Vol.2018-NL-235 No.2
Vol.2018-SLP-121 No.2
2018/5/13

日本語韻律構造を考慮した prosody-aware subword embeddingとDNN多方言音声合成への適用

高道 慎之介[1,a)]　秋山 貴則[2]　猿渡 洋[1,b)]

概要：統計的音声合成の急速な発展に伴い，日本共通語のみに留まらない次世代音声合成が必要とされている．その 1 つとして我々は，特定話者の声色であらゆる日本語方言の音声を合成する多方言音声合成の研究を進めており，本稿は，その韻律コンテキスト生成を扱う．韻律規則が古くから研究されている日本共通語では，規則及び辞書ベースの韻律コンテキスト生成が可能だが，多様な方言を扱う多方言音声合成において，各方言に対してそのような韻律規則を決定することは現実的ではない．一方，音声コーパスから韻律コンテキストを教師なしに推定する prosody-aware word embedding が提案され，英語音声合成における有効性が報告されている．しかしながらこの手法は，方言に含まれる未知語の韻律コンテキストを適切に生成できず，また，利用する韻律情報に過不足がある．これに対し本稿では，日本語韻律構造を考慮した prosody-aware subword embedding を提案する．日本語テキストは言語モデル尤度と日本語韻律構造に基づいて subword 系列に分割されるため，未知語を既知 subword 系列に効果的に分割できる．本稿では更に，この提案法を多方言音声合成に適用し，方言混合 subword tokenizer と多方言 subword embedding を提案する．実験的評価では，日本共通語および 20 方言の音声合成において提案法の有効性を示す．

Shinnosuke Takamichi[1,a)]　Takanori Akiyama[2]　Hiroshi Saruwatari[1,b)]

1. はじめに

統計的音声合成 [1] は統計モデルを使用して音声を合成する方法であり，人間と計算機の自然な音声コミュニケーションを可能にする技術である．特に，近年の深層学習 (deep neural network: DNN) 技術 [2], [3] は，主要言語のテキストを高音質に読み上げることを可能にした．この進展に伴い，音声なりすまし検出 [4] や音声認識 [5], [6] との統合，発話間変動の考慮 [7] やユニバーサル音声合成 [8] など，次世代音声合成に向けた研究が広くすすめられている．

次世代音声合成として期待されるものの１つが方言音声合成である．方言による音声コミュニケーションでは，声色などの話者性のみならず，方言の地域性が語彙・発音・韻律等に付与される．この地域性を考慮した音声合成技術は，異なる方言話者間の円滑なコミュニケーション [9], [10] や，合成音声への新たなキャラクタ性の付与にも応用可能である．これに対し我々は，話者性と地域性を分離する多方言音声合成に関する研究を進めている．具体的には，図 1 に示すように，あらゆる方言テキストに対し，特定話者（もしくは，方言と独立に制御される話者）の音声を合成する技術である．話者性と地域性を分離することで，それらを独立に制御できる柔軟な音声合成が可能になる．また，複数方言の音声情報を一括にモデル化することで，モデルパラメータ共有・適応 [11], [12] に基づく地域性の制御・補間が可能になると予想される．これは，方言の音声言語処理の条件である (1) 少量の方言音声資源のみで学習できることと (2) 異なる方言に容易に接続できること [9] を満たすための重要な技術である．これらの利点を持つ多方言音声合成の実現に向け，本稿では，方言間で異なる語彙・発音・韻律情報のうち，韻律情報を扱う．

韻律規則が古くから研究されている日本共通語では，規則及び辞書ベースの韻律コンテキスト生成が可能である．しかしながら，多様な方言を扱う多方言音声合成では，各

[1] 東京大学 大学院情報理工学系システム情報学専攻, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[2] 東京大学 計数工学科, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[a)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[b)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

方言に対してそのような韻律規則を決定することは現実的ではない．そこで我々は，多方言音声コーパスから，韻律コンテキストを自動抽出する方法を試みる．この実現には (1) 多方言音声コーパスの収集と (2) コンテキスト生成モデルの構築が必要である．(1) に関して我々は，クラウドソーシングによる音声コーパス収集を実施し，CPJD コーパスを構築した [13]．これは，ネイティブ方言話者により作成・発話されたテキスト，音声，及び各方言の地理情報からなるコーパスである．(2) に関して Ijima ら [14] は，音声コーパスから韻律コンテキストを教師なしに自動推定する prosody-aware word embedding を提案した．この手法では，各単語からそれに対応する韻律情報を予測する DNN を学習し，その bottleneck 特徴量を韻律コンテキストとする．この手法は英語音声合成における有効性が確認されているが，未知語に対する韻律コンテキストを適切に生成できず，既知単語から未知単語への適応も困難である．故に，方言固有の単語に対する学習が困難となり，また，多方言音声合成における語彙数増加に伴い学習が困難になる．更に，用いられる韻律情報の複雑さは予測元の単語の複雑さによらず一定であるため，韻律情報に過不足が生じる．

これらに対し本稿では，日本語韻律構造を考慮した prosody-aware subword embedding を提案する．日本語テキストは，部分文字列の言語モデル尤度と日本語アクセント句境界に基づいて subword 系列に分割されるため，未知語を既知 subword 系列に効果的に分割できる．また，学習時に，subword 内モーラ数を考慮した変調フィルタリングを行うことで，過不足のない韻律情報を利用する．本稿では，この手法を日本共通語に適用した後，多方言音声合成のための韻律コンテキスト生成に拡張する．多方言音声合成における韻律コンテキスト生成では，方言間で共有する方言混合 subword tokenizer と，方言情報により条件づけられる多方言 subword embedding を提案する．方言混合 subword tokenizer は，多方言コーパスの言語モデル尤度と韻律構造を考慮して学習され，多方言 subword embedding は，各方言の韻律情報を単一モデルで表現する．実験的評価では，日本共通語および 20 の方言の音声合成において提案法を評価する．その結果，(1) 日本共通語の音声合成において，提案法は従来法の音質を上回ること，また，(2) いくつかの方言において，方言混合 subword tokenizer と多方言 subword embedding は，日本共通語の韻律コンテキストの使用より自然な方言アクセントを生成できることを示す．

2. Prosody-aware word embedding

DNN 音声合成 [15] では，テキストから抽出された特徴量（コンテキスト）から音声特徴量を予測する DNN を構築する．このコンテキストは，音素などの音韻コンテキスト，アクセント等の韻律コンテキスト，また，時間位置を表す継続長コンテキストなどが含まれる．韻律規則が広く研究されている日本共通語などにおいて，この韻律コンテキストは，規則または辞書ベースなどにより推定される [16]．

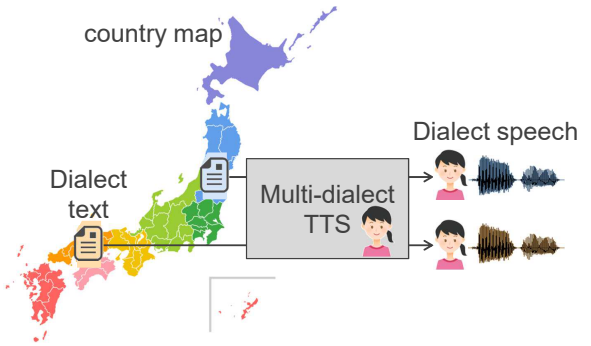

図 1 多方言音声合成の概要．特定話者（もしくは，方言と独立に制御される話者）の声色で，指定された方言とそのテキストから音声を合成する．TTS は text-to-speech の略．

Prosody-aware word embedding [14] は，単語列と F_0 系列から韻律コンテキストを教師なしに推定する方法である．この手法ではまず，各単語と連続 F_0 系列をアライメントし，単語毎の連続 F_0 系列が固定長となるよう，F_0 をリサンプリングする．その後，単語の one-hot ベクトルから，リサンプリング後の F_0 系列の離散コサイン変換（DCT）の低次成分を推定する DNN（以降，embedding モデル）を学習する．韻律コンテキストはこの bottleneck 特徴量として得られる．この手法は，単語数の増加に応じて学習が困難になり，また，方言固有の単語などの未知単語へのモデル適応が困難である．更に，シラブル数（日本語の場合はモーラ数）が単語毎に異なるにも関わらず固定次数の DCT 成分を利用しているため，不要な F_0 情報をモデル化，もしくは，必要な F_0 情報を無視している（例えば，'a' と 'linguistic' の F_0 情報を同程度使用することは，明らかに不自然である）．

3. 日本語韻律構造を考慮した prosody-aware subword embedding

提案法の構成を図 2 に示す．入力テキストは，部分文字列の言語モデル尤度と日本語アクセント句境界を考慮して subword 系列に教師なし分割されるため，提案法は，従来法における未知語問題を緩和できる．また，各 subword に対応する連続 F_0 系列は，固定長にリサンプリングされた後に subword 内モーラ数に応じて変調フィルタリングされるため，提案法は，各 subword に応じた適切な F_0 情報を利用できる．

3.1 日本語アクセント句境界を考慮した subword tokenizer

Subword 分割 [17] は，生文または単語列における低頻出語を部分文字列に教師なし分割することで，未知語問題を緩和する手法である．本論文では，生文に対する言語モデ

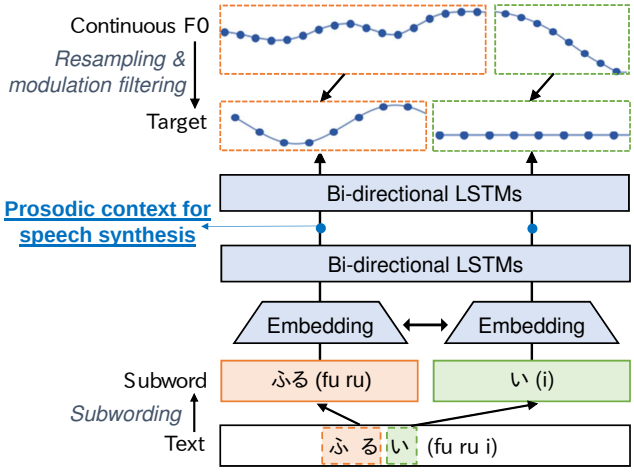

図 2 Prosody-aware subword embedding の構成．Subword 系列から，リサンプリングと変調フィルタリングを施した連続 F_0 を予測し，その bottleneck 特徴量を韻律コンテキストとする．LSTM は，long short-term memory の略．

表 1 Subword 分割の例．'/' はアクセント句境界．アクセント句を考慮することで，アクセント句をまたぐ subword 分割を防ぐ．

生文	本当な/のかも/しれない
subword	本当　なの　かもしれない
subword（アクセント句境界を考慮）	本当　な　の　かも　しれない

ル尤度に基づく教師なし subword 分割法 [18] を適用する．しかしながら，言語モデルに基づく分割は音声情報を無視するため，構築された subword tokenizer は，複数のアクセント句にまたがる subword 分割を行う（表 1 中段）．この subword から韻律コンテキストを生成する場合，複数のアクセント句に対応する韻律コンテキストを単一の subword を用いて表現することになるため，推定精度が低下する．そこで本稿では，アクセント句境界を考慮した subword 分割法を提案する．具体的には，subword 分割の学習データのアクセント句境界を既知として，アクセント句境界を超える部分文字列を言語モデルの計算から除外する．この処理により，subword 分割時にはアクセント句への事前トークナイズを必要とせずに，アクセント句境界にまたがる文字列を積極的[*1]に部分文字列に分割する（表 1 下段）．

3.2 Subword 内モーラ数を考慮した変調フィルタリング

日本語はモーラ等時性言語であり，モーラ毎に変化する高低のアクセントを持つ．故に，連続 F_0 系列を subword 毎に予測する際に，subword 内モーラ数だけの等時間間隔位置における高低アクセント以外の微細構造は不要である．この微細構造の除去は，連続 F_0 系列に対する変調フィルタリング（変調スペクトル [19] に対するフィルタリング）で実現できる．ここで，m モーラの

[*1] 必ず分割するわけではない．これは，分割時に事前トークナイズを行わないためである．

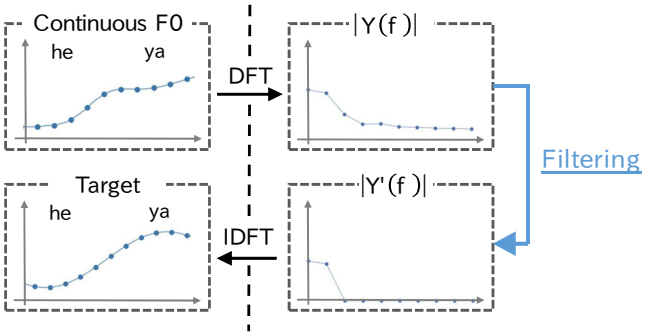

図 3 変調フィルタリングの例（2 モーラの場合）．DFT は discrete Fourier transform の略．

subword に対応する，T フレームの連続 F_0 系列を考える．この連続 F_0 系列の変調スペクトル（音声パラメータ時系列のフーリエ変換）を $[Y(0),\cdots,Y(f),\cdots,Y(T-1)]^\top$ とする．f は，変調周波数インデックスである．これに対し，モーラ数に応じて不要な成分を削除するフィルタ $C = [C(0),\cdots,C(f),\cdots,C(T-1)]^\top$ を設計する．$C(f)$ は，次式で与えられる．

$$C(f) = \begin{cases} 1 & (f \le f_{\text{th}} \text{ or } f \ge T - f_{\text{th}}) \\ 0 & (\text{otherwise}) \end{cases} \quad (1)$$

$$f_{\text{th}} = \begin{cases} 0 & (m = 1) \\ \frac{m+1}{2} & (\text{otherwise}) \end{cases} \quad (2)$$

フィルタ後の変調スペクトルは $Y(f)' = Y(f)C(f)$ で与えられ，embedding に用いる F_0 系列は，この変調スペクトルの逆離散フーリエ変換で得られる．図 3 に 2 モーラ（$m = 2$）の場合の例を示す．2 次以上の変調周波数成分を削除することで，各モーラ位置に相当する高低の情報を保持していることが分かる．上記の subword embedding には，subword 系列と連続 F_0 系列のアライメントが必要である．本稿では，文字列と音素列のアライメントと，音素列と連続 F_0 系列のアライメントを独立に計算しておくことで，subword と連続 F_0 系列のアライメントを得る．アライメント後，subword 系列から変調フィルタリング後の連続 F_0 系列を予測する embedding モデルを学習する．この際，embedding モデルとして recurrent neural network (RNN) を用いることでアクセント結合を考慮する．

4. 多方言音声合成への適用

3 節の prosody-aware subword embedding を多方言音声合成に拡張するために，全ての方言で共有する方言混合 subword tokenizer と，方言により条件づけられる多方言 subword embedding を提案する．

4.1 方言混合 subword tokenizer

共通語と方言で言語モデルの傾向が異なるため，日本共通語の学習データから学習された 共通語 subword tokenizer

表 2 標準語 subword tokenizer に含まれず，方言混合 subword tokenizer に含まれる subword の例

subword	左の subword を含む方言
どす	京言葉
ずら	遠州弁
だば	津軽弁，秋田弁
やけん	土佐弁，阿波弁，伊予弁，福岡弁

は，方言テキストに頻出する単語・フレーズを部分文字列に分割してしまい，後段の embedding モデルの韻律予測性能を低下させる．これに対し本稿では，日本共通語（東京方言）及び多方言音声コーパスから，単一の subword tokenizer を学習する．この方言混合 subword tokenizer は，表 2 に示すように，方言に頻出するフレーズを subword として保持する．

4.2 多方言 subword embedding

方言混合 subword tokenizer により分割された subword とそれに対応する F_0 から，多方言 subword embedding モデルを学習する．このとき，複数方言の韻律コンテキストを一括にモデル化するために，embedding モデルを方言情報により条件付ける．具体的には，隠れ層における最初の RNN （図 2 における，bottleneck 特徴量を抽出する前の LSTM）に対して，方言インデックス d に対応するベクトル $\vec{q}_d$ を入力する．このベクトルとして本稿では，(1) 方言コード，または (2) 地理情報を用いる．

4.2.1 方言コード（離散表現）

話者コードを用いる方法 [11] と同様に，方言を one-hot ベクトルで離散的に表現する．$\vec{q}_d$ は，d 次元目の値を 1，それ以外を 0 とするベクトルであり，その次元数は学習データに含まれる方言数に等しい．

4.2.2 地理情報（連続表現）

各方言は固有のアクセントを有するが，地理的に隣接した方言同士は同様のアクセントカテゴリに分類されることが多い [20]．そこで $\vec{q}_d$ を，当該方言の利用地域における中心都市の地理緯度・経度からなる 2 次元の連続値ベクトルとする．

5. 実験的評価

5.1 実験条件

Subword tokenizer の学習には sentencepiece [18] を利用し，subword 語彙数は未知語タグを含む 4,000 とする．日本共通語のアクセント句境界は，open jtalk [21] を用いて推定する．音声データのサンプリング周波数は 16 kHz，音声分析のフレームシフトは 5 ms とする．Embedding モデル学習時には，話者の違いを正規化するため，連続 F_0 系列を発話毎に平均 0，分散 1 に正規化する．リサンプリング後の連続 F_0 の系列長は 64 とする．従来法 [14] においては 1 次から 10 次の DCT 係数を予測する．従来法において，語彙数の爆発に伴う合成音声品質の極端な劣化が認められたため，従来法においても word embedding ではなく subword embedding を行う．Subword と音素のアライメントには，fast_align [22] を利用した．Embedding モデルは，2 つの bi-directional LSTM と，それらを接続する bottleneck 層の Feed-Forward neural network で構成される．各 LSTM のユニット数は 256 であり，bottleneck 層の活性化関数は ReLU [23] である．Bottleneck 特徴量の次元数は 64 とする．最適化手法には，Adam [24] を用いる．

表 3 実験に使用した方言名と都道府県の一覧．地域毎に分けて示している．太字は主観評価実験に使用した方言を指す

北海道	**北海道弁 (1)**
東北	津軽弁 (2)，秋田弁 (3)，いわき弁（4)
関東	埼玉弁 (5)
中部	金沢弁 (6)，**遠州弁 (7)**，福井弁 (8)
近畿	**京言葉 (9)**，**奈良弁 (10)**，**大阪弁 (11)**
中国	**岡山弁 (12)**，出雲弁 (13)，**広島弁 (14)**
四国	**阿波弁 (15)**，**土佐弁 (16)**，**伊予弁 (17)**
九州	**福岡弁 (18)**，**宮崎弁 (19)**，諸県弁 (19, 20)
沖縄	なし

音響モデルは，3 層の隠れ層からなる Feed-Forward neural network である．隠れ層のユニット数は 512，活性化関数は ReLU である．入力コンテキストは，通常用いられる 190 次元の quin-phone と 3 次元の音素内継続長ベクトルに加え，前後及び当該 subword の韻律コンテキスト（計 192 次元）と，9 次元の当該 subword 内継続長ベクトルを使用する．継続長は，自然音声のものを利用する．予測する音声特徴量は，0 次から 39 次のメルケプストラム係数，5 周波数帯域における平均非周期成分，連続 F_0，有声／無声ラベル，及びそれらの動的特徴量から成る 94 次元ベクトルである．学習時には，音声特徴量を平均 0，分散 1 に正規化する．

5.2 Prosody-aware subword embedding の評価

まず，日本共通語において 2 節の従来法 [14] と 3 節の提案法の合成音声を比較する．subword tokenizer と subword embedding の学習データは，新聞記事読み上げ音声コーパス（JNAS）[25] 15,676 文及び JSUT コーパス [26] 5,390 文である．音響モデルの学習データは，JSUT コーパス 5,390 文である．評価データは，学習データに含まれない JSUT コーパス 600 文である．比較手法は以下の通りである．

(1) **Conventional:** 2 節の従来法 [14]
(2) **Proposed:** 3 節の提案法（変調フィルタリングのみ）
(3) **Proposed (acc):** 3 節の提案法（変調フィルタリング＋アクセント句境界の考慮）

5.2.1 客観評価

F_0 系列の予測精度を比較するために，各手法の合成音声と自然音声の連続対数 F_0 の root mean squared error

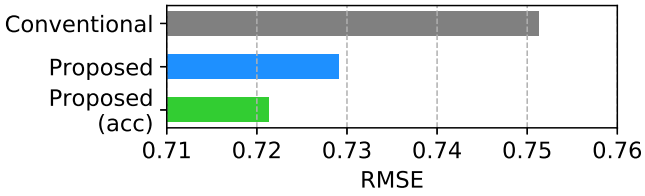

図 4 自然音声と合成音声の連続対数 F_0 の RMSE．従来法と比較して，提案法の RMSE が小さいことを確認できる．

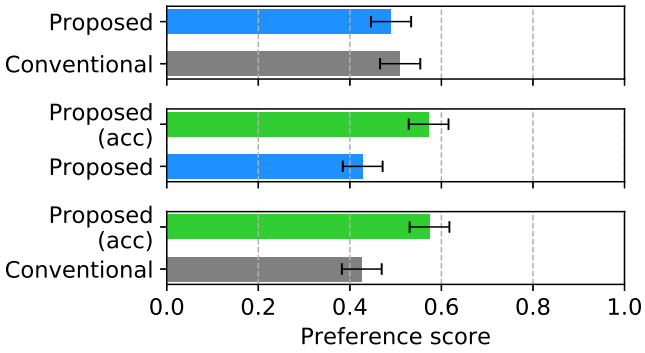

図 5 合成音声の自然性に関する主観評価結果（エラーバーは 95 %信頼区間）．提案法のスコアは従来法を上回ることが分かる．

(RMSE) を計算した．ただし，計算には平均 0，分散 1 で正規化した連続対数 F_0 を使用した．図 4 に結果を示す．従来法と比較して，提案法の RMSE が小さいことを確認できる．また，その値は，アクセント句境界を考慮することで更に小さくなることが分かる．

5.2.2 主観評価

合成音声品質に関するプリファレンス AB テストを実施した．図 5 に結果を示す．評価人数は，各評価において 50 人である．アクセント句境界を考慮しない場合，提案法と従来法に有意差は見られないが，アクセント句境界を考慮することにより，合成音声品質が有意に向上することが分かる．以上の結果より，日本共通語において，提案する prosody-aware subword embedding の有効性が確認された．

5.3 方言混合 subword tokenizer と多方言 subword embedding の評価

多方言音声合成において 4 節の提案法を評価する．Subword tokenizer と subword embedding の学習データは，JNAS・JSUT コーパス，更に，CPJD コーパス [13] のうち音素と F_0 のアライメントを取得できた，表 3 の 20 方言，4,534 文である．JNAS と JSUT は，東京方言として扱う．これらの学習データでは方言毎にデータ量の偏りがあるが，subword tokenizer の学習時においてデータ量に応じた言語モデルの重み付けは行わない．音響モデルは，5.2 節と同一である．評価として，合成音声品質に関するプリファレンス AB テストを実施した．この評価者として，当該方言の利用地域の在住歴が 3 年以上であり各方言アクセントを十分に判断できる者を方言毎に 1 名確保した．ただし，CPJD コーパスに含まれる全ての方言に対する評価者を確保することは困難であったため，表 3 に太字で示す 12 方言に対してのみ評価を実施した．評価データは，各方言に対して CPJD コーパス 20 文である．このデータは，学習データと重複せず，CPJD コーパスからランダムに選ばれた．

表 4 各方言の合成音声の自然性に関するプリファレンステストの集計結果．中央の 2 つの数字は，手法 A または B のプリファレンススコアが，もう一方のスコアより高かった方言の数を表す．

A	A is better	B is better	B
Common	8	4	Dialect code
Common	7	5	Geography
Dialect code	5	6	Geography

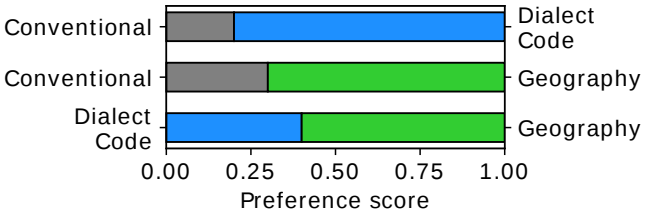

図 6 合成音声の自然性に関するプリファレンススコア (宮崎弁)．

比較手法は以下の通りである．

(1) **Common:** 共通語 subword tokenizer & subword embedding (5.2 節の “Proposed” に対応)
(2) **Dialect code:** 方言混合 subword tokenizer & 多方言 subword embedding (方言コード)
(3) **Geography:** 方言混合 subword tokenizer & 多方言 subword embedding (地理情報)

共通語韻律コンテキスト（“Common”）では，アクセント句境界の考慮が可能（5.2 節の “Proposed (acc)” に相当）だが，多方言におけるアクセント句境界の推定が未解決課題であるため，全ての比較手法で変調フィルタリングのみを適用した．

評価結果の集計を表 4 に，結果の一例として宮崎弁におけるプリファレンススコアを図 6 に示す．いくつか（方言コードは 4 つ，地理情報は 5 つ）の方言において，“Common” よりも提案法の音質が主観的に高いと判断された．以上より，いくつかの方言において，提案法による方言韻律コンテキストが，共通語韻律コンテキストより自然な方言アクセントを生成できることが明らかになった．また，方言コードと地理情報を比較すると，1 方言において，地理情報を用いた合成音声品質が高いと判断された．

6. まとめ

本稿では，音声合成の韻律コンテキストを教師なしに推定するために，日本語アクセント句境界を考慮した subword

tokenizer と subword 内モーラ数を考慮した変調フィルタリングに基づく，prosody-aware subword embedding を提案した．また，この提案法を多方言音声合成に拡張し，方言混合 subword tokenizer と，多方言 subword embedding を提案した．実験的評価では，日本共通語と日本語多方言音声合成において有効性を確認した．今後は，embedding の改善，及び，多方言音声合成における音響モデルの学習法を検討する．

謝辞： 本研究の一部は，セコム科学技術支援財団，JSPS 科研費 17H06101，18K18100 の助成を受け実施した．

参考文献

[1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," vol. abs/1609.03499, 2017. [Online]. Available: https://arxiv.org/abs/1703.10135

[3] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[4] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 755–767, 2018.

[5] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 1268–1272.

[6] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while speaking: Speech chain by deep learning," in *Proc. ASRU*, Okinawa, Japan, Dec. 2017.

[7] S. Takamichi, K. Tomoki, and H. Saruwatari, "Sampling-based speech parameter generation using moment-matching network," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 3961–3965.

[8] S. Sitaram, A. Parlikar, G. K. Anumanchipalli, and A. W. Black, "Universal grapheme-based speech synthesis," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 3360–3364.

[9] N. Hirayama, K. Yoshino, K. Itoyama, S. Mori, and H. G. Okuno, "Automatic speech recognition for mixed dialect utterances by mixing dialect language models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 32, no. 2, pp. 373–382, 2015.

[10] K. Yoshino, N. Hirayama, S. Mori, F. Takahashi, K. Itoyama, and H. G. Okuno, "Parallel speech corpora of japanese dialects," in *Proc. LREC*, Portoroz, Slovenia, May 2016, pp. pp. 4652–4657.

[11] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2278–2282.

[12] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*, Brisbane, Australia, Apr. 2015, pp. 4460–4464.

[13] S. Takamichi and H. Saruwatari, "CPJD corpus: Crowdsourced parallel speech corpus of Japanese dialects," in *Proc. LREC*, Miyazaki, Japan, May 2018.

[14] Y. Ijima, H. Nobukatsu, R. Matsumura, and T. Asami, "Prosody aware word-level encoder based on BLSTM-RNNs for DNN-based speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 764–768. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-521

[15] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[16] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[17] R. Senrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. ACL*, Berlin, Germany, 2016, pp. 1715–1725. [Online]. Available: http://www.aclweb.org/anthology/P16-1162

[18] "sentencepiece." [Online]. Available: https://github.com/google/sentencepiece

[19] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Postfilters to modify the modulation spectrum for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.

[20] 木部暢子, "方言アクセントの誕生," **国語研プロジェクトレビュー**, no. 2, pp. 23–35, Jul. 2010.

[21] "Open JTalk http://open-jtalk.sp.nitech.ac.jp/."

[22] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *Proc. NAACL*, Atlanta, U.S.A., Jun. 2013, pp. 644–648.

[23] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, Lauderdale, U.S.A., Apr. 2011, pp. 315–323.

[24] D. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.

[25] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

[26] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," vol. abs/1711.00354, 2017. [Online]. Available: https://arxiv.org/abs/1711.00354