Vol.2021-MUS-131 No.19
Vol.2021-SLP-137 No.19
2021/6/18

唇画像からの音声生成における話者依存性の分析

金澤 尚希[1]　鈴木 基之[2]

概要：近年，様々な手法のサイレント音声技術が研究されているが，精度の改善が目下の課題である．これらの手法の内，本研究では唇画像からの音声生成に注目した．現在，DNN を用いて唇画像からスペクトル特徴量や基本周波数を推定して音声生成を行う手法では話者性が生成した音声の品質にどの程度の影響を与えるか考慮されていない．そのため，本研究では特定話者モデル，不特定話者モデル，話者適応モデルに評価実験を行った．その結果，話者適応モデルの正答率は 71%となり，不特定話者モデルより 10 ポイント程度向上することがわかった．

1. はじめに

近年，ディープニューラルネットワーク（DNN）やリカレントニューラルネットワーク（RNN）の発達により音声認識技術の精度は向上しつつある。しかし、音声認識技術には音を立てることが推奨されない静かな図書館や周囲の音を拾ってしまう人混みの中のような特定の環境下では使用することができないという弱点が存在する．

そこで最近，これらの問題を解決する方法の 1 つとして音声を使わずに発話内容を認識させるサイレント音声認識が研究されている．サイレント音声認識技術には発話する際の唇画像系列を用いた読唇システム [1] や発話する際の口周囲の表面筋電信号の変化を利用したもの [2]，発話する際の舌の動きを超音波診断装置を用いて取得し，それを利用するもの [3] など様々な手法が提案されている．しかし，これらのシステムは文献 [2] のように日本語 5 母音の認識で約 69.6%，文献 [3] のようにスマートスピーカーを操作するスピーチコマンドを用いての音声認識成功率で 65%ほどと認識精度が低く実用に耐えないものが多い．そのため，認識精度の向上が目下の課題である．

本研究では，サイレント音声認識技術の中でも読唇システムに注目している．この手法はカメラで撮影される映像をシステムの入力とするため，筋電信号や超音波などを利用する手法のように特別な機材が必要なく，実用化すればスマートフォンのような一般に普及している機材でも利用することができると考えられることが利点の一つである．

読唇システムには文献 [1] のように出力が単語列のものと文献 [4] のように出力が音声波形のものが存在する．出力が単語列の場合，出力する単語列をあらかじめ定義しておき，入力から定義した単語列に最も近いものを推定し出力とする．そのため，シチュエーションごとに使用されると予測できる単語のみを出力単語として定義することで認識精度を高めることができるという利点がある．しかし，出力単語として定義されているもの以外の推定はできず，推定結果が間違っていた場合，その結果から正しい情報を得ることは困難であるという欠点もある．

一方，出力を音声波形とする場合，入力からスペクトル，基本周波数，非周期性指標といった特徴量を推定し，それらから音声波形を合成して出力する．そのため，出力が単語列の場合のように推定結果が一意に定まったものを人が認識するのではなく，生成された音声を直接人が聞くことによって，発話内容を理解することができる．この場合，人が文脈などを考慮することで認識精度を向上させることができる可能性がある．しかし，合成された音声が必ずしも人が聞いて内容を理解できる音声になるとは限らないという欠点も存在する．

本研究では読唇システムを通話のように遠く離れた相手との会話を行うというような人間が出力を直接受け取ることを想定しているため，読唇システムの出力を単語列ではなく音声を出力する方法を採用している．

このような手法を用いた研究として文献 [4] がある．この研究で提案されているシステムは，まず LipNet[1] の構造を利用して唇画像系列からスペクトル特徴量系列を推定し，その結果からさらに，DNN を用いて基本周波数系列を推定する．そして，これらの推定され特徴量と入力の唇画像に対応する音声から抽出した非周期性指標系列を用いて音声波形を生成する．この研究では，同じ音素の発話をする際，話者によって口の形が異なるなどの話者によって

[1] 大阪工業大学大学院情報科学研究科
[2] 大阪工業大学

ⓒ 2021 Information Processing Society of Japan

異なる要素が認識精度に影響を及ぼす可能性があると考えられる．しかし，DNN を学習する際に M2TINIT[5] の男性 1 名で行っている．そのため，生成した音声の品質にどの程度の影響を及ぼすのか検証する必要がある．また，この研究では非周期性指標系列を元音声から抽出しているため，唇画像系列だけから音声生成できない．そのため，唇画像系列のみから音声を生成することができるように，非周期性指標系列の推定方法も考える必要がある．

そこで本研究では特定話者から学習した DNN モデル (特定話者モデル)，不特定多数の話者から学習した DNN モデル (不特定話者モデル)，不特定話者モデルのパラメータを一部，特定話者で再学習した DNN モデル (話者適応モデル) を用いてそれぞれのモデルから生成された音声の正答率を比較することで話者性が唇画像からの音声生成においてどの程度の影響を及ぼすか検証する．

2. 話者性を考慮した唇画像からの音声生成法

2.1 従来の手法

伊藤らの研究では唇画像系列から単語列を出力する LipNet[1] の構造を利用している．LipNet は文単位の読唇術を可能にした DNN モデルであり，伊藤らの研究では唇画像系列からスペクトル特徴量系列を推定する際に LipNet の構造を利用して回帰モデルとして作成している．このネットワークでは入力として RGB3 チャネルの画像系列を用いるので，CNN は 3 次元構造の STCNN(Spatio-Temporal Convolutional Neural Networks) を用いている．そのため，CNN に用いる Pooling 層や Dropout 層も 3 次元のものとなっている．具体的なネットワーク構成は図 1 である．3 次元畳み込み層を 3 つ連続し使用し，それぞれの畳み込み層は pooling 層により層を縮小している．その後に双方向 GRU 層を 2 層用いて，活性化関数として線形係数を用いることで出力層からメルケプストラム系列を得ている．

また，推定されたメルケプストラム系列から基本周波数を推定するためのネットワークは，素子数 256 の隠れ層を持つ LSTM を 2 層用いて構成されている．

これらのネットワークを用いて基本周波数系列とスペクトル特徴量系列を推定し，唇画像系列に対応する音声データから STRAIGHT[8] を用いて特徴量抽出した非周期性指標系列と合わせて音声合成を行う．伊藤らの研究ではこのように非周期性指標系列に関しては推定先の音声から抽出した値を用いているため，唇画像系列のみから音声生成することができない．また，この場合には推定したメルケプストラム系列や基本周波数系列と推定先の音声から抽出した非周期性指標系列では時間的なずれが生じて音声が乱れる可能性があるため，非周期性指標系列も併せて推定する必要があると考えられる．

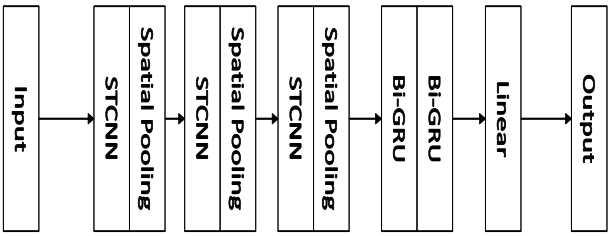

図 1 LipNet 構造を利用したネットワーク

2.2 音声生成方法の改良

本研究ではでは従来法から以下の点を変更している．

- 非周期性指標も入力された唇画像から推定
- 有声/無声区間の推定

2.2.1 非周期性指標の推定

2.1 節でも述べたように伊藤らの研究では唇画像系列のみから音声を生成することができないため，非周期性指標系列の推定方法も考える必要がある．本研究ではメルケプストラム系列から基本周波数系列を推定する方法と同様にメルケプストラム系列から非周期性指標系列を推定している．このようにすることで音声波形を合成する際に必要な要素をすべて唇画像系列から生成することができ，推定されたメルケプストラム系列と時間的同期がとれた非周期性指標系列が推定できると考えられる．

2.2.2 有声/無声区間の推定

本研究ではメルケプストラム系列から基本周波数系列を推定する際とメルケプストラム系列から非周期性指標系列を推定する際には併せて有声/無声区間の推定を行う．発話中の無声区間を推定することで有声音として生成することを防ぐことができると考えられる．無声区間はメルケプストラム系列や非周期性指標系列を推定する際と同様に隠れ層 256 の LSTM を 2 層用いて無声区間と有声区間の分類問題として推定している．そして，これらの推定結果それぞれと無声区間の推定結果を併せて新たに基本周波数系列と非周期性指標系列を求める．具体的には基本周波数の時系列の内，推定された無声区間の部分の値を無声区間を表す値に置き換えている．非周期性指標系列に関しても同様の手順で新たに非周期性指標系列を作成する．

2.2.3 システムの流れ

改良を加えたシステムの流れが図 2 である．まず，初めに唇画像系列からメルケプストラム系列を推定する．その際，唇画像系列からスペクトル特徴量系列を推定するネットワークに関しては従来の研究と同様に LipNet の構造を利用したものを使用している．このとき，入力に用いる唇画像 1 枚につき，1 フレームのメルケプストラムを生成するため，入力である唇画像系列と教師データであるメルケプストラム系列では 1 対 1 の対応をとる必要がある．そのため，唇画像系列に対して線形補間を行い，唇画像系列とメルケプストラム系列で 1 対 1 の対応がとれるようにしている．

次に，推定したメルケプストラム系列を入力として基本周波数系列，有声/無声区間，非周期性指標系列を推定する．その後，2.2.2 節で述べたように有声/無声区間と基本周波数系列，非周期性指標系列を合わせることで有声/無声区間を考慮した基本周波数系列と非周期性指標系列を生成する．そして，これらの推定されたメルケプストラム系列，基本周波数系列，非周期性指標系列を音声分析合成変換システム WORLD[9] によって音声波形に変換する．

2.3 推定モデルにおける話者性の考慮

本研究が提案しているシステムでは唇画像系列を入力とするため，特定話者モデルが不特定話者モデルより，唇画像の個人性の影響を大きく受けると考えられる．そのため，個人性が生成音声の正答率向上につながるのであれば，特定話者モデルのほうが不特定話者モデルに比べて正答率が高くなると考えられる．しかし，特定話者モデルを学習するために必要な十分なデータ量を用意することは困難であるため，不特定話者モデルのパラメータを一部特定話者で再学習した話者適応モデルも用いて性能比較を行う．

2.3.1 特定話者モデル

このモデルは特定の話者 1 名から図 2 の DNN すべてをを学習する．特定の話者 1 名から学習するため，その話者の個人性の影響を強く受ける．評価は学習に使用した特定の話者 1 名のデータを用いて音声を生成して行う．

2.3.2 不特定話者モデル

このモデルは多数の話者から図 2 の DNN すべてを学習する．多数の話者から学習するため，特定の個人の影響は受けづらく，複数の話者で共通する特徴を学習する．評価は学習に使用した話者以外の話者のデータを用いて音声を生成して行う．

2.3.3 話者適応モデル

このモデルは不特定話者モデルの重みの一部について再学習を行う．再学習に使用する不特定話者モデルはあらかじめ学習しているため，話者適応用データは特定話者モデルを大量のデータで学習するときに比べ，少ないデータ数で学習することができる．本研究で再学習する DNN は唇画像系列からスペクトル特徴量を推定するもののみで，他の DNN は不特定話者モデルと同じものを用いている．再学習する際には 2 つの双方向の GRU のみ学習を行う．これによって，時間的な特徴である口の動かし方について学ぶことを期待している．前段の CNN3 層に関しては発話内容による口の形を学習しており，そこに関しては話者性による影響が少ないと考えられるため，今回は再学習の対象としていない．

3. 読唇システムにおける話者依存性の調査

3.1 使用したコーパス

本研究で使用するコーパスは雑音環境下連続数字発話タスク評価データベース (CENSREC-1-AV)[6] である．コーパスの詳細は表 1 のようになっている．発話内容は 0〜9 の数字 7 桁以下で，それぞれの数字の読みは表 2 の通りになっている．音声データは音声区間の前後の非音声区間が 0.8 秒となるように切り出されたもので 0.8 秒に満たないものは可能な限り無音区間を確保している．このデータベースから，不特定話者モデルの学習用として 10 名，評価用話者としてそれ以外の 5 名を用いた．

表 1 コーパスの詳細

話者数	女性 15 名
発話数	1,155 発話 (77 発話/話者)
音声	1 種類 (クリーン音声)
画像	1 種類 (クリーンカラー)

表 2 数字の読み

単語	読み
1	/ichi/
2	/ni/
3	/saN/
4	/yoN/
5	/go/
6	/roku/
7	/nana/
8	/hachi/
9	/kyu:/
O	/maru/
Z	/zero/

3.2 画像データの補間

本研究で提案するシステムに入力する唇画像系列は唇周囲の 50 × 30 ピクセルを切り出したものを用いている．音声データとそれに対応する唇画像系列ではフレームレートが異なるため唇画像系列に線形補間を行っている．線形補間では画像間の時間が均等になるように約 6 倍の量に補間し，音声データと唇画像系列でフレームが 1 対 1 の対応がとれるようにしている．

3.3 モデルの学習に使用したデータ

3.3.1 特定話者モデル

このモデルでは、評価用話者と学習話者は同じであり，話者 1 人につき 1 つのモデルを学習した．学習に使用した発話数は 1 名ごとに 66 発話である．この 5 人の話者は CENSREC-1-AV の学習セットから不特定話者モデルに使用しない女性 5 人を使用する．評価データは学習に使用した話者と同じで学習に使用していない発話を使用した．

3.3.2 不特定話者モデル

10 名の学習話者のデータを用いて学習を行った．学習に

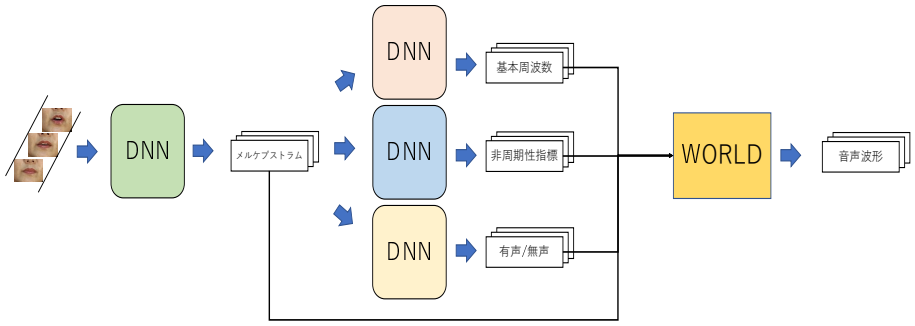

図 2 唇画像列から音声を生成する流れ

使用した話者の発話数はそれぞれの話者ごとに 77 発話で合計 770 発話である．

3.3.3 話者適応モデル

5 名の評価話者データそれぞれを用いて，不特定話者モデルの話者適応を行った．再学習に使用した発話数はそれぞれの話者ごとに 11 発話で 11 発話合わせてすべての数字を含むように発話を選択している．

3.4 生成した音声の聴取実験

本研究では唇画像系列から音声を生成する際に話者性が与える影響について検証するために，複数のモデルから生成される音声と加工をしていない元音声を用いて聴取実験を行った．それぞれの話者とモデルの組み合わせごとにすべての数字が最低二回以上含まれるように発話を選択している．生成した音声データは元音声も含めて 150 発話であった．実験参加者は 12 名 (男性 11 名，女性 1 名) で，静かな環境でヘッドフォンなどを着用して 20 発話分聴き，その音声が何と言っているか回答するという実験を行った．発話内容は 0〜9 の日本語数字で聞き取れない部分は判別不能として「X」と回答する．判別不能の「X」は発話全体の内，聞き取れなかった箇所のみを置き換える形で回答し，「X」の個数は実験参加者が音声を聴いた際に数字があると感じた回数記入する．1 人につき，それを最大 6 セットまで行い，結果として合計 840 発話分に対する回答を得た．

3.5 モデルごとの性能評価

モデルごとに正答率を比較した結果が表 3 である。正答率は式 1 を用いて計算される．例えば，表 4 のように推定結果が「5553Z54」で正解が「5453949」のようなデータが存在するとする．表 4 の「I」が挿入誤り，「D」が脱落誤り，「S」が置換誤り，「C」が正解である．

$$\text{正答率 (\%)} = \frac{\text{C の個数} - \text{I の個数}}{\text{C の個数} + \text{D の個数} + \text{S の個数}} \times 100 \qquad (1)$$

元音声の正答率は 100%でそれぞれのモデルで推定する音声の加工していない元音の音声は十分聞き取れる音声であることがわかる．モデル同士で正答率を比較すると，話者適応モデルが最もよい結果で 71%となった．逆に，不特定話者モデルが最も悪い結果で 62%となり，特定話者モデルは 65%となり不特定話者モデルよりは良いが話者適応モデルよりは悪い結果となった．

表 3 モデルごとの正答率の比較

モデル	正答率
特定話者	65%
話者適応	71%
不特定話者	61%
元音声	100%

表 4 誤りの例

正解データ	5	4	5	3	9		4	9
推定結果	5	5	5	3	Z	5	4	
数字の対応	C	S	C	C	S	I	C	D

不特定話者モデルの結果が最も低いことから話者性を考慮することで正答率が向上すると考えられる．話者適応モデルと特定話者モデルの結果については，話者適応モデルの方が正答率が高くなっているが，モデルの性質上，特定話者モデルのほうが話者 1 人のみから学習でき，口の動かし方や発話の速度といった話者性をより学習できていると考えられるため，逆の結果になると考えていた．実際はそのようにならなかったが，その原因として特定話者モデルの学習データが不足していたことが考えられる．話者適応モデルの再学習に利用している不特定話者モデルの学習データは 770 発話を用いているのに対して，特定話者モデルは 1 つのモデルあたり 66 発話しか用いていない．話者適応モデルは不特定話者モデルの学習の時点で発話内容による口の形などを学習できたが，特定話者モデルの方はデータ数が少なかったため発話内容による口の形などが十分に学習できていなかったと考えられる．この違いが特定話者モデルと話者適応モデルの正答率の逆転を起こしたと考えられる．

3.6 話者ごとの性能評価

話者ごとに正答率を比較した結果が表 5 である．モデルを区別せず正答率を計算した平均の正答率は FIH，FIG，FEI，FES，FEK の順に高くなっており，正答率が高い FIG，FIH の 2 人と正答率が低い FEI，FEK，FES の 3 人

表 5　話者ごとの正答率の比較

話者	平均	特定話者	話者適応	不特定話者
FIG	87%	94%	79%	87%
FIH	89%	95%	89%	78%
FEI	53%	40%	61%	69%
FEK	51%	62%	63%	30%
FES	52%	50%	58%	53%

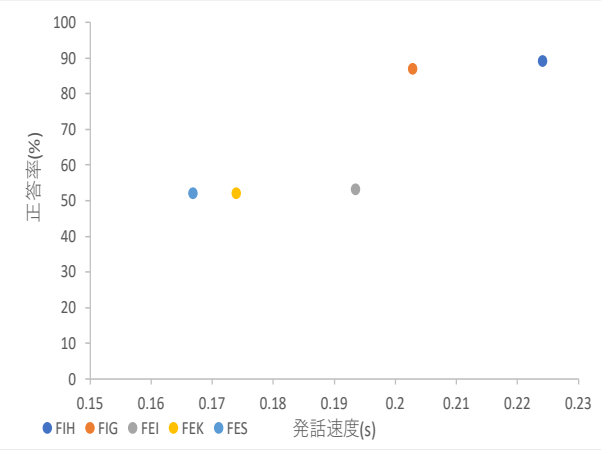

図 3　話者ごとの発話速度の比較

では 30 ポイント以上離れていることがわかる．

話者ごとに話し方などに違いがないか音声を聞き比べたところ，これらの音声では発話速度や数字の区切り方に違いがあると感じられたため，これらの正答率の違いについて発話速度と数字間の無音の挿入回数に注目して比較する．数字間の無音の挿入回数は多ければ数字をを 1 つずつ分けて発話していることがわかると考えられる．

3.6.1　話者ごとの発話速度の違い

本研究では発話速度を比較するために，発話する際の 1 モーラあたりの時間に注目した．これによって発話全体での発話速度ではなく，数字ごとにどの程度の時間をかけているか測ることができると考えられる．また，数字ごとにモーラ数が異なるため，1 数字あたり時間ではなく 1 モーラあたりの時間を採用した．図 3 は話者ごとの発話速度を比較したものである．総発話時間，総モーラ数，1 モーラあたりの発話時間は話者のテストデータの元音声から発話前後の無音区間を除いた区間から計測した値である．1 モーラあたりの時間は総発話時間を総モーラ数で割ることで求めている．

1 モーラあたりの時間について比較すると，正答率が高い順に 1 モーラあたりの時間が長くなっていることがわかる．このように 1 モーラあたりの時間が長ければ正答率が高くなる可能性があるが，FIG と FEI のように 1 モーラあたりの時間にに大きな差はないが，正答率は 30 ポイント以上離れているものも存在するため，この要素のみが正答率に影響するとは考えられない．

表 6　発話内に挿入された無音数

話者	正答率 (%)	挿入された無音の数
FIH	89	13
FIG	87	7
FEI	53	0
FES	52	1
FEK	51	3

3.6.2　発話内の無音挿入回数

本研究では発話内の無音挿入回数を音声認識システムの Julius を用いて計測している．Julius に用いられる音響モデルは dictation-kit のものを用いており，単語辞書は 0〜9 のみを定義している．また，文法規則として，数字 1 つから始まり数字の後には数字またはショートポーズが続くように定義している．このシステムによってショートポーズが挿入された回数を求め，それを数字間の無音の挿入回数としている．表 6 は発話内の無音挿入回数を話者ごとに比較したものである．

これを見ると，FIH や FIG のような正答率が高い話者は数字間の無音の挿入回数は多くなっていることがわかる．挿入回数が多いほど数字が区切られ前後の数字と繋がることで他の音に聴こえることが避けられると考えられる．逆に正答率が低い FEI，FES，FEK は無音の挿入回数が少なくなっている．そのため，数字間の無音の挿入回数は正答率に大きな影響を与える可能性がある．

これらのことより，1 モーラあたりの時間が長く数字間の無音の挿入回数が多い音声，すなわち，発話速度が遅く，数字を 1 つずつ区切った音声のほうが正答率が高くなる可能性があるとわかった．

4.　おわりに

本研究では，唇画像から音声生成を行うシステムにおいて特定話者モデル，不特定話者モデル，話者適応モデルを構築し，唇画像からの音声生成における話者性が及ぼす影響について分析を行った．実験の結果，話者適応モデルによる生成音声の正答率が不特定話者モデルによる生成音声の正答率を 10 ポイント上回ったことから，話者性を考慮することで正答率を向上させることができるとわかった．

話者ごとに正答率を比較した結果，話者によって正答率に大きな差があることがわかった．これについて，1 モーラあたりの発話時間と数字間に挿入されている無音の数について注目し分析した結果，発話速度が遅く，数字 1 つずつ区切った音声のほうが正答率が高くなる可能性があるとわかった．

これらのことより，日本語数字音声に対しては話者性を考慮することで正答率を向上させることができるとわかったが，一般日本語音声に関してはどの程度の影響があるか不明なため，今後は，一般日本語音声に対する話者性の影響を調べることが必要だと考えられる．

ⓒ 2021 Information Processing Society of Japan

参考文献

[1] Y. Assaeletal., “LipNet : end-to-end sentence-level lipreading,” arXiv:1611.01599, (2016)
[2] 福見 稔, “乾式電極を用いた口唇 EMG による個人認証” 電気学会・電子情報システム部門大会,(2018)
[3] 暦本 純一, 木村 直紀, 河野 通就,“SottoVoce:超音波画像と深層学習による無発声音声インタラクション” 情報処理学会 インタラクション，(2019)
[4] 伊藤 大貴，滝口 哲也，有木 康雄,“LipNet 構造を用いた唇画像から音声への変換” 日本音響学会研究発表会講演論文集，2018 号, ROMBUNNO.2-Q-30, (2018)
[5] 酒向 慎司, 徳田 恵一, 益子 貴史, 小林 隆夫, 北村 正,“HMM に基づいた視聴覚テキスト音声合成—画像ベースアプローチ” 情報処理学会論文誌, vol.43, no.7, pp.2169-2176, 2002)
[6] 田村 哲嗣, 宮島 千代美, 北岡 教英, 武田 一哉, 山田 武志, 滝口 哲也, 柘植 覚, 山本一公, 西浦 敬信, 中山 雅人, 傳田 遊亀, 藤本 雅清, 松田 繁樹, 小川 哲司, 黒岩眞吾, 中村 哲,“音下マルチモーダル音声認識評価基盤 CENSREC-1-AV の構築” 情報処理学会研究報告 [音声言語情報処理], vol.2010-SLP-82, no.7,(2010).
[7] 伊藤 大貴，滝口 哲也，有木 康雄,“CNN-LSTM を用いた唇画像から音声への変換” 日本音響学会 2017 年秋季研究発表会講演論文集, 2-Q-27, (2017)
[8] Kawahara, H., M asuda-Katsuse, 1. and do Cheveigne, A.(1999)“Restructur ing speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based FO extraction: Possible role of a reptitive structure in sounds,”Speech Communication,27 187 207.
[9] Morise, M., Yokomori, F. and Ozawa, K.: WORLD:a vocoder-based high-quality speech synthesis systemfor real-time applications,IEICE Trans. Inf. Syst.,Vol. E99-D, No. 7, pp. 18771884 (2016).

ⓒ 2021 Information Processing Society of Japan