口唇口蓋裂者の音声認識のためのデータ拡張方式の検討 *

☆冨士原健斗, 高島遼一 (神戸大), 杉山千尋, 田中信和,
野原幹司, 野崎一徳 (大阪大), 滝口哲也 (神戸大)

1　はじめに

口唇口蓋裂は先天的な障害であり，口唇や口蓋 (口の中の天井部分) に割れ目が生じている症状を指す．言葉を発する際，人間は口内の空気の流れを変化させることによって発音する．そのため，口腔に関する疾患は発声に悪影響を及ぼし，口唇口蓋裂においても症例に応じた構音障害との関連が指摘されている [1]．Fig. 1 に健常者 (上図) と口唇口蓋裂者 (下図) の発話「一週間ばかり，ニューヨークを取材した」のスペクトログラムを示す．口唇口蓋裂者の音声の特徴として，空気が鼻腔に漏れてしまうことによる開鼻声が挙げられる．開鼻声は健常者の音声に比べて，低周波成分が強く，高周波成分が弱くなる傾向が指摘されている [2]．

音声認識システムは広く普及し，携帯電話やスマートスピーカーなど生活の様々な場面で利用されている．一般的な音声認識システムは健常者を対象として作られたものであるため，健常者と異なる特性を持つ口唇口蓋裂者の音声の認識は困難である．したがって，口唇口蓋裂者専用の音声認識システムを構築する必要がある．しかし，口唇口蓋裂者にとって，大量の学習データを収録することは大きな負担となる．また，収録した音声には人手でラベルを付与する必要があるが，口唇口蓋裂者は健常者音声に比べて聞き取りが難しいため，正確なラベルを付与することが困難である．したがって，音声認識システムの構築は健常者に比べて少量の学習データで行うことが求められる．

利用できるデータ量を増やす方法の 1 つとして,「データ拡張」が挙げられる．ここでデータ拡張とは，収録されたデータを加工し，新しいデータを生成することを指す．これにより，音声認識システムの学習データを増量し，精度を向上させる手法が存在する [3]．また，文献 [4] では，拡張後のデータから拡張前のデータを復元するモデルを事前学習することで，より良い音声の特徴表現を獲得する「自己教師あり学習」を提案しており，自己教師あり学習に加えてデータ拡張によるデータ増量を行うことで，さらに音声認識性能が向上することを示している．

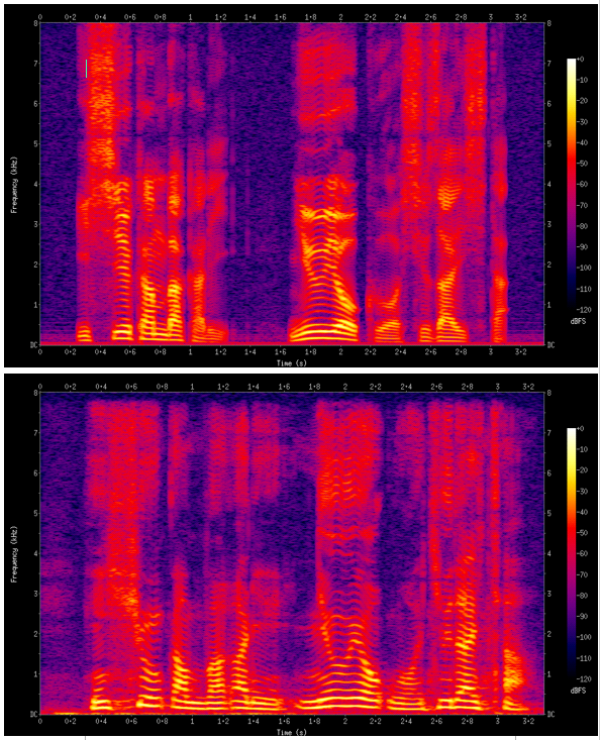

Fig. 1　Example of spectrogram uttered for /i q sh u: k a N b a k a r i ny u: y o: k u o sh u z a i sh i t a/ of a physically unimpared person (top) and a person with cleft lip and cleft palate (bottom).

健常者を対象にした従来のデータ拡張や自己教師あり学習の手法は，口唇口蓋裂者を対象にした場合も，一定の効果を発揮することが期待されるが，必ずしも口唇口蓋裂者の発話の性質に合わせた学習が行えるとは言えない．さらに音声認識システムの性能を改善するためには，低周波数帯域に集中する情報をより効果的に読み取ることが必要だと考えられる．

そこで，本研究では新しいデータ拡張の手法として，周波数方向にデータを変形する「周波数

* An investigation of data augmentation method for speech recognition of cleft lip and cleft palate, by Kento Fujiwara, Ryoichi Takashima (Kobe University), Chihiro Sugiyama, Nobukazu Tanaka, Kanji Nohara, Kazunori Nozaki (Osaka University), Tetsuya Takiguchi (Kobe University)

伸縮」を検討する．学習データの周波数情報を低域へ圧縮することで，口唇口蓋裂者の発話の性質がより強くなったデータを生成し，音声認識システムの頑健性を高めることができると期待される．また，同様の処理を自己教師あり学習に用いることで，更に性能を改善することができると期待される．周波数伸縮や従来のデータ拡張，自己教師あり学習を用いて特定話者音声認識モデルを学習し，音素認識で評価を行い，提案法の有効性を確認する．

2　関連研究

健常者の音声認識タスクのために提案されてきた，従来手法について詳述する．Daniel ら [3] は，音声認識時にデータ拡張を行う手法として SpecAugment を提案した．SpecAugment では，ニューラルネットワークに入力されるメルフィルタバンク特徴に対してデータ拡張を行う．データ拡張は 3 種類あり，時間方向にマスクをかける「時間マスク (Time masking)」，周波数方向にマスクをかける「周波数マスク (Frequency masking)」，基準点が移動するようにデータを伸縮させる「時間伸縮 (Time wapring)」を行う．マスクの幅や伸縮による移動量は，パラメータとしてデータセットに合わせて設定する．それぞれの処理は，音声セグメントの部分的な欠落，周波数情報の部分的な欠落，時間方向の変形に対応している．本来よりも認識が難しい学習データが生成されることで，音声認識モデルが頑健になり，性能が改善されることが報告されている．

Weiran ら [4] は，SpecAugment と同様のデータ拡張を利用した自己教師あり学習の手法を提案している．この手法では，エンコーダーデコーダーネットワークを，拡張されたデータから本来のデータを復元するというタスクで学習する．これにより，人手によるラベルを用いず，頑健に特徴を抽出するエンコーダーが得られる．学習を終えたエンコーダーを音声認識システムに組み込むことで，性能が改善されることが報告されている．また，音声認識システムの学習を行う際に SpecAugment によるデータ増量を行うことで，さらなる性能改善が報告されている．

本研究では，文献 [4] の手法をベースとして，さらに口唇口蓋裂者の音声の性質に適したデータ拡張方式について検討する．

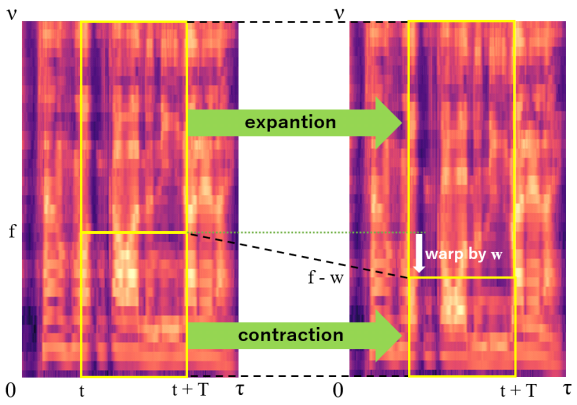

Fig. 2　Overview of our proposed method.

3　提案手法

3.1　周波数伸縮

Fig. 2 に，提案する周波数伸縮 (Frequency warping) の概要を示す．処理の手順は [3] に倣い，基準点となる次元と基準点の移動量を選んだ後，それに従ってデータを伸縮させる．ただし，低周波成分が強くなる傾向の度合いは発話中に変化する可能性があるため，音声認識モデル学習時のデータ手法として本手法を用いる際は，伸縮する時間帯を限定する．

入力は τ フレーム，ν 次元のサイズを持つメルフィルタバンク特徴とする．まず，処理の基準となる次元 f を $[W_{max},\ \nu]$ から一様乱数によって選択し，伸縮による f の移動量 w を $[W_{min}, W_{max}]$ から一様乱数によって選択する．W_{min}, W_{max} はパラメータとして設定する．次に，伸縮を行う時間帯の左端 t を $[0,\ \tau - T_{max}]$ から一様乱数によって選択し，時間帯の長さ T を $[T_{min}, T_{max}]$ から一様乱数によって選択する．$t+T$ が時間帯の右端となる．T_{min}, T_{max} はパラメータとして設定する．その後，選択された時間帯のデータを $T \times f$, $T \times (\nu - f)$ のサイズに二分する．それぞれを圧縮，伸長することで，$T \times (f - w)$, $T \times (\nu - f + w)$ のサイズに変形し，周波数方向に結合する．

以上の操作により，τ フレーム，ν 次元のサイズを持ち，低周波数帯域が圧縮されたデータが生成される．なお，伸縮の処理は Pytorch の関数 Image.resize によって行う．

3.2　自己教師あり学習

Fig. 3 に，今回提案する自己教師あり学習の概要を示す．エンコーダーデコーダーネットワーク

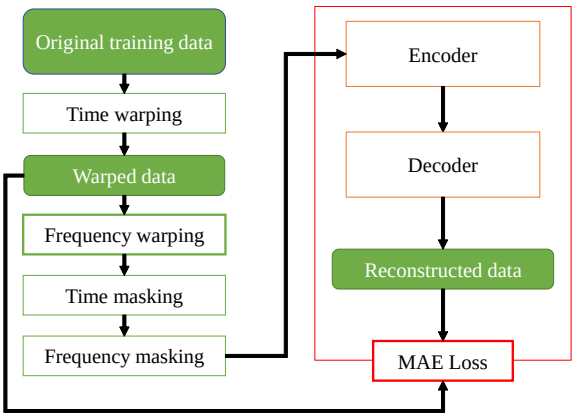

Fig. 3　Overview of Self-supervised Learning.

への入力データに対し，データのパターンを増やすために時間伸縮を行い，続いてタスクを生成するために周波数伸縮と時間マスク，周波数マスクを行う．損失として，再構成された出力と時間伸縮された直後の入力データを比較し，データ全体で平均絶対誤差を求めて学習を行う．自己教師あり学習を終えたエンコーダーは，パラメータを固定し，音声認識モデルに入力する特徴量抽出器として用いる．

なお，自己教師あり学習を行うにあたっては，可能な限り困難なタスクを生成することが性能改善に貢献すると考えられる．そこで，自己教師あり学習における周波数伸縮は，データ増強として用いる場合と異なり，時間帯を限定せずにデータ全体を対象にする．

4　評価実験

4.1　実験条件

データ拡張を用いた特定話者音素認識を行い，音素誤り率で評価した．評価話者として，口唇口蓋裂者男性 2 名 (SPK1/SPK2) を対象にした．データとして，ATR 研究用日本語音声データベース [6] に含まれる音素バランス文の読み上げ音声を，それぞれ 495 文と 503 文収録した．アノテーションが完了しているデータが一部しか存在しない状況を仮定し，音声認識システムの学習には 200 文のみを用いた．このうち 50 文を開発データ，50 文をテストデータ，100 文を学習データとした．エンコーダーの自己教師あり学習には残るデータを用い，このうち 50 文を開発データとし，話者 SPK1 は 245 文，話者 SPK2 は 253 文を学習データとした．自己教師あり学習には，ラベル情報を使用していない点に注意されたい．

Table 1　Parameters for each augmentation.

	ASR	Self-supervised
T_w	0〜200	0〜200
F_w	0〜20	0〜20
T_d	−50〜50	−150〜150
W_{min}, W_{max}	0〜2	0〜10
T_{min}, T_{max}	50〜100	τ 〜 τ

音声データのサンプリング周波数は 16kHz であり，音響特徴量として，フレームシフト 10ms，窓幅 25ms で抽出された 40 次元のメルフィルタバンク特徴を用いた．

音声認識システムには，音素を出力単位とする CTC を用いた [5]．モデルの中間層は 2 層の双方向 GRU で構成され，各層で入力フレームを 2 分の 1 にサブサンプリングした．出力層は音素 38 種類に未知音素を加えた 39 次元とした．バッチサイズは 5，初期学習率は 0.001 とし，最適化には Adam[7] を用いた．

自己教師あり学習を行う際に用いるエンコーダーデコーダーネットワークは，エンコーダーを双方向 LSTM4 層，デコーダーを全結合層 2 層で構成し，活性化関数として ReLU を用いた．バッチサイズは 10，初期学習率は 0.001 とし，最適化には Adam を用いた．

Table 1 に，音声認識システムの学習 (ASR) とエンコーダーネットワークの自己教師あり学習 (Self-supervised)，それぞれにおける各データ拡張の設定を示す．T_w は時間マスクの幅，F_w は周波数マスクの幅，T_d は時間伸縮の移動量に対応し，周波数伸縮と同様，処理を行う度に表記の範囲から一様乱数で大きさを決定した．話者 SPK1 と SPK2 で設定は共通とした．

4.2　実験結果

まず，自己教師あり学習を用いず，通常のメルフィルタバンク特徴で学習した音声認識システムの結果を Table 2 に示す．ベースラインとして，データ拡張を一切行わずに学習した場合の結果を示した (None)．データ拡張を行う場合は，各拡張を単独で用いた場合と，SpecAugment に相当する従来の 3 種類を組み合わせた場合 (3types)，従来手法に周波数伸縮を加えた 4 種類を組み合わせた場合 (4types) を比較した．

Table 2 PER [%] of the speaker-dependent cleft lip and palate model.

Augmentation	SPK1	SPK2
None	25.54	25.88
Time masking	22.83	23.13
Frequency masking	22.60	22.97
Time warping	20.78	21.85
Frequency warping	22.37	22.78
3types	19.85	19.95
4types	19.43	19.60

単独でデータ拡張を用いた場合，いずれもベースラインと比べて性能の改善が見られた．口唇口蓋裂者の音声特性を反映した周波数伸縮により，音声認識の性能が向上することが分かる．また，従来手法が口唇口蓋裂者の音声認識でも有効であることが分かる．時間伸縮が特に高い性能を示した要因として，今回使用した学習データでは，個々のスクリプトに対する発話が 1 回だけであったことが挙げられる．時間伸縮が，同じスクリプトで話速や個々の音素継続長が異なる発話を擬似的に生み出し，学習データの多様性を高めたと考えられる．

データ拡張を 3 種類や 4 種類で組み合わせて利用した場合，更に性能が改善した．周波数伸縮は，従来手法と同時に利用できる手法であることが分かる．

次に，自己教師あり学習によって抽出した特徴量で音声認識システムを学習した場合の結果を Table 3 に示す．話者ごとに，自己教師あり学習は 2 通りの方法で行い，周波数伸縮を用いず従来手法と同様に学習した場合 (baseline)，周波数伸縮を含めた 4 種類を用いて学習した場合 (ours) を比較した．また，音声認識を行う際にデータ拡張を利用する場合としない場合での性能を比較した．

baseline と ours では，いずれも ours がより高い性能を示した．自己教師あり学習においても，周波数伸縮は従来手法と同時に利用できる手法であることが分かる．

5 まとめ

本研究では，口唇口蓋裂者の音声認識のためのデータ拡張方式を検討した．健常者を対象にした従来手法が，口唇口蓋裂者の音声認識においても有効であることを確認した．新たな拡張の手法として周波数伸縮を提案し，口唇口蓋裂者の音声の性質に頑健な学習を行うことで性能改善が得られることを確認した．また，周波数伸縮は従来手法と同時に利用することで更に性能改善を得られる手法であることを確認した．今後の課題として，更に性能を高めるための新たなデータ拡張や，より適切な処理の手順について検討する．

Table 3 PER [%] of the speaker-dependent cleft lip and palate model using self-supervised learning.

	SPK1		SPK2	
Augmentation	baseline	ours	baseline	ours
None	25.74	24.61	23.90	22.93
3types	19.39	18.65	19.71	19.05
4types	19.16	17.96	19.44	17.85

参考文献

[1] 道健一，“口腔疾患による言語障害の診断と治療に関する臨床的研究,” 日本口腔科学会雑誌，35 巻 (4 号)，pp. 1035-1076，1986.

[2] 片岡竜太，“開鼻声の定量的評価法に関する研究 - 周波数特性と主観評価量との関連について - ,” 日本口蓋裂学会雑誌，13 巻 (2 号)，pp. 204-216，1988.

[3] D. S. Park *et al.*, “SpecAugment: A simple data augmentation method for automatic speech recognition,” arXiv preprint arXiv: 1904.08779，2019.

[4] W. Wang *et al.*, “Unsupervised pre-training of bidirectional speech encoders via masked reconstruction,” ICASSP，2020.

[5] A. Graves *et al.*, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” ICML，1990.

[6] A. Kurematsu *et al.*, “ATR Japanese speech database as a tool of speech recognition and synthesis,” Speech Communication，9，pp. 357-363，1990.

[7] D. Kingma，J. Ba，“Adam: A method for stochastic optimization,” ICLR，2015.