統計的食道音声強調における ポーズ位置不一致データを活用したモデル学習*

☆岸本真由美, 土井啓成, 戸田智基, Sakriani Sakti, 中村哲 (奈良先端大・情報)

1 はじめに

喉頭摘出者のための発声補助技術の一つとして，食道音声から通常音声への変換（Esophageal-Speech-to-Speech: ES-to-Speech）[1] が提案されており，その有効性が示されている．この手法では，食道音声と多数の健常者による通常音声の同一内容発話データを用いて，変換モデルの学習が行われる．品質の高い学習データを構築するためには，食道音声のポーズ位置に合わせた健常者音声の収録が有効であるが，各食道音声に対して健常者音声を再収録する必要があるため，多大な労力を有する．本報告では，食道音声とポーズ位置が一致していない健常者音声を，学習データとして効果的に使用する手法を提案し，実験的評価結果からその有効性を示す．

2 統計的手法に基づく食道音声強調

2.1 食道音声から通常音声への変換

ES-to-Speech では，食道音声のスペクトルセグメント特徴量から，健常者音声のスペクトル，F_0，非周期成分 [2] の推定を各々行う GMM を計 3 つ使用する．GMM の学習データには，同一内容発声の食道音声と健常者音声に対して，動的計画法によるフレームアライメントを行うことで得られるパラレルデータを用いる．動的計画法では，文単位でスペクトル距離尺度が最小となるフレームアライメントを求める．なお，スペクトル距離尺度の計算には，食道音声のスペクトルを直接用いるのではなく，学習された GMM に基づき変換されたスペクトルを用いる．GMM の学習処理とフレームアライメントを繰り返し行うことで，フレームアライメントの精度を向上させる [3]．

ES-to-Speech において，一対多 EVC を導入することで，食道音声の話者性を改善することができる．学習処理では，食道音声と同一内容発声の多数の健常者音声を用いて多数のパラレルデータを作成し，固有声 GMM（Eigenvoice GMM: EV-GMM）を学習する [4]．適応処理では，所望の声質を持つ健常者音声の少量かつ任意の発話データを用いて，EV-GMM の教師無し適応を行い，その声質へと変換する GMM が得られる．なお，一対多 EVC に基づく ES-to-Speech では，スペクトル推定および非周期成分推定を各々行う EV-GMM を計 2 つ学習する．一方で，F_0 推定に関しては，食道音声のピッチに対応した F_0 パターンを推定するために，食道音声のピッチを模して発声された健常者音声を別途収録し，それを用いて一対一 VC と同様の処理により F_0 推定用の GMM を学習する．適応処理においては，推定 F_0 パターンに対して対数領域において線形変換を行うことで，全体の平均および分散のみを適応する．

変換処理では，動的特徴量および系列内変動を考慮した最尤系列変換法 [5] を用いる．推定された F_0 及び非周期成分を用いて混合励振源を生成し，推定されたスペクトルを畳み込み，変換音声を合成する．

2.2 学習データにおけるポーズ位置の影響

学習データとして，ポーズ位置が一致していない同一内容発声の音声データを用いると，パラレルデータにおいて無音フレームと有音フレームの対応付けが行われる．このようなパラレルデータを用いて学習された GMM を用いると，無音フレームの音響特徴量から有音フレームの音響特徴量への変換や，その逆の変換などが生じることになる．ポーズ周辺の音素においても，異なる調音結合の影響を受ける音素間で対応付けが行われるため，不適切な対応関係が学習される．また，フレームアライメントは動的計画法により自動的に行うため，ポーズ位置以外の箇所のフレームアライメントの精度を劣化させる可能性もある．これらは変換音声の品質を大幅に劣化させる要因となる．この問題を防ぐために，従来では，ポーズ位置が一致した同一内容発声の食道音声と健常者音声のパラレルデータを使用している [1]．

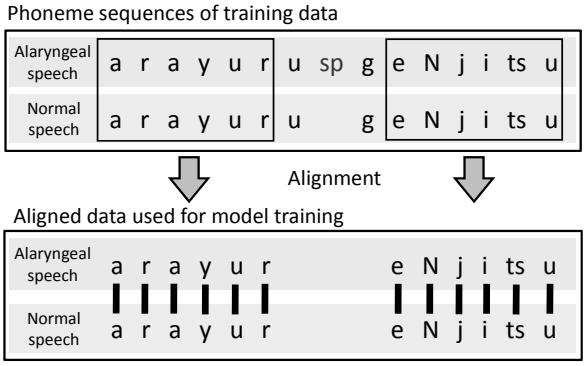

Fig. 1 Proposed process of removing mismatched pause and its preceding and suceeding phonemes.

3 ポーズ位置不一致データへの対応

ポーズ位置が不一致な音声データにおいても，ポーズから十分に離れた部分の音響特徴量は，ポーズの有無による調音結合の影響が十分に小さくなるため，GMM の学習データとして有効に利用できると考えられる．そこで，ポーズ位置不一致データを学習に使用するための手法として，不一致ポーズ及びその調音結合の影響を受ける周辺の音素を除外して，パラレルデータを構築する手法を提案する．

提案手法では，音素セグメンテーションを行うことで，除外するフレームを決定する．まず，元話者と目標話者の各発話データに対して，音素系列を用意する．本報告では，同一内容発話であるため，ポーズ以外の音素は等しいとして，ポーズ位置のみ各発話データに応じて人手で挿入する．次に，健常者音声データで学習された不特定話者用隠れマルコフモデル（Hidden Markov Model: HMM）を用いた Viterbi アライメントを行うことで，音素の時間情報を付与する．食道音声に対するアライメント処理では，ポーズ位置不一致データで学習された GMM に基づき健常者音声へと変換された音声を用いることで，アライメント精度の劣化を低減させる．得られた時間情報を元に，不一致が生じているポーズおよびその周辺音素に対応するフレームを削除する．ポーズ不一致箇所において分割された時系列データ毎に動的計画法によるフレームアライメントを行い，パラレルデータを構築する．例として，ポーズおよびその前後 1 音素を除外する処理を Fig. 1 に示す．

*Model training using training data including mismatched pause positions in statistical esophageal speech enhancement. by KISHIMOTO, Mayumi, DOI, Hironori, TODA, Tomoki, SAKTI, Sakriani, NAKAMURA, Satoshi (NARA INSTITUTE of SCIENCE and TECHNOLOGY)

4 実験的評価

4.1 実験条件

元話者の音声として男性喉頭摘出者 1 名の食道音声を収録し，ポーズ位置が一致した音声データとして男性健常者 23 名及び女性健常者 17 名の音声を収録する．この内，30 名分（男性 18 名，女性 12 名）を EV-GMM の学習用に使用し，残りの 10 名分（男性 5 名，女性 5 名）を評価用に使用する．また，ポーズ位置が一致していない音声データとして，男性健常者 15 名及び女性健常者 15 名の音声を，EV-GMM の学習用に使用する．発話内容は音素バランス文 50 文であり，内 40 文を学習用，残り 10 文を評価用とする．サンプリング周波数は 16 kHz とする．

スペクトル特徴量として，0 次から 24 次のメルケプストラム係数を用いる．食道音声に対しては，メルケプストラム分析 [6] を，健常者音声に対しては，STRAIGHT 分析 [7] をそれぞれ用いる．食道音声のスペクトルセグメント特徴量は，スペクトル推定及び非周期成分推定時においては当該フレーム及び± 8 フレームを用いて生成し，F_0 推定においては当該フレーム及び± 16 フレームを用いて生成する．スペクトルセグメント特徴量の次元数は 50 とする．また，音源特徴量として，STRAIGHT によって抽出された対数 F_0 と 5 帯域 (0-1, 1-2, 2-4, 4-6, 6-8 kHz) の非周期成分を用いる．シフト長は 5 ms とする．

客観評価実験では，提案手法の有効性をメルケプストラム歪みで評価する．ポーズ位置一致データ 4，8，15，30 名分の 4 パターン（*Matched*），ポーズ位置一致データ 30 名分とポーズ位置不一致データ 30 名分の計 60 名分（*Matched+Mismatched*），ポーズ位置一致データ 30 名と提案手法を適用したポーズ位置不一致データ 30 名分の計 60 名分（*Matched+DelPau*）の 3 種類 6 パターンのデータを用いて，EV-GMM の学習を行う．なお，提案手法での削除フレームは不一致ポーズとその前後 2 音素分とする．提案手法における音素セグメンテーションには Julius [8] の音素セグメンテーションキット，音響モデルは不特定話者用トライフォン HMM を用いる．

主観評価実験では，*Matched*（30 人分），*Matched+Mismatched*，*Matched+DelPau* の 3 種類の学習データを用いて生成された変換音声を用いて，音質および話者性に関する対比較実験を行う．音質に関する実験では，被験者に 2 種類の変換音声サンプルをランダムな順で提示し，より音質が高いサンプルを強制的に選択させる．話者性に関する実験では，被験者に目標話者音声サンプルを提示した後に，2 種類の変換音声サンプルをランダムな順で提示し，目標音声と話者性がより近いサンプルを強制的に選択させる．被験者は健聴者 10 名であり，防音室内にてヘッドフォン両耳受聴により，評価を行う．

4.2 実験結果

客観評価実験結果を Fig. 2 に示す．ポーズ位置一致データの結果（*Matched*）から，一対多 EVC に基づく食道音声強調において，学習に用いる話者数を増加させることで，変換精度を改善できることが分かる．この傾向は，*Matched+Mismatched* にも見られるが，その改善効果は小さくなる．一方で，提案手法によるフレーム削除を行うことで，より大きな改善効果が得られることが分かる（*Matched+DelPau*）．

主観評価実験結果を Fig. 3 に示す．ポーズ位置一致データのみを用いた場合（*Matched*）と比較し，提案手法を適用せずにポーズ位置不一致データを併用すると（*Matched+Mismatched*），音質が劣化している．併用時には，無音フレームと有音フレーム間で不適切な対応付けがなされたパラレルデータを用いて EV-GMM が学習されるため，独特な雑音が変換音声にしばしば混入し，音質劣化が生じる．一方で，併用時に提案手法を適用することで（*Matched+DelPau*），その問題を回避することができ，*Matched* と比較して，より音質が高く，話者性も改善された変換音声が得られることが分かる．

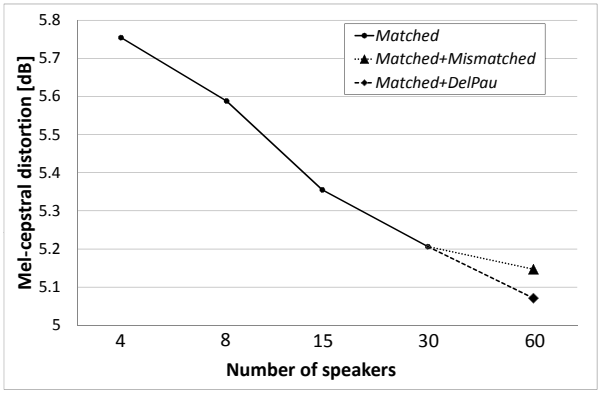

Fig. 2 Mel-cepstral distortion as a function of the number of pre-stored target speakers in one-to-many EVC for alaryngeal speech enhancement.

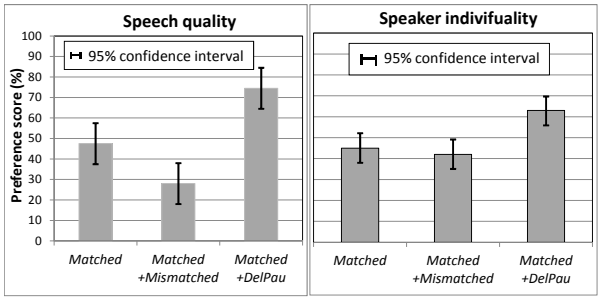

Fig. 3 Results of preference tests.

以上より，一対多 EVC に基づく ES-to-Speech において，提案手法を用いることで，ポーズ位置不一致データを効果的に使用することが可能となる．

5 おわりに

本報告では，食道音声とポーズ位置が一致していない健常者音声を，一対多 EVC に基づく ES-to-Speech におけるモデル学習に使用する手法を提案した．実験結果から，提案手法により，ポーズ位置一致データのみでなく，ポーズ位置不一致データも併用した変換モデル学習処理が可能となり，学習に使用する健常者音声データ数を増やすことで，変換音声の音質及び話者性が改善されることを示した．

謝辞 本研究の一部は，科研費補助金若手研究（A）により実施したものである．

参考文献

[1] Doi *et al.*, IEICE Trans. Inf.& Syst., E93-D (9), 2472-2482, 2010.
[2] Kawahara *et al.*, MAVEBA 2001, 2001.
[3] 戸田 他, 信学技報, SP2004-107, 67-72, 2004.
[4] Ohtani *et al.*, IEICE Trans. Inf.& Syst., E93-D (6), 1589-1598, 2010.
[5] Toda *et al.*, IEEE Trans. ASLP, 15 (8), 2222-2235, 2007.
[6] Tokuda *et al.*, Proc. ICSLP, 1043-1045, 1994.
[7] Kawahara *et al.*, Speech Commun., 27 (3-4), 187-207, 1999.
[8] Lee *et al.*, Proc. EUROSPEECH 2001, 1691-1694, 2001.