Vol.2015-MUS-107 No.58
2015/5/24

DNN-HMMを用いた歌声の自動歌詞認識の検討

川井 大陸[1,a)]　山本 一公[1,b)]　中川 聖一[1,c)]

概要：朗読音声, 話し声音声, 歌声音声の音韻分布間距離を調べることで, 歌声音声の音声認識の困難さを明らかにした. また伴奏なし歌声音声を音声認識するために, GMM-HMM と DNN-HMM をそれぞれ用いて性能を比較した. これまでに提案した歌詞言語モデルや発音拡張辞書, 発話スタイル適応, 話者適応と併用することで単語認識精度 35.4% (音節認識精度 52.2%) を示して, 従来の認識精度を大きく上回る結果を得た.

1. はじめに

歌声の自動歌詞認識の第一段階として, 本稿では伴奏なし日本語歌唱に対する大語彙歌詞認識を行う. 歌声の自動歌詞認識は音声認識タスクの中でも難しい試みの一つである. その理由の一つとして歌声コーパスの不足が挙げられる. 伴奏なしの歌声データベース (DB) は研究目的で一般公開されているものがほとんどない. これまでの研究では独自に用意した DB を使って実験している [1], [3], [8], [10], [11], [12], [15].

歌詞の認識は難しいため, 多くの研究ではテストセットの歌詞を用いたクローズドな言語モデルを作成することで認識性能を向上させている [1], [3], [12], [15]. 歌声に頻出する音韻の引き伸ばしも歌詞認識を難しくする要因である. 文献 [12] では歌声に頻出する長音区間を除去することで認識率の向上を図った.

オープンデータを対象とした大語彙歌詞認識はあまり研究されていない. 文献 [11] では大規模な歌詞コーパスで学習した言語モデルと少量の歌声で適応学習した朗読音声の GMM-HMM 音響モデルを用いた. 結果は男性ボーカルに対して音素認識精度 34.9%, 男女ボーカルに対して単語認識精度 12.4%を示した. また文献 [8] では歌詞特有の特徴であるフレーズの繰り返し情報を活用している. フレーズの繰り返し部分に対して出力結果を統合することでより信頼できる書き起こしを生成できる. 結果は男女ボーカルに対して音素認識精度 27.0%, 単語認識精度 9.5%を示した. いずれの研究も低い認識精度を示しており歌詞認識の難しさが伺える.

近年, 音声認識タスクで Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) の代わりに Deep Neural Network-Hidden Markov Model (DNN-HMM) を用いることで大きな性能改善が得られている [17]. この理由は, DNN-HMM が次元数の大きなマルチフレーム特徴量を柔軟に学習できるためである.

本稿では, まず朗読音声, 話し声音声, 歌声音声の音韻分布間距離を調べることで, 歌声音声の自動歌詞認識の困難さを明らかにする. 一般に歌声は朗読音声と比べてピッチバリエーションが多かったり, 歌唱フォルマントと呼ばれる 2.8kHz 付近に強く表れる周波数が存在することが分かっている [13]. 本稿では, 歌声の特徴パラメータを統計的に分析し, 朗読音声や話し声音声と比較する. 続いて, 伴奏なし日本語歌唱の自動歌詞認識を行うために, 歌詞に適応した言語モデル, 音響モデル, 発音辞書を検討する. 最後に, GMM-HMM と DNN-HMM の比較を行い, DNN-HMM が歌声の自動歌詞認識タスクでも有効であることを示す.

2. データーベース (DB)

収集した歌詞 DB と歌声 DB を表 1 に示す.

歌詞 DB 作成のために, 歌詞投稿サイト”ピアプロ”[9] に投稿された歌詞 13 万曲分を収集した. これを MeCab[6] で形態素解析することによりテキストを単語毎に分割し読みを付与した. この歌詞 DB には同一音楽の歌詞が複数含まれる場合がある. そのため各歌詞テキストの先頭 20 単語をチェックして, 既に存在する歌詞だった場合はその歌詞を除去した. また同様の方法でテストセットの歌詞も DB から除去した. 歌詞テキストは改行を手掛かりにして区切った. 日本語のみの歌詞 DB を作成するために, アルファベットが含まれる区間を除去している.

テストセットに使う歌声のために, カラオケ音源を含む JPOP 男性ボーカル 7 名 7 曲を収集した. これらのオリジナル音源とカラオケ音源の振幅の差分から歌声だけを抽出した. さらに音節ごとに手動で時間情報付きのアノテーションを行った. また, 大語彙歌詞認識に用いるための単語単位の書き起こしも用意した. 歌声データは 0.5 秒以上の

[1] 豊橋技術科学大学
Toyohashi University of Technology
[a)] kawai@slp.cs.tut.ac.jp
[b)] kyama@tut.jp
[c)] nakagawa@tut.jp

表 1 構築した音声-言語 DB

(a) 言語 DB

名前	単語数	詳細
新聞 DB	206.7M	毎日新聞記事 91 年 1 月-94 年 9 月 (45 ヶ月分)
歌詞 DB	28.6M	ピアプロ歌詞 130K

(b) 音声 DB

名前	曲数	人数	時間	詳細
テストセット	7	7	19:01	市販 JPOP
話者適応音声	7	7	19:53	市販 JPOP
発話スタイル適応音声	40	40	1:39:28	ピアプロ曲
NN 朗読音声	6	4	4:33	大学で収録
NN 歌声音声	6	4	14:27	大学で収録

無音区間を手掛かりにして区切られている. 複数の歌詞がオーバーラップする区間, 英語歌詞を含む区間, スキャット (ex. ラーラー) 区間, 0.5 秒以上の無音区間の除去を行った.

話者適応データに使う歌声のために, テストセットと同じボーカルの JPOP7 名 7 曲を収集した. 歌声への MAP 適応に使うデータのために, "ピアプロ" に投稿された男性ボーカル 40 名 40 曲を収集し, テストセットと同様の前処理を行った. 朗読-歌声特徴変換に用いるニューラルネットワーク (NN) の学習には, 本大学の歌唱経験者 4 名より計 6 曲分の歌詞の朗読音声と歌声音声のペアを収録した. また音素ごとに手動で時間情報付きのアノテーションを行った.

3. 発話スタイル別音響モデルの差異分析

歌声認識の困難さを調査するために, 3 種類の発話スタイル (朗読, 話し声, 歌声) の音韻間の差異を調べる. 音節の全状態全混合分布の共分散行列に対する行列式の平均を音節モデルの分散の大きさ DET とする.

$$DET(a) = \frac{1}{M}\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N}|\Sigma\{P(S_a^i, j)\}| \qquad (1)$$

S_a^i : 音節 a の第 i 状態
$P(S_a^i, j)$: 音節 a の状態 i の第 j 分布
$\Sigma\{P(S_a^i, j)\}$: 音節 a 状態 i 第 j 分布の共分散行列
$|\Sigma|$: Σ の行列式
M : 状態数
N : 混合数

音節間距離は山本らの方法 [14] を用いて, 2 つの状態混合分布間の最小距離を用いて, 次式で 2 つのモデル間の距離 D を計算する.

$$D(a,b) = \frac{1}{M}\sum_{i=1}^{M}\min_{j,k} d\{P_a(S_a^i, j), P_b(S_b^i, k)\} \qquad (2)$$

$d(P_a, P_b)$: 分布間の距離

表 2 音節間距離計算に使用する音声 DB

READ	read-speech (ASJ+JNAS [5]; 133 話者 22K 発話)
SPON	spontaneous-speech (CSJ [7]; 797 講演 222K 発話)
SUNG	sung-speech (テストセット+発話スタイル/話者適応データセット; 47 話者 54 曲)

分布間の距離 d にはユークリッド距離とバタチャリア距離の 2 つを使用した. 表 2 に示す音声 DB を使って 5 状態 8 混合 116 音節のコンテキスト独立 GMM-HMM を学習した.

発話スタイル別母音別の分散の大きさを表 3 に示す (M=5, N=8). 分散の大きさを比較すると READ > SUNG > SPON となっている.

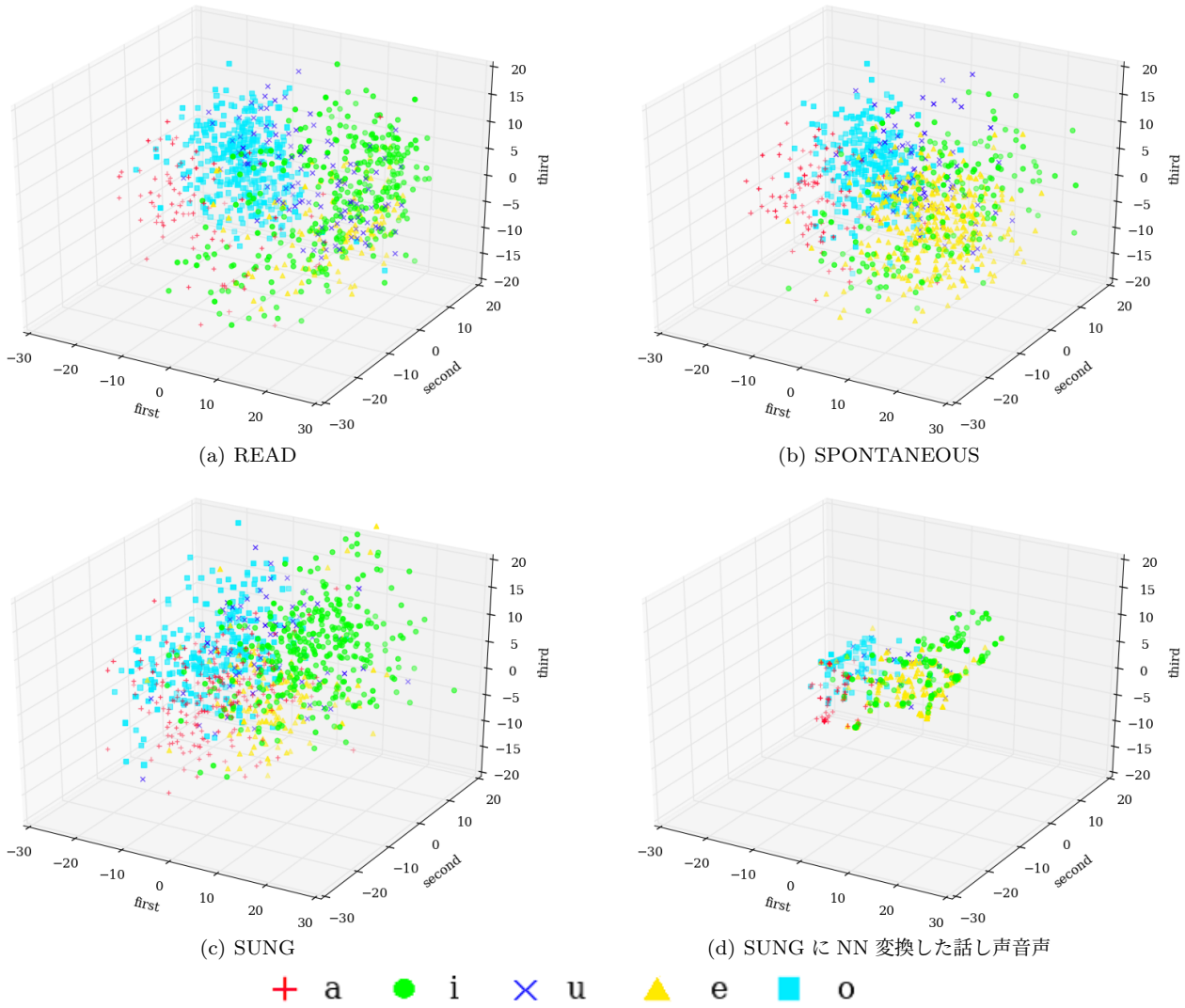
母音別の発話スタイル間距離を表 4 に示す. ここで, A_B は発話スタイル A と発話スタイル B 間の距離を意味する. READ_SPON, READ_SUNG, SPON_SUNG を比較すると, READ_SPON の距離が最も近い. これは, SUNG の分布が他の発話スタイルと比べて大きく異なることを意味している. READ_SUNG と SPON_SUNG を比較すると, SPON_SUNG の方が距離が近いことが分かる. このことから, 音響モデルの学習に使える歌声音声が少ない場合, 朗読音声より話し声音声を代わりに使った方が良いことが分かる.

発話スタイル別の母音間距離を表 5, 6 に示す. 各発話スタイルの平均母音間距離を大きい順に並び替えると, ユークリッド距離尺度では READ >> SPON $\simeq$ SUNG となった. 一方でバタチャリア距離尺度では READ >> SUNG > SPON となった.

以上の結果から, 話し声と比較して各母音の分散が大きいこと, 朗読音声と比較して歌声音声は母音間の距離が小さくなっていることが分かり, このことが歌声音声の自動歌詞認識を難しくしていると考えられる.

朗読音声と歌声音声の発話スタイル間距離 (バタチャリア距離) は歌声音声の母音間距離と同じくらい離れていることが分かる. 一方, 話し声音声は朗読音声より歌声音声との発話スタイル間距離が小さいことが分かる. このことから擬似的に歌声音声を作る場合, 朗読-歌声変換をするより, 話し声-歌声変換の方が上手く変換しやすいと考えられる.

発話スタイル別の 5 母音の音響パラメータの分布図を図 1 に示す. 音響特徴パラメータは MFCC12 次元で, 主成分分析で 3 次元に圧縮した図である. 累積寄与率は 50.2%であり, 3 次元では不十分であるが参考にはなる. 参考までに, 次節で述べるニューラルネットワークによる話し声音声を歌声音声に変換した音声も示した. 図からも明らかなように, 音声の認識は朗読音声, 話し声音声, 歌声音声, NN 変換した話し声音声の順に難しくなることが予想される.



(a) READ (b) SPONTANEOUS

(c) SUNG (d) SUNG に NN 変換した話し声音声

図 1 5 母音の MFCC12 次元を主成分分析で 3 次元に圧縮した結果

表 3 各母音の共分散行列の行列式の平均

母音	READ	SPON	SUNG
a	7.74E+19	7.28E+17	2.93E+18
i	1.23E+18	4.00E+17	7.42E+18
u	2.58E+18	5.32E+17	8.89E+18
e	1.51E+17	2.09E+17	7.69E+17
o	5.37E+18	5.65E+17	7.36E+18
Ave.	1.73E+19	4.87E+17	5.47E+18

表 4 母音別発話スタイル間ユークリッド距離 (括弧内はバタチャリア距離)

母音	READ_SPON	READ_SUNG	SPON_SUNG
a	6.36 (0.24)	10.7 (0.85)	9.54 (0.50)
i	5.63 (0.18)	8.61 (0.39)	7.94 (0.34)
u	6.70 (0.25)	9.72 (0.62)	6.39 (0.33)
e	6.51 (0.30)	9.42 (0.60)	7.98 (0.43)
o	5.39 (0.21)	10.23 (0.71)	7.43 (0.37)
Ave.	6.12 (0.24)	9.74 (0.64)	7.85 (0.40)

4. 歌詞認識システム [16]

4.1 言語モデル

言語モデルは，Palmkit[4] を用いて，語彙サイズ 2 万語の単語 2 グラム，3 グラム，4 グラム言語モデルを作成した．スムージング手法は，いずれも Witten-Bell 法を用いた．

4.2 長母音に適した発音辞書

歌声特有の「音韻の引き延ばし」によって生じる挿入誤りは母音が連続する形で現れる．そこで発音辞書を拡張することで連続する母音を捉えられないか検討した．各単語の読みに対して全ての音節の母音が 2 個まで連続できるように発音辞書に発音変形を追加した．

実際に”ちょうちょ”という単語を発音拡張した例を表 7 に示す．この場合，発音拡張した後の発音の総数は 8 ($= 2^3$) となる．

表 5 発話スタイル別母音間ユークリッド距離

(a) READ

	i	u	e	o	Ave.
a	19.72	18.34	15.75	16.37	17.55
i		11.25	5.22	21.64	14.46
u			11.29	6.22	11.78
e				21.20	13.36
o					16.36
Ave.					14.70

(b) SPONTANEOUS

	i	u	e	o	Ave.
a	13.02	14.24	14.10	11.68	13.26
i		6.94	4.00	15.59	9.89
u			8.30	8.26	9.44
e				15.05	10.36
o					12.65
Ave.					11.12

(c) SUNG

	i	u	e	o	Ave.
a	14.37	11.50	8.03	8.95	10.71
i		11.40	8.22	13.89	11.97
u			11.94	10.22	11.27
e				12.34	10.13
o					11.35
Ave.					11.09

表 6 発話スタイル別母音間バタチャリア距離

(a) READ

	i	u	e	o	Ave.
a	2.15	1.55	1.64	1.33	1.67
i		0.75	0.19	2.60	1.42
u			0.69	0.19	0.80
e				2.69	1.30
o					1.70
Ave.					1.38

(b) SPONTANEOUS

	i	u	e	o	Ave.
a	0.84	0.89	1.16	0.67	0.89
i		0.30	0.11	1.10	0.59
u			0.37	0.35	0.48
e				1.13	0.69
o					0.81
Ave.					0.69

(c) SUNG

	i	u	e	o	Ave.
a	1.22	0.94	0.47	0.59	0.80
i		0.71	0.46	1.26	0.91
u			0.86	0.71	0.80
e				1.20	0.75
o					0.94
Ave.					0.84

4.3 GMM-HMM 音響モデル

音響モデルは 116 音節のコンテキスト独立 5 状態 64 混合 GMM-HMM を使用する. 左コンテキスト依存モデルは歌声の MAP 適応による性能改善がコンテキスト独立モデルほど得られなかったため使用していない. 特徴量は MFCC, MFCC の $\Delta, \Delta\Delta$, 対数パワーの $\Delta, \Delta\Delta$ の計 38 次元を用いる. 学習データには, CSJ[7] (話し声音声) を用いた. 音響モデルの発話スタイル適応には, MAP 適応とニューラルネットワークによる朗読-歌声変換の 2 種類を試みた.

4.3.1 MAP 適応

MAP 適応は追加学習の一方法である. 学習済みの音響モデルのパラメータを事前情報として, 適応データを使って事後確率が最大になるようにパラメータを更新する. 我々は大量の話し声音声を使って話し声音声モデルを学習し, 少量の歌声音声を使って適応を行った. ラベル付きデータを利用して各ガウス分布の平均ベクトルと対角分散行列を学習した.

4.3.2 朗読-歌声変換

同一音素系列の同一話者の朗読-歌声 MFCC のペアを使ってニューラルネットワークを学習することで, 朗読音声を擬似的な歌声音声に変換できないか検討した. 朗読-歌声変換をするネットワークは入力層, 中間層 2 層, 出力層から構成される. 朗読 → 歌声変換をするネットワークでは, 入力層 (朗読音声), 出力層 (歌声音声) はユニット数 12 (MFCC12 次元に対応) の線形関数を持つ. ネットワークの中間層は 24 ユニットのシグモイド関数を持ち, 各層はバイアスユニットを持つ. 学習データには, 同一人物の同一コンテキスト上に出現する朗読-歌声音素のペア 4 名 6 曲を用いた. 母音に対しては可能な限り全フレームを用いて, 子音に対しては先頭 3 フレームを用いた.

作成したニューラルネットワークを用いて, 朗読音声特徴量の MFCC に対して朗読 → 歌声変換をした後, その歌声変換済み特徴量を使って音響モデルを学習する. 実際には, 学習された NN で話し声音声を歌声変換した方が良かったのでこれを用いた. これは 3 節で述べた話し声, 歌声変換の方が上手く変換しやすい予想と合致している.

表 7 発音拡張の具体例

ちょうちょ	chō	cho
発音辞書 (original)	cho u	cho
発音拡張辞書 (extension)	cho u u cho u ⋮ cho o u u	cho cho o cho o

4.4 DNN-HMM 音響モデル

(a) モデル構造と学習方法

DNN-HMM は入力層 429 ユニット, 隠れ層 3 層でそれぞれ 1024 ユニット (Rectified Linear Unit), 出力層 645 ユニットの計 5 層で構成される. 入力特徴量は MFCC, MFCC の Δ, ΔΔ, 対数パワー, 対数パワーの Δ, ΔΔ の計 39 次元を 11 フレーム用いる. 出力層のユニット数はコンテキスト独立 HMM の状態数 (5 状態 ×123 音節+3 状態 ×10 音節) と一致する. 学習データの状態ラベルは GMM-HMM を用いて強制アライメントにより作成した. DNN は事前学習はせずファインチューニングのみで学習した.

(b) 発話スタイル/話者適応

発話スタイル適応 DNN を作成するために, 我々は話し声音声に歌声音声を加えて DNN を学習した. 発話スタイル適応には歌声音声 40 名 40 曲を使用した.

話者適応にはテストセットそれぞれに対して同一話者の歌声 1 曲を使用した. 話者適応のデータは少量なので発話スタイル適応と同時に実行した.

5. 実験

5.1 実験条件

実験に用いる音声の分析条件はサンプリング周波数 16kHz, フレーム窓長 25ms, フレームシフト長 10ms である. 抽出した特徴量は MFCC, MFCC の Δ, ΔΔ, 対数パワーの Δ, ΔΔ の計 38 次元である (但し, DNN-HMM の場合はパワーも含めた 39 次元). 実験の評価には男性ボーカル 7 名 7 曲のテストセットを用いた.

ベース音響モデルは, CSJ[7] で学習した 116 音節のコンテキスト独立 64 混合 GMM-HMM で構成される. 学習に用いた CSJ のデータは男性の講演音声から 797 講演 222K 発話である. DNN の学習にも同様のデータを用いた.

歌声への MAP 適応は発話スタイル適応 ($\mathrm{MAP_{SONG}}$) と話者適応 ($\mathrm{MAP_{SPK}}$) の 2 つを試みた. 発話スタイル適応には男性ボーカル 40 名 40 曲を用いた. 話者適応にはテストセットと同一話者の男性ボーカル 7 名 7 曲を用いた. 朗読-歌声 MFCC 変換のニューラルネットワークは, 4 名 6 ペアの朗読-歌声音声を使って, 4.3.2 節で説明した方法をベースモデルに適用した.

言語モデルは単語 2 グラム, 3 グラム, 4 グラム言語モデルを作成した. 学習データは, 1991 年から 1994 年までの毎日新聞記事 (45 ヶ月, 206.7M 単語) で構成される新聞 DB と 130K 歌詞で構成される歌詞 DB である. 大語彙歌詞認識で使う発音辞書は, 言語モデルの学習データから頻出上位 2 万語を使用した.

デコーダは SPOJUS++[2] を用いた. 言語重みは 1, 10, 15, 20, 25, 30 から, ワードペナルティは-30, -20, -10, 0 からシステムごとにテストセットの平均認識精度が最大になるように値を選択している.

5.2 実験結果

テストセットの書き起こしに対する言語モデルの単語未知語率 (OOV) とパープレキシティ (PP) を表 8 に示す. 新聞モデルは単語未知語率 13.1%, パープレキシティ 194, 歌詞モデルは単語未知語率 1.8%, パープレキシティ 107 となっており歌詞言語モデルの方が優れた性能を示した. 大語彙歌詞認識の結果を Table 9 に示す. 言語モデルには新聞コーパスと歌詞コーパスで学習した単語 3 グラム言語モデルを用いた.

表 9 の AM の列で, "GMM"は GMM-HMM, "DNN"は DNN-HMM を意味する. 下付き文字は音響モデルの学習データを表しており"SPON"は CSJ コーパス, "NN"はニューラルネットワークで歌声変換した CSJ コーパス, SONG は発話スタイル適応データ (歌声データ), SPK は話者適応データを意味する. LEXICON の Original は 1 通りの発音表記を表し, Extension は長音に対する複数の発音表記を許した場合である. ベースラインは News+Original+$\mathrm{GMM_{SPON}}$ で単語認識精度 4.4%を示した. 新聞言語モデルの代わりに歌詞言語モデルを使用することで単語認識精度が 4.7%向上している.

発話スタイル適応データのみで GMM-HMM の学習を行ったところ, CSJ コーパスと比べて学習データ数は大きく減少したが性能は改善した ($\mathrm{GMM_{SONG}}$). DNN-HMM でも同様の傾向が得られた ($\mathrm{DNN_{SONG}}$). 先行研究 [11] で使われた発話スタイル適応手法によって単語認識精度 14.6%に向上した ($\mathrm{GMM_{SPON}} + \mathrm{MLLR_{SONG}}$). 発話スタイル適応手法を MAP 適応にすることでより高い認識精度を示すことが分かった ($\mathrm{GMM_{SPON}} + \mathrm{MAP_{SONG}}$).

発音辞書を拡張することで, 単語認識精度が 2.3%向上している. 言語重みと挿入ペナルティを揃えて比較してみたところ, 挿入誤りの減少が顕著に見られたことから, 音韻の引き延ばしによって湧き出した母音を発音辞書を拡張することで捉えられるようになったと考えられる.

NN による朗読 → 歌声変換した音響モデルでは, 変換後の特徴量を使ってモデルを学習すると分散や母音間ケプストラム距離が小さくなることがわかり, これが原因で認識性能が低下した (図 1 参照). 一方で NN 変換した音響モデルに対して発話スタイル適応すると, NN 変換していない場合より認識精度が 0.9%向上した. さらに話者適応をすることで単語認識精度 29.1%を示した.

話し声音声で学習した DNN-HMM は単語認識精度 22.1%を示した. この結果は GMM-HMM より良かった. DNN-HMM を CSJ コーパスに発話スタイル適応データを加えて学習すると単語認識精度 34.2%を示した. さらに, 話者適応データを加えることで単語認識精度 35.4%を示した. この結果は, 我々が知っている限りでは発表されている論文の中で最高性能を示している.

歌詞音節 3 グラム言語モデルを用いて音節認識した結果を表 10 に示す. $\mathrm{DNN_{SPON+SONG+SPK}}$ は節認識精度 52.2%で最高性能を示した (音素認識精度 63.5%).

表 8 LM の評価結果 (単語未知語率とパープレキシティ)

学習データ	N-gram	OOV[%]	PP
News	2	13.1	218
206.7M 単語	3	13.1	194
	4	13.1	210
Lyrics	2	1.8	134
28.6M 単語	3	1.8	107
	4	1.8	114

表 9 大語彙歌詞認識の結果 [%]

LM	LEXICON	AM	Cor	Acc
News	Original	GMM_{SPON} (Base)	5.1	4.4
Lyrics		GMM_{SPON} (Base)	9.8	9.1
		GMM_{SONG} (Base)	17.3	14.4
		$+MLLR_{SONG}$	18.9	14.6
		$+MAP_{SONG}$	26.9	22.9
		DNN_{SPON}	18.7	13.6
		DNN_{SONG}	37.0	30.3
		$DNN_{SPON+SONG}$	38.6	32.2
	Extension	GMM_{SPON} (Base)	14.9	11.4
		$+MAP_{SONG}$	33.6	25.0
		GMM_{NN}	14.2	10.5
		$+MAP_{SONG}$	34.3	25.9
		$+MAP_{SONG+SPK}$	35.4	29.1
		DNN_{SPON}	27.6	22.1
		$DNN_{SPON+SONG}$	40.8	34.2
		$DNN_{SPON+SONG+SPK}$	42.5	35.4

表 10 音節認識の結果 [%]

LM	LEXICON	AM	Cor	Acc
Lyrics	Extension	GMM_{SPON}(Base)	30.9	28.8
		$+MAP_{SONG+SPK}$	50.7	47.5
		DNN_{SPON}	46.3	43.2
		$DNN_{SPON+SONG+SPK}$	57.8	52.2

6. おわりに

本稿で我々は歌声に適応した言語モデル, 音響モデル, 発音辞書を検討した. 新聞言語モデルより歌詞言語モデルの方が認識性能が向上した. 発音辞書に発音バリエーションを追加することで音韻の引き延ばしを捉えられるようになり性能が改善した. 発話スタイル適応データで学習した GMM-HMM は話し声音声で学習した GMM-HMM より高い性能を示したが, 話し声音声で学習した GMM-HMM に対して発話スタイル適応をしたモデルは更に良くなった. DNN-HMM でも同様の結果になった. NN による朗読 → 歌声変換は発話スタイル/話者適応と併用することで初めて効果が発揮された. DNN-HMM は GMM-HMM より高い性能となり, 最も高い性能を示した (単語認識精度 35.4%, 音節認識精度 52.2%, 音素認識精度 63.5%). 我々が知る限り, この認識精度はこれまでの研究の中で最も高い.

今後は, NN による歌声変換の改善を行う予定である.

参考文献

[1] Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: Singer identification based on accompaniment sound reduction and reliable frame selection, *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pp. 329–336 (2005).

[2] Fujii, Y., Yamamoto, K. and Nakagawa, S.: Large vocabulary speech recognition system: SPOJUS++, *Proceeding of 11th WSEAS International Conference MUSP*, Vol. 11 (2011).

[3] Hosoya, T., Suzuki, M., Ito, A., Makino, S., Smith, L. A., Bainbridge, D. and Witten, I. H.: Lyrics Recognition from a Singing Voice Based on Finite State Automaton for Music Information Retrieval, *Proc. ISMIR*, pp. 532–535 (2005).

[4] Ito, A.: Palmkit, http://palmkit.sourceforge.net/.

[5] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus, *Proc. Int. Conf. on Spoken Language Processing*, pp. 722–725 (1998).

[6] KUDO, T.: MeCab : Yet Another Part-of-Speech and Morphological Analyzer, *http://mecab.sourceforge.net/* (2005).

[7] Maekawa, K., Koiso, H., Furui, S. and Isahara, H.: Spontaneous Speech Corpus of Japanese., *Proc. 2nd LREC*, European Language Resources Association, pp. 947–952 (2000).

[8] McVicar, M., Ellis, D. and Goto, M.: Leveraging repetition for improved automatic lyric transcription of popular music, *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3141–3145 (2014).

[9] Media, C. F.: Piapro, http://piapro.jp/.

[10] Mesaros, A.: Singing voice identification and lyrics transcription for music information retrieval invited paper, *Speech Technology and Human - Computer Dialogue (SpeD), 2013 7th Conference on*, pp. 1–10 (online), DOI: 10.1109/SpeD.2013.6682644 (2013).

[11] Mesaros, A. and Virtanen, T.: Automatic Recognition of Lyrics in Singing, *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2010, No. 1, p. 546047 (online), DOI: 10.1155/2010/546047 (2010).

[12] Sasou, A. and Goto, M.: Japan Patent, JP4576612B (2007).

[13] Sundberg, J., 榊原 健一 (監訳), 伊藤みか, 小西知子, 林良子 (訳)：歌声の科学, 東京電機大学出版局 (2007).

[14] 山本一公, 中川聖一：発話スタイルによる話速・音韻間距離・ゆう度の違いと音声認識性能の関係 (音声情報処理: 現状と将来技術論文特集), 電子情報通信学会論文誌. D-II, 情報・システム, II-パターン処理, Vol. 83, No. 11, pp. 2438–2447 (2000).

[15] 佐宗 晃, 後藤真孝, 速水 悟, 田中和世：ARHMM に基づいた音声分析手法と歌声認識による評価 (聴覚・音声及び一般), 電子情報通信学会技術研究報告. SP, 音声, Vol. 105, No. 199, pp. 19–24（オンライン), 入手先 ⟨http://ci.nii.ac.jp/naid/110003298740/⟩ (2005).

[16] 川井大陸, 山本一公, 中川聖一：朗読音声-歌声音声の特徴量変換と話者適応を用いた歌詞認識の性能向上の検討, 情報処理学会研究報告. SLP, 音声言語情報処理, Vol. 2014, No. 2, pp. 1–6 (2014).

[17] 関 博史, 中川聖一：音節単位 DNN-HMM の音声認識の評価, 日本音響学会春季研究発表会講演論文集, pp. 179–182 (2014.3).