来手法と提案手法の合成音声を用いた主観評価結果（エラーバーは 95 %信頼区間）

Fig. 4　Results of subjective evaluation for comparing conventional VC and proposed VC integrating posterior probability conversion. Error bar indicates the 95 % confidence intervals.

する．また，学習データにおける無音フレームの 80 %を除去する．最適化アルゴリズムとして，学習率 0.01 の AdaGrad [25] を用いる．音声認識・合成のモデルは，隠れ素子数を 256 とした bi-directional long short-term memory (LSTM) [9, 10] である．事後確率変換モデルにおいて，encoder は隠れ素子数を 256 とした bi-directional LSTM であり，decoder は隠れ素子数を 256 とした uni-directional LSTM である．

コンテキストラベルとして，quin-phone を用いる．音声認識モデルの学習時には，quin-phone を先行音素や後続音素などの 5 つの組に分割し，各音素に対する cross-entropy の和を最小化する [26]．本稿では，スペクトル特徴量とその動的特徴量のみを認識・合成に用いる．提案手法では，F_0 を線形変換し [4]，事後確率変換後の事後確率系列を用いた DTW によりその長さを伸縮する．平均非周期成分は，DTW による系列長変換のみを施す．事後確率変換時には，系列長を決めるために自然音声の音素継続長を使用する．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は既知とし，当該音素内の事後確率を変換する音素非依存の事後確率変換モデルを構築する．最終的なコンテキスト事後確率系列は，各音素毎の推定結果を連結させることで得られる．

以降では，事後確率変換と dual learning の有効性を実験的に評価する．

4.2　事後確率変換モデルの有効性に関する評価

4.2.1　客 観 評 価

客観評価指標として，目的話者の自然音声と合成音声のメルケプストラム歪みを計算する．系列のアライメント法として，従来手法 [9] では DTW を用いる．提案手法では sequence-to-sequence 学習を用いる． 評価結果を図 3 に示す．従来手法

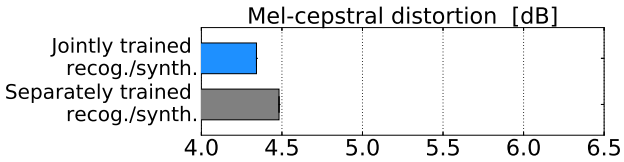

図 5　Auto-encoding におけるメルケプストラム歪み．従来の個別学習と提案する同時学習による認識・合成を比較している．

Fig. 5　Mel-cepstral distortion in auto-encoding case. The conventional separately trained and proposed jointly trained recognition and synthesis modules are compared.

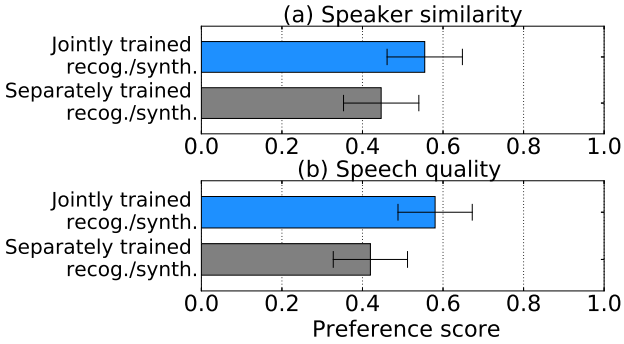

図 6　Dual learning を用いた場合の主観評価結果．従来の個別学習と提案する同時学習による認識・合成を比較している（エラーバーは 95 %信頼区間）．

Fig. 6　Results of objective evaluation for comparing conventional separately-trained and proposed jointly-trained recognition and synthesis modules. Error bars indicate the 95 % confidence intervals.

と比較して，提案手法による大幅な改善がみられる．これは，DTW による歪みの増加を避けたためであるが，提案手法の音声変換時に継続長を既知としたため，この評価値は提案手法における理想値に相当することに注意する．

4.2.2　主 観 評 価

提案手法による影響を確認するため，合成音声の音質に関するプリファレンス AB テスト，及び，話者性に関する XAB テストを実施する．被験者数は各評価に対して 7 名である．

評価結果を図 4 に示す．図 4(a) より，提案手法による話者性の改善がみられる．一方で，図 4(b) より，音質に関しては劣化している．この原因に関して，いくつかの合成音声において，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．

4.3　Dual learning の有効性に関する評価

4.3.1　音声認識・合成モデルの dual learning

ここでは，音声認識・合成モデルを同時に学習させる dual learning の有効性を検証する．まず，音声認識・合成モデルを通じて入力音声特徴量を再構築させる auto-encoding の結果に対してメルケプストラム歪みを計算し，従来手法 [9] と比較した．計算結果を図 5 に示す．この図より，dual learning によるメルケプストラム歪みの改善が確認できる．

次に，dual learning による合成音声の品質改善を確認するための主観評価を行う．評価方法は 4.2.2 節と同様である．評価結果を図 6 に示す．評価結果より，dual learning による合

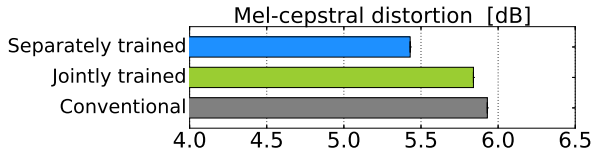

図 7　(1) 従来手法，(2) 音声認識・事後確率変換・音声合成の個別学習，(3) dual learning を用いた場合のメルケプストラム歪み

Fig. 7　Mel-cepstral distortion of three methods: (1) conventional VC, (2) proposed VC using separately trained recognition/synthesis, and (3) proposed VC using jointly trained recognition/synthesis/conversion.

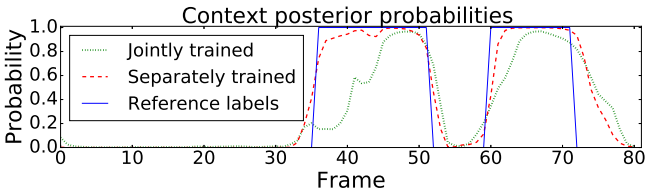

図 8　推定されたコンテキスト事後確率系列の例

Fig. 8　An example of posterior probability sequence.

成音声の音質及び話者性の改善が確認できる．

4.3.2　事後確率変換モデルを介した dual learning

事後確率変換モデルを介した音声認識・合成モデルの dual learning の有効性を検証する．評価指標はメルケプストラム歪みであり，比較対象は，(1) 従来手法 [9]，(2) 個別のモデル学習（4.2.1 節における提案手法），(3) dual learning の 3 つである．

評価結果を図 7 に示す．従来手法と比較すると，dual learning による歪みの改善が確認できる．しかし，個別に学習させた場合と比較すると劣化がみられる．この原因を調査するために，音声認識モデルの出力として得られるコンテキスト事後確率の推定結果を分析した．図 8 に推定結果の一例を示す．この図より，全てのモデルを個別に学習させた場合では，ハードな事後確率が推定されており，当該区間においてほぼ 1 に近い値で，それ以外の区間ではほぼ 0 に近い値となっていることが確認できる．一方で，dual learning を行なった場合では，よりソフトな事後確率が推定されていることがわかる．この結果についてさらなる調査が必要であるが，これが劣化の原因の一つであることが予想される．

5. おわりに

本稿では，コンテキスト事後確率の sequence-to-sequence 学習に基づく音声変換と，音声認識・事後確率変換・音声合成モデルの同時学習を行う dual learning を提案した．実験的評価より，(1) コンテキスト事後確率変換による合成音声の話者性の改善，(2) 音声認識・合成モデルの dual learning による合成音声の音質及び話者性の改善を確認した．今後は，sequence-to-sequence 学習を行う場合の系列長の決定法について調査する．

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．

文　献

[1] A. B. Kain, J.-P. Hosom, X. Niu, J. P. H. van Santen, M. F.-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 743–759, 2007.

[2] F. Rudzicz, "Acoustic transformations to improve the intelligibility of dysarthric speech," in *Proc. SLPAT*, Edinburgh, Scotland, Jul. 2011, pp. 11–21.

[3] S. Aryal and R. G.-Osuna, "Can voice conversion be used to reduce non-native accents?," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 7929–7933.

[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[6] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. ICASSP*, Taipei, Taiwan, Apr. 2009, pp. 3893–3896.

[7] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, Seattle, U.S.A., May 1998, pp. 285–288.

[8] D. Sunderman, H. Hoge, A. Bonafonte, H. Ney, A. W. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *Proc. ICASSP*, Toulouse, France, May 2006.

[9] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, Seattle, U.S.A., Jul. 2016.

[10] L. Sun, S. Kang, K. Li, and H. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 322–326.

[11] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences," in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 287–291.

[12] K. Cho, D. Bahdanau, F. Vougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. EMNLP*, Doha, Qatar, Oct. 2014, pp. 1724–1734.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NIPS*, Montreal, Canada, Dec. 2014, pp. 3104–3112.

[14] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma, "Dual learning for machine translation," in *Proc. NIPS*, Barcelona, Spain, Dec. 2016, pp. 820–828.

[15] W. Wang, S. Xu, and B. Xu, "First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention," in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2243–2247.

[16] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. NIPS*, Vancouver, Canada, Dec. 2006, pp. 153–160.

[17] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APSIPA ASC*, Jeju, Korea, Dec. 2016.

[18] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in

Proc. NIPS, Montreal, Canada, Dec. 2014, pp. 3581–3589.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. ICLR*, Banff, Canada, Apr. 2014.

[20] Y. Sagisaka, K. Takeda, M. Abe, S. Katagiri, T. Umeda, and H. Kuawhara, "A large-scale Japanese speech database," in *ICSLP90*, Kobe, Japan, Nov. 1990, pp. 1089–1092.

[21] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3–4, pp. 187–207, 1999.

[22] H. Kawahara, Jo Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *MAVEBA 2001*, Firentze, Italy, Sep. 2001, pp. 1–6.

[23] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.

[24] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, "The NAIST text-to-speech system for the Blizzard Challenge 2015," in *Proc. Blizzard Challenge workshop*, Berlin, Germany, Sep. 2015.

[25] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, Jul. 2011.

[26] B. Huang, D. Ke, H. Zheng, B. Xu, Y. Xu, and K. Su, "Multi-task learning deep neural networks for speech feature denoising," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 2464–2468.