Vol.2020-MUS-127 No.33
Vol.2020-SLP-132 No.33
2020/6/7

VOCALOID曲の歌唱におけるブレス位置の自動推定

内藤 悟嗣[2,a)]　齋藤 佑樹[1,b)]　高道 慎之介[1,c)]　齋藤 康之[2,d)]　猿渡 洋[1,e)]

概要：本研究では機械学習を用いて，VOCALOID 楽曲のブレス位置の自動推定法を提案する．VOCALOID 曲は，作曲者が意図的にブレス音を挿入しない限り，楽曲中にブレス音が存在しない．そのため，ユーザが VOCALOID 曲を歌唱する際には，適切な位置にブレスを挿入する必要がある．本稿では，VOCALOID 楽曲の言語的・歌声的特徴を統合した統計モデルに基づくブレス位置予測を行い，実験的評価でその効果を検証する．

Satoshi Naito[2,a)]　Yuki Saito[1,b)]　Shinnosuke Takamichi[1,c)]　Yasuyuki Saito[2,d)]
Hiroshi Saruwatari[1,e)]

1. はじめに

VOCALOID [1] とはヤマハが開発した歌声合成技術であり，10 年以上もの歴史を持っている．現在，VOCALOID 曲は数多くのユーザの創作活動に使用されている一方で，カラオケ楽曲として歌唱活動にも使用されている．しかしながら，歌声合成ではその性格上，作曲者が意図的にブレス音を挿入しない限り，楽曲中にブレス音が存在しない．そのため，ユーザが VOCALOID 曲を歌唱する際には，適切な位置にブレスを挿入する必要がある．この課題に対し，VOCALOID 楽曲の歌詞に自動的にブレス位置を付与できれば，ユーザ歌唱の支援になると考えられる．

そこで本研究では，機械学習を用いて VOCALOID 曲の適切なブレス位置を推定する方法を提案する．機械学習の特徴量として，意味論的分断を考慮した言語的特徴量と，非歌唱区間を考慮した歌声的特徴量を用いる．実験的評価では，言語的特徴のみ，歌声的特徴のみ，また，言語的・歌声的特徴を用いた推定結果を評価する．

2. 機械学習を用いたブレス位置推定

ブレス位置はフレーズ（楽曲的なまとまり）間に位置する可能性が高い [2]．また，フレーズはボーカルの言語的・歌声的特徴に関連すると予想され，それを言及する関連論文も存在する [3]．本研究ではまず，VOCALOID 曲の歌詞と歌声を用意し，ブレス位置を付与する．その後，ブレス位置において言語的・歌声的変化が生じると仮定し，所望の位置の言語的・音声的特徴量からブレス挿入の有無を予測する統計モデルを学習する．

2.1 データセットの作成

まず，歌詞に対してブレス位置を付与する．言語的に挿入可能なブレス位置は文節境界のみに存在すると仮定し，文節分割された歌詞の分割境界に対してブレス挿入の有無の 2 値ラベルをアノテーションする．本アノテーションは全て手動で行う．作業者の違いによりブレス挿入基準が変わることをさけるために，単一の作業者で本アノテーションを実施する．次に，歌声データを作成するために，VOCALOID 曲とそのオフボーカル版（ボーカルのみを含まないトラック）を用意する．VOCALOID 曲の周波数振幅からオフボーカル版の周波数振幅を減算し，VOCALOID 曲の周波数位相を付与したものを歌声データとする．ここでは，歌詞と別途のブレス位置を付与する．具体的には，非歌唱時間が 0.4 [sec] 以上の箇所にブレス挿入のラベルを付与する．

2.2 言語的特徴によるブレス位置推定

ブレスは，歌詞の文節間での係り受けが起こりにくい場所に位置する．すなわち，文節に分割された歌詞が意味論

[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] 木更津工業高等専門学校 情報工学科, 2-11-1 Kiyomidai Higashi, Kisarazu-shi, Chiba 292-0041, Japan.
[a)] j16427@kisarazu.kosen-ac.jp
[b)] yuuki_saito@ipc.i.u-tokyo.ac.jp
[c)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp
[d)] saito@j.kisarazu.ac.jp
[e)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

的に分断される位置に挿入されると予想される．そこで，係り受け関係の強さを意味するスコア [4] を用いた推定を行う．まず，歌詞を文節単位で分割する．そして，各々の当該文節間とその前後文節間の 3 つのスコアを入力として，当該文節間のブレス挿入の有無を推定する識別器を学習する．識別器には SVM（support vector machine）を用いてブレス挿入が可能であれば (1)，そうでないのなら (0) と定義し学習をする．推定時には，スコアを学習済みの SVM へ与え，ブレス挿入の事後確率を計算し，一定値以上の事後確率を持つ文節間にブレスを挿入する．

2.3 歌声的特徴によるブレス位置推定

歌声的特徴によるブレスは，長い非歌唱区間に挿入される可能性がある．そのため，この挿入は，2.2 節のような意味論的分断とは別の単位で発生すると考えられる．そこで，歌声から得られるスペクトログラムを用いて，モーラ単位の推定を行う．まず，ラベリングされたブレスが挿入可能な時間と不可能な時間を中心に前後 1 [sec] 間のスペクトログラムをそれぞれ算出する．そして，得られたスペクトログラムを入力として，2.2 節と同様にブレス挿入が可能であれば (1)，そうでないのなら (0) と定義し，2 値分類を CNN (convolution neural network) を用いて学習する．損失関数は交差エントロピー誤差を用いる．t をブレス挿入ラベル，y を推定された事後確率とするとき，損失 E は式 (1) で求められる．

$$E = -\sum_{k=1}^{K} t_k \log y_k \qquad (1)$$

ここで，添字の k はデータインデックス，K はデータ総数である．

推定にはまず，歌詞と歌声の時間的対応付け（アライメント）は，音響モデルに基づく音素アライメントを行うことで獲得する．次に歌詞のそれぞれのモーラの発声開始時間をアライメントによって取得し，それらの前後 1 [sec] 間のスペクトログラムを入力として学習済みの CNN へ与え，各モーラにおけるブレス挿入の事後確率を計算し，一定値以上の事後確率を持つモーラ間にブレスを挿入する．

2.4 言語・歌声情報を用いたブレス位置推定

2.2 節と 2.3 節の推定から各々得られる事後確率を入力として，新たにブレス挿入が可能であれば (1)，そうでないなら (0) と定義し，2 値分類を SVM を用いて学習する．推定には学習時と同様，2.2 節と 2.3 節の推定から得られる各々の確率を学習済みの SVM へ与え，ブレス挿入が可能であれば (1)，そうでないなら (0) と推定する．この推定は，2.3 節と同様にモーラ毎に行われる．ただし，言語情報に基づく方法は文節毎，歌声情報に基づく方法はモーラ毎に行われるため，文節間に対応しないモーラ間の推定においては，言語情報に基づく事後確率を 0 とする．

表 1 2.2 節の識別器のパラメータ

分類方法	c	γ
非線形 svm	7.563	9.784

表 2 2.3 節の識別器のパラメータ

バッチサイズ	エポック数	オプティマイザ	学習率
10	10	Adam [7]	0.001

表 3 2.4 節の識別器のパラメータ

分類方法	c
線形 svm	1.512

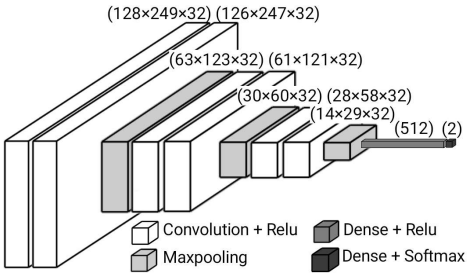

図 1 CNN の構造

3. 実験的評価

3.1 実験条件

データセットとして，歌詞・VOCALOID 曲・オフボーカル版を piapro [5] から入手可能な 30 曲を用意した．2.2 節の係り受け解析には cabocha [4] を，2.3 節の音素アライメントには Julius segmentation-kit [6] を用いた．

2.2 節で用いた識別器のパラメータを表 1 に示す．また，2.3 節で用いた識別器のパラメータを表 2 に示す．加えて，このときの CNN の構造を図 1 に示す．さらに，2.4 節で用いた識別器のパラメータを表 3 に示す．c は誤分類をどの程度許容するかを示しており，値が小さいほど許容しやすくなる．γ は非線形カーネルパラメータを示しており，値が大きくなるほどクラスを分類する決定境界が複雑になる．また，これらのハイパーパラメータは実験的に決定した．

3.2 実験結果

leave one out 法により曲の推定を行うとき，言語的特徴によるブレス位置推定の混同行列を表 4 に，歌声的特徴によるブレス位置推定の混同行列を表 5 に，言語的・歌声的特徴によるブレス位置推定の混同行列を表 6 に示す．ここでは，事後確率が 50.0 % 以上のときを (1)，50.0 % 未満を (0) とする．表 4 と表 5 を比較すると，言語的特徴の方が歌声的特徴よりも高い精度でブレス挿入位置を推定できることがわかる．また，表 6 より，言語的特徴・歌声的

表 4 言語的特徴によるブレス位置推定の混同行列

正解結果 \ 推測結果	ブレスあり	ブレスなし	Recall
ブレスあり	546	393	0.581
ブレスなし	1045	1992	
Precision	0.343		-

表 5 歌声的特徴によるブレス位置推定の混同行列

正解結果 \ 推測結果	ブレスあり	ブレスなし	Recall
ブレスあり	70	133	0.345
ブレスなし	227	2559	
Precision	0.236		-

表 6 言語・歌声的特徴によるブレス位置推定の混同行列

正解結果 \ 推測結果	ブレスあり	ブレスなし	Recall
ブレスあり	137	21	0.867
ブレスなし	484	2347	
Precision	0.2206		-

特徴の両方を用いたブレス位置推定の Accuracy（正解率）は 83.1 % ，Recall（再現率）は 86.7 %，Precision（適合率）は 22.1 % という結果が得られた．再現率に比べ適合率が小さいことから，本研究の手法は，適切な位置にブレスを挿入できる一方で，不要な箇所にも挿入することが分かる．

4. 検討事項

2.1 節において，非歌唱区間が 0.4 [sec] 以上ある場合にはブレスをする確率が高いと定義したが，本来人間が歌唱する際に発声間の時間が小さいなら短いブレス，発声間の時間が大きいなら長いブレスを入れているだろう．本研究ではブレスがあるか否かで 2 分類を行っていたためブレスの深さは考慮されていなかった．したがって，短いブレス，長いブレス，ブレスなしの 3 値分類を行うことで自然なブレスを推定できると考えられる．また，音素アライメントの segmentation-kit を用いるためには，日本語・英語・記号混じりの歌詞から読みを適切に推定する必要がある．本研究では cabocha による読み推定を実施したが，そこに多くの推定エラーが含まれ，アライメントの精度を落としてしまっていた．そのため，入力する歌詞はあらかじめ変換前と変換後を入力とする機構を用意することで精度の改善につながると考えられる．さらに，segmentation-kit に含まれる音響モデルは話声から学習されているため，歌声の分析においてはアライメント精度が落ちてしまっていた．したがって，歌声に特化した音素アライメントの機構と置き換えることでも精度の改善につながると考えられる．

5. おわりに

本研究は，ユーザによる VOCALOID 曲を歌唱する際のブレス位置推定を目的に，機械学習により適切なブレス位置を推定する手法の提案と検証をした．結果として正解率 83.1 % ，再現率 86.7 % ，適合率 22.1 %の精度で推定できた．

謝辞： 本研究は，東京大学 GAP ファンドプログラム「音声合成技術の研究開発・商用利用を加速させる音声コーパスの設計・構築」，JSPS 科研費 17H00749 の支援を受けた．

参考文献

[1] H. Kenmochi and H. Ohshita, “VOCALOID - commercial singing synthesizer based on sample concatenation,” in *Proc. INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 4011–4012.
[2] T. Nakano, M. Goto, J. Ogata, and Y. Hiraga, “Acoustic characteristics of breath sounds in solo vocal and their application to automatic breath detection,” in *IPSJ SIG Notes*, 2008, pp. 83–88.
[3] 中村敏枝, “音楽における「間」と呼吸について,” in **日本音響学会音楽音響研究会資料**, 1994, pp. 1–8.
[4] T. Kudo and Y. Matsumoto, “Japanese dependency analysis using cascaded chunking,” in *Proc. CoNLL*, 2002, pp. 63–69.
[5] “piapro,” https://piapro.jp/.
[6] “Speech segmentation toolkit using Julius,” https://github.com/julius-speech/segmentation-kit.
[7] D. Kingma and B. Jimmy, “Adam: A method for stochastic optimization,” in *arXiv preprint arXiv:1412.6980*, 2014.