舌亜全摘出者の音韻明瞭度改善のための中間層同士の関係を利用した知識蒸留方式の検討

Study of Knowledge Distillation Method using the Relationship between Middle Layers for Improving the Speech Intelligibility of Glossectomy Patients

高島 和嗣
Kazushi Takashima

岡山大学 阿部研究室
Abe Laboratory, Okayama University

概要 本研究では，舌亜全摘出者の音韻明瞭度改善のために知識蒸留を用いた方式を検討する．補助情報として音素ラベルを用いる方式を教師モデルとし，知識蒸留をおこなうことによって，合成時に音素ラベルを必要としない生徒モデルを作成する．教師モデルと生徒モデルの中間層の出力の誤差を学習に用いる．本報告では，教師モデルの知識をより強く引き継ぐために，中間層同士の関係を利用する方式の検討をおこなった．

1 はじめに

舌亜全摘出者は癌治療などのために舌を半分以上摘出した患者であり，構音機能に障害が残る．そのため，舌亜全摘出者の発する音声は健常者の音声よりも聞き取りづらく，音声によるコミュニケーションが困難である．

声質変換とは，ある特定の話者が発声した音声を，発話内容を保持しつつ，別の話者が発声した音声に聞こえるように変換を技術である [1]．これまで，GMM(Gaussian Mixture Model)[2] による声質変換を用いた改善方式や，DNN(Deep Neural Network) と差分スペクトル法を用いた改善方式 [3] によって，音韻明瞭度の改善が検討されてきた．しかし，舌摘出者の曖昧な発言による一対多変換の問題が存在するため，十分な音韻明瞭度改善は実現できていない．

一対多変換の問題とは，舌亜全摘出者の構音が不完全であるため，健常者のある音素が舌亜全摘出者の複数の音素に対応してしまう問題である．この問題を解決するために，補助情報として音素ラベル系列を用いる声質変換方式が提案されている [4]．評価実験の結果，音素ラベル系列を用いることによって，摩擦音等の子音の音韻明瞭度が大きく改善できることが示された．しかし，[4] の方式では変換時に，発話内容に対応する音素ラベル系列を事前に用意しておく必要があるため，実用的ではない．[4] の方式の変換モデルを教師モデル，合成時に音素ラベルを必要としない変換モデルを生徒モデルとし，知識蒸留 [5][6] をおこなうことによって，合成時に音素ラベルを必要としないが教師モデルと同等の性能の生徒モデルを作成する方式を検討してきた [7]．評価実験の結果，知識蒸留を行わない方式と比較して，摩擦音等の子音の音韻明瞭度が改善されていることが示された．しかし，[7] の方式では，教師モデルの知識を十分に引き継ぐことができていない．教師モデルの知識をより強く引き継ぐために，中間層同士の関係を利用する知識蒸留の方式 [8] を検討する．[8] では，中間層同士の関係を知識蒸留に利用することで，[6] の中間層の出力の損失を学習に利用する方式よりも良い性能を表し，教師モデルの知識をより強く引き継いでいることが示されている．本研究では，[8] の方式を利用することにより，教師モデルの性能により近くなるような生徒モデルの作成を目指す．

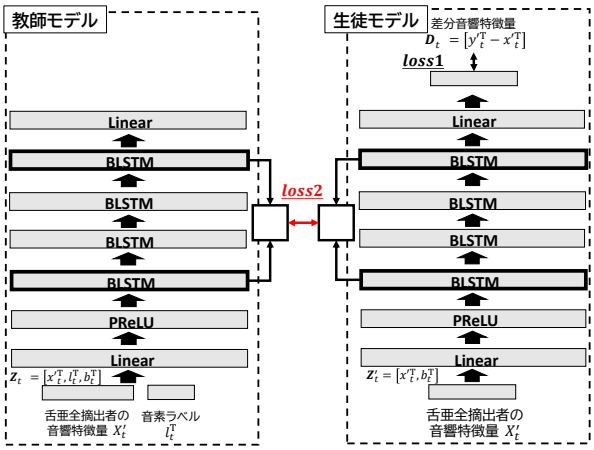

図 1: 提案方式

2 提案方式

提案方式の概要を図 1 に示す．本稿では，図 1 の太線で囲まれた 1 層目と 4 層目の中間層の出力から FSP Matrix を生成し，教師モデルと生徒モデルの FSP Matrix の誤差を loss2 として学習を行う．また，[7] で最も良い性能を示した loss1，loss2 を合計した値を最小化するように生徒モデル全体を学習する方法を用いて知識蒸留をおこなう．

3 評価実験

3.1 実験条件

音声データは 1 名の男性話者が発声したものであり，この話者が通常に発声した健常者音声と舌を固定する器具 [9] を装着して発声した疑似舌摘出者音声を用いて評価実験をおこなった．本稿での実験条件を表 1 に

表 1: 実験条件

データセット条件	
音声サンプリング周波数	20 kHz
音声分析	WORLD 音声分析
音声フレームシフト長	5 ms
学習データセット	ATR 音素バランス A-H セット 400 文
検証データセット	ATR 音素バランス I セット 50 文
評価データセット	ATR 音素バランス J セット 51 文
変換モデルに関する条件	
ネットワーク構成	
[I, L-128, PReLU, BLSTM-128, BLSTM-128, BLSTM-128, BLSTM-128, L-128, O-25]	
I: 入力層 (Input layer)	PLA なし: I-25, PLA あり: I-70
O: 出力層 (Output layer)	PLA: 補助情報としての音素ラベル
L: 線形層 (Linear layer)	
PReLU(Parametric Rectified Linear Unit) : 活性化関数	
(付随する数字はユニット数を表す)	
損失関数	平均二乗誤差
最適化手法	Adam ($\alpha = 0.001$, $\beta 1 = 0.9$, $\beta 2 = 0.999$)
ミニバッチサイズ	2 sentences

示す．発話様式については，学習データセットと検証データセットはフレーズ発声，評価データセットは連続発声を用いている．また，評価データである J セット 53 文の内，ノイズがあった音声を除いた 51 文を使用した．

3.1.1 客観評価実験

客観評価実験ではメルケプストラム歪みを用いて評価する．なお，評価に用いたメルケプストラムは変換音声を WORLD 分析して抽出した．評価実験として以下の 4 種類の声質変換方式を比較する．

- baseline : 音響情報のみを用いる方式
- PLA : 音響情報に加えて，補助情報として音素ラベルを用いる方式
- KD : PLA を教師モデルとして知識蒸留をおこなう方式 (4 層目の中間層の出力の誤差を利用)
- KD_flow : PLA を教師モデルとして知識蒸留をおこなう方式 (FSP Matrix の誤差を利用)

各声質変換方式の変換音声のメルケプストラム歪みを図 2 に示す．図 2 より，baseline を比較して，知識蒸留をおこなう KD_flow のほうがメルケプストラム歪みが小さくなっていることが分かる．これは，知識蒸留をおこなうことによって，教師モデルである PLA に近くなったためと考えられる．一方で，KD と KD_flow を比較すると，FSP Matrix を利用する KD_flow はメルケプストラム歪みが大きくなっていることが分かる．

これは，ニューラルネットワークでは低レイヤーはより一般的なデータの特徴，高レイヤーはよりタスク特有の特徴を含む傾向にあるため，FSP Matrix を用いることにより，低レイヤーの情報が高レイヤーの情報を曖昧にしたことが原因であると考えられる．

4 まとめと今後の課題

本報告では，舌亜全摘出者の音韻明瞭度改善のための中間層同士の関係を利用する知識蒸留方式について検討した．生徒モデルの性能向上のため，中間層同士の関係を表現する FSP Matrix を教師モデルと生徒モデルでそれぞれ求め，FSP Matrix の誤差を生徒モデルの学習に利用した．評価実験の結果より，従来の知識蒸留の方式と比較して，FSP Matrix の誤差を用いた知識蒸留による音韻明瞭度の大幅な改善は見られなかった．

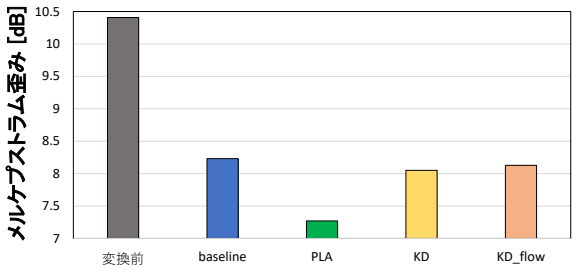

図 2: 各声質変換方式のメルケプストラム歪み

今後の課題としては，今回使用したデータセットではネットワークに対してデータ数が少なかった可能性が考えられるため，使用するデータ数を増加することが挙げられる．

参考文献

[1] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, “Voice conversion through vector quantization,” in *Proc. ICASSP*, pp. 655–658, Apr. 1988.

[2] 田中慧, 原直, 阿部匡伸, 皆木省吾, “GMM に基づく声質変換を用いた舌亜全摘出者の音韻明瞭性改善の検討,” 日本音響学会講演論文集, pp. 141–144, 2-5-8, Sep. 2016.

[3] H. Murakami, S. Hara, M. Abe, M. Sato, S. Minagi, “Naturalness Improvement Algorithm for Reconstructed Glossectomy Patient’s Speech Using Spectral Differential Modification in Voice Conversion,” in *Proc. INTERSPEECH*, pp. 2464–2468, Sep. 2018.

[4] 村上博紀, 原直, 阿部匡伸, “舌亜全摘出者の音韻明瞭度改善の Bidirectional LSTM-RNN に基づく音素補助情報を用いた声質変換方式の検討,” 日本音響学会講演論文集（春）, 2-P-32, Mar. 2019.

[5] G. Hinton, O. Vinyals, and J. Dean, ”Distilling the knowledgein a neural network”, arXiv preprint arXiv:1503.02531, 2015.2, 7

[6] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. ”Fitnets: Hints for thin deep nets”, In ICLR, 2015.

[7] 高島和嗣, 原直, 阿部匡伸, “音素情報を知識蒸留する舌亜全摘出者の音韻明瞭度改善法,” 日本音響学会講演論文集（秋）, 1-3Q-10, Sep. 2021.

[8] J. Yim, D. Joo, J. Bae, J. Kim, ”A gift from knowledge distillation: Fast optimization, network minimization and transfer learning”,Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , pp. 4133–4141, 2017

[9] H. Murakami, S. Hara, M. Abe, M. Sato, and S. Minagi, “Naturalness Improvement Algorithm for Reconstructed Glossectomy Patient’s Speech Using Spectral Differential Modification in Voice Conversion,” *Proc. INTERSPEECH*, pp. 2464–2468. (2018)