Vol.2014-SLP-104 No.2
2014/12/15

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

川井 大陸[1,a)]　山本 一公[1,b)]　中川 聖一[1,c)]

概要：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の自動歌詞認識を行う．このために歌声に適応した言語モデル，音響モデル，発音辞書を使うことで伴奏なし独唱の自動歌詞認識を検討する．言語モデルには歌詞をうまく捉えるために歌詞コーパスで学習した単語 N グラム言語モデルを使用した．音響モデルの学習には，歌声データ不足を補うため少量の歌声データを用いて 2 種類の適応化をした．1 つ目は MAP 適応による音響モデルの適応学習である．MAP 適応では 40 名 40 曲の歌声データを使う方法と，1 曲の話者適応データを使う方法を試みた．2 つ目は朗読 MFCC と歌声 MFCC のペアを使って学習したニューラルネットワークによる特徴変換である．歌声で頻繁に表れる「伸ばす音」に対処するため，発音辞書のバリエーションを増やした．性能評価には，事前に伴奏音を除去した JPOP 男性 7 名 7 曲の楽曲を用いる．実験の結果，提案システムは音節認識精度 46.1%（音素認識精度 59.0%），単語認識精度 25.9%を示し，新聞言語モデルと話し声音響モデルに基づく従来のシステムより良い性能を示した．

キーワード：歌詞認識，朗読-歌声変換，MAP 適応，長母音化

Abstract: As a first step, we consider Japanese lyrics recognition in monophonic singing that contains no musical instruments. To express singing well, we attempt to use an n-gram language model using a lyrics corpus, singing-adapted GMM-HMM-based acoustic models and plural pronunciation lexicons for vowel-lengthening. We attempted to adapt the read-speech AMs to sung-speech AMs using two approaches. One is MAP adaptation and the other is neural network-based feature transformation. For adapting to singing, we use 40 pieces of music sung by 40 male singers. For adapting to speaker, we use a piece of music sung by a male singer who is the same speaker as a singer of a test data. To deal with the property of singing offten involving lengthening the duration of each vowel, we augment the pronunciation variations. Evaluation is performed on a test set that contains 7 pieces of commercial music sung by 7 male singers. As a result of experiments, our system showed syllable accuracy of 46.1% (phoneme accuracy of 59.0%) and word accuracy of 25.9% in male monophonic Japanese singing. This result showed higher accuracy than a conventional system based on the newspaper LM and the read-speech AM.

Keywords: lyrics recognition, read-sung speech transformation, MAP adaptation, vowel-lengthening

1. はじめに

近年，インターネットの普及によって，我々が利用できる音楽は急激に増加している．これに伴い，音楽データに対する情報検索の需要が高まっている．音楽を特定するための重要な情報として歌手，ジャンル，旋律，歌詞がある．

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．音響モデルの学習には，大規模なアノテーション済み音声データが必要となる．Wang ら [1] は，話し声データベースで学習した音響モデルを用いて歌声の歌詞認識を行った．この研究ではテストデータの歌詞テキストを言語モデルに用いることで高い認識精度を達成している．佐宗ら [2] は，少量の歌声で適応学習した朗読音声の GMM-HMM 音響モデルを用いることで，歌声のデータ不足に対処した．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 [3] 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．細谷ら [4] は，認識する歌詞データベース（238 曲）が事前に分かっているという制限のもとで，単語 3 グラム言語モデルや有限状態オートマトンを構築した．この制限のため，単語 3 グラム言語モデルを用いた場合は 48.2%，有限状態

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

オートマトンを用いた場合は約 77.4%の単語認識精度を得ている．歌詞の認識は難しいため，多くの研究は認識率向上のためテストセットの歌詞を使って言語モデルを作成しており，クローズドな言語モデルの使用により当然認識率は向上する [2], [3].

大語彙歌詞認識の研究として，文献 [5] 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 音響モデルを用いて歌声の歌詞認識を行った. 結果は単語認識精度 12.4%と低い値を示しており，歌詞認識の難しさが伺える．また文献 [6] 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することで性能改善をしている．この研究では，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している．本稿では，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，歌声に適応した音響モデル，発音辞書を検討する．

2. データベース

公開されている歌声コーパスや市販の音楽 CD の多くは歌声と伴奏がミックスダウンされている．しかしながら，我々は歌声の自動歌詞認識の第一段階として伴奏音を含まない歌声を必要としている．また，言語モデルを作成するために歌詞データベースが必要となる．そこで，歌声音声-歌詞言語データベースを新たに構築した．

歌詞データベース作成のために，歌詞投稿サイト”ピアプロ（http://piapro.jp/）に投稿された歌詞 13 万曲分を収集した．そして MeCab 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．この歌詞データベース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する．また同様の方法でテストセットの歌詞もデータベースから除去する．歌詞テキストは改行を手掛かりにして区切った．そして日本語のみの歌詞データベースを作成するために，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．そして “歌声りっぷ”（http://www.vector.co.jp/soft/win95/art/se127635.html.）を用いて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 秒以上の無音区間を手掛かりにして区切られている．そして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 JPOP6 名 6 曲を収集した．ただし，7 名のうち 1 名だけ話者適応データが入手できなかった．MAP 適応学習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．収集したデータベースをまとめた表を**表 1** に示す．

3. 歌詞認識システム

3.1 言語モデル

言語モデルは, Palmkit（http://palmkit.sourceforge.net/）を用いて，音節 3 グラム言語モデル，単語 2 グラム，3 グラム，4 グラム言語モデルを作成した．音節言語モデルの語彙サイズは 116 語，単語言語モデルの語彙サイズは 2 万語とした．スムージング手法は，いずれも Witten-Bell 法を用いた．

言語モデルの評価尺度には単語未知語率 (OOV) とパープレキシティ (PP) を用いる．単語未知語率は式 (1) で定義される．

$$OOV\ rate = \frac{number\ of\ OOVs}{number\ of\ total\ words}. \tag{1}$$

テストセットの単語列 $w_1...w_n$ の生成確率を $P(w_1...w_n)$ とすれば，パープレキシティは式 (2) で定義される．

$$PP = P(w_1...w_n)^{-\frac{1}{n}}. \tag{2}$$

3.2 音響モデル

3.2.1 GMM-HMM

音響モデルは GMM-HMM モデルを使用する．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．音響モデルの歌声適応には，MAP 適応学習を用いる方法とニューラルネットワークを用いる方法の 2 種類を試みた．

3.2.2 MAP 適応

MAP 適応 [7] 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．我々は大量の朗読音声を使って朗読音響モデルを学習し，少量の歌声音声を使って適応学習を行った．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．

3.2.3 朗読-歌声変換

事前調査の結果，朗読音声と歌声音声の MFCC の音素間ケプストラム距離は朗読音声同士の音素間ケプストラム距離よりも大きいことがわかった．つまり，朗読音声と歌声音声は音素間の MFCC が大きく異なる．そこで同一音素系列の朗読-歌声 MFCC のペアを使ってニューラルネッ

表 1 収集した歌声音声-歌詞言語データベース
(a) 言語データベース

用途	曲数	単語数	説明
言語モデル学習データ	130K	28.6M	“ピアプロ” より収集した歌詞テキスト

(b) 音声データベース

用途	曲数	話者数	再生時間	説明
テストセット	7	7	19 分 1 秒	市販 JPOP より収集
話者適応データ	6	6	17 分 24 秒	市販 JPOP より収集
MAP 適応データ	40	40	1 時間 39 分 28 秒	“ピアプロ” より収集
NN 学習用朗読音声	6	4	4 分 33 秒	大学で収録
NN 学習用歌声音声	6	4	14 分 27 秒	大学で収録

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，またその逆方向の変換が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．歌声 → 朗読変換をするネットワークでは，入力層（歌声音声）はユニット数 12（MFCC12 次元），もしくは 13（MFCC12 次元+ピッチ情報に対応）で，出力層（朗読音声）はユニット数 12（MFCC12 次元に対応）の線形関数を持つ．いずれのネットワークも中間層は 24 ユニットのシグモイド関数を持つ．また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．**図 1** はマッピングの概略図である．以下の 3 つの変換法を比較する．

- 朗読 → 歌声データ変換
  （read-speech to sung-speech for data; r2s$_{\mathrm{d}}$ ）
  朗読特徴量の 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．
- 朗読 → 歌声モデル変換
  （read-speech to sung-speech for model; r2s$_{\mathrm{m}}$）
  朗読特徴量で音響モデルを学習した後，モデルの MFCC 平均ベクトルに対して朗読 → 歌声変換をする．$\Delta, \Delta\Delta$ パラメータを変換することもできる．まずベースモデルの $\Delta, \Delta\Delta$ パラメータを使って前後 2 フレームの MFCC を推定する．そして，これらの MFCC に対してニューラルネットワークによる歌声変換を行う．この結果得られる近似的な MFCC パラメータ 5 フレームを使って Δ パラメータを推定する．$\Delta\Delta$ パラメータも同様の方法で推定が可能である．
- 歌声 → 朗読データ変換
  （sung-speech to read-speech; s2r）

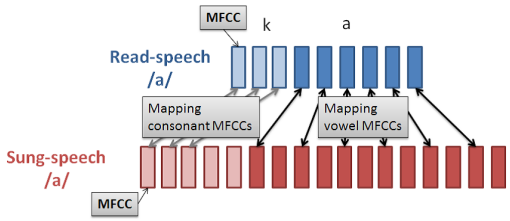

図 1 朗読-歌声音声のアライメント方法

歌声のテストセットに対して歌声 → 朗読変換をする．歌声 → 朗読変換ネットワークは次の 2 つの方法で作成した．1 つ目は，12 ユニットの入力層に対して歌声音声の MFCC12 次元を用いる場合である（s2r$_{\mathrm{mfcc}}$）．2 つめは，13 ユニットの入力層に対して歌声音声の MFCC12 次元+ピッチ情報を用いる場合である（s2r$_{+\mathrm{pitch}}$）．ピッチ情報をネットワークの学習データに用いることで，ピッチに依存したより信頼できる歌声 → 朗読変換が期待できる．ピッチの推定は Praat[8] を用いて 75-600Hz の範囲に制限して推定した．そしてピッチが推定できた区間のみを使ってネットワークの学習を行った．ピッチの抽出ができない歌声区間では，入力ユニット数 12 のネットワークを用いる．

3.3 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．**図 2** に示すように，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．ただし，計算コストの問題より 1 単語あたりの音節数が 10 以下の単語しか発音変形を追加していない．つまり，1 単語あたり最大 1024 種類の発音が割り当てられることになる．

3.4 デコーダ

歌声の歌詞認識は，一般的な音声認識と同じ方法で，音

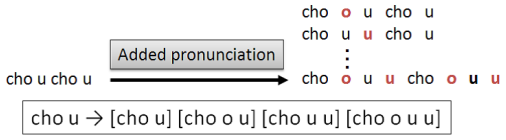

図 2 発音拡張

声特徴量 $\boldsymbol{o}$ が分かっている時，最も生起確率の高いラベル系列 $\boldsymbol{l}$ を推定する問題と言える．この問題は式 (3) で定式化される．

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} \log P(\boldsymbol{o}|\boldsymbol{l}) + \alpha \log P(\boldsymbol{l}) + \beta|\boldsymbol{w}| \qquad (3)$$

デコーダは音響モデル，言語モデル，発音辞書によって構成されている．言語重み α（言語スコアと音響スコアのバランスを調整するもの）とワードペナルティ β（単語が多数挿入されることを防ぐもの）を適切に設定することで歌声の認識性能を向上できる．デコーダは SPOJUS++[9] を用いた．

4. 実験

4.1 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 $\Delta, \Delta\Delta$，対数パワーの $\Delta, \Delta\Delta$ の計 38 次元である．認識実験の評価は 2 節で説明した男性ボーカル 7 名 7 曲のテストセットを用いた．このテストセットは 19 分の歌声が含まれている．

ベース音響モデルは，CSJ[10] 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 話者 222K 発話分である．928 コンテキストの左コンテキスト依存 64 混合 GMM-HMM も同様の学習データを用いて学習した．

MAP 適応学習は，2 節で説明した男性ボーカル 40 名 40 曲を用いる．朗読-歌声 MFCC 変換のニューラルネットワークは，2 節で説明した 4 名 6 ペアの朗読-歌声音声を使って，3.2.3 節で説明した 3 つの方法を適用した．

言語モデルは，音節 3 グラム言語モデル，単語 2 グラム，3 グラム，4 グラム言語モデルを作成した．学習に用いたデータは，1991 年から 1994 年までの毎日新聞記事 (45 ヶ月，206.7M 単語分) と 2 節で説明した 130K の歌詞で構成される歌詞データベースである．言語モデルの音節単位の認識実験で使う発音辞書は，学習データに存在する全 116 音節を使用する．大語彙歌詞認識の実験で使う発音辞書は，言語モデルの学習データから頻出上位 2 万語を使用する．

デコーダのパラメータは，言語重み α：1，10，15，20，25，30 とワードペナルティ β：-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値を選択している．

表 2 言語モデルの性能評価

学習データ	N-gram	OOV[%]	PP
新聞	2	13.1	218
	3	13.1	194
	4	13.1	210
歌詞	2	1.8	134
	3	1.8	107
	4	1.8	114

表 3 コンテキス独立（CI）モデルと左コンテキスト依存（CD）モデルの認識結果 - 音節認識精度 [%]

Base		MAP20		MAP40	
CI	CD	CI	CD	CI	CD
20.4	20.7	34.2	30.6	36.2	33.4

4.2 実験結果

4.2.1 歌詞言語モデルの評価

言語モデルごとのテストセットの歌詞に対する未知語率 (OOV) とパープレキシティ (PP) を**表 2** に示す．この結果，歌詞コーパスを使用した単語 3 グラム言語モデルが未知語率 1.8%, パープレキシティ 107 で最も良い性能を示した．新聞コーパスの言語モデルによる新聞の文に対するパープレキシティは 50-100 程度なので，言語モデルとしては十分な精度といえる [11]．

4.2.2 MAP 適応による音節認識実験結果

続いて音響モデルの評価のために男性ボーカル 7 名 7 曲に対して音節単位の認識実験を行った．言語モデルには歌詞コーパスで学習した音節 3 グラム言語モデルを用いた．

コンテキスト独立モデル (CI) とコンテキスト依存モデル (CD) を使ってテストセットに対する音節認識精度を**表 3** に示す．Base は朗読音声 (CSJ) で学習したベース音響モデルである．MAP20 は 20 名 20 曲，MAP40 は 40 名 40 曲の適応データでベースモデルに対して MAP 適応している．コンテキスト独立モデルとコンテキスト依存モデルを比較したところ，Base ではコンテキスト依存モデルのほうが高い認識精度を示している．一方，MAP 適応した場合はコンテキスト独立モデルのほうがより高い認識精度を示しており，36.2%の認識精度を得た．適応データの量を 2 倍にしてもコンテキスト独立-依存モデルの認識精度は同じくらい差がある．よって，以降の実験ではコンテキスト独立モデルを使用する．

4.2.3 朗読-歌声変換による音節認識実験結果

4 名 6 曲の朗読-歌声音声ペアより 3 名を学習データ，1 名をテストデータとしてニューラルネットワークの学習を行った場合の学習回数ごとの平均二乗誤差（MSE）を**図 3** に示す．学習を繰り返す度に誤差が小さくなっている．

ニューラルネットワークで朗読-歌声変換した場合の音節認識精度を**表 4** に示す．尤度は，言語重みとワードペナルティを揃えた場合の対数尤度のフレーム平均を表す．テ

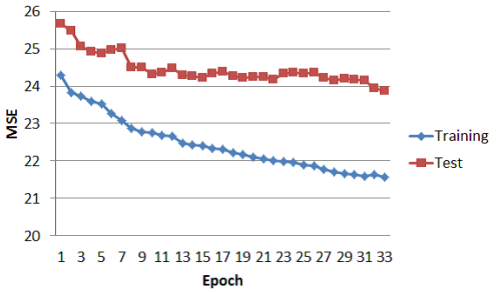

図 3 朗読 → 歌声変換ネットワークの学習曲線

表 4 NN による朗読-歌声変換したモデルやテストセットで認識した結果 - 音節認識精度 [%] と対数尤度のフレーム平均

モデル	テストセット	Acc	尤度
CI	JPOP	20.4	-63.7
CI	$\text{JPOP}_{\text{s2r}_{\text{mfcc}}}$	18.0	-56.4
CI	$\text{JPOP}_{\text{s2r}_{+\text{pitch}}}$	18.5	-56.3
$\text{CI}_{\text{r2s}_{\text{d}}}$	JPOP	19.5	-64.3
$\text{CI}_{\text{r2s}_{\text{m}}}$	JPOP	19.4	-64.1

ストセット JPOP を用いて，朗読モデル CI とニューラルネットワークによる朗読 → 歌声変換モデル $\text{CI}_{\text{r2s}_{\text{d}}}, \text{CI}_{\text{r2s}_{\text{m}}}$ を比較したところ，朗読モデルの方が認識精度 20.4%で朗読 → 歌声変換モデルよりも 1%程精度が高いことが分かった．また朗読モデル CI を用いて，歌声テストセット（JPOP）を認識した場合とニューラルネットワークによる歌声 → 朗読変換したテストセット（JPOP_{s2r}）を認識した場合を比較したところ，歌声テストセットをそのまま認識した方が歌声 → 朗読変換後の音声を認識するより 1-2%精度が高いことが分かった．

それぞれの対数尤度のフレーム平均値を比較したところ，歌声テストセットを用いた場合と比較して歌声 → 朗読変換したテストセットの方が尤度が大きいことがわかった．対数尤度値を見る限りでは歌声 → 朗読変換は上手くできていると思われるが，認識精度が改善しなかった理由の解明には至らなかった．

歌声テストセットを歌声 → 朗読変換する時，ピッチ情報を入力層に含めたネットワーク（$\text{s2r}_{+\text{pitch}}$）ほうが含めないネットワーク（$\text{s2r}_{\text{mfcc}}$）より認識精度，尤度共に高いことが分かり，期待通りの結果となったが，変換処理を行わないベースモデルより悪くなった．ベースモデル (CI) と朗読 → 歌声変換したモデル (CI_{r2s}) を比較するとベースモデルの方が尤度が大きいことが分かった．

MFCC の $\Delta, \Delta\Delta$ パラメータを近似変換する方法を試みたが認識精度の改善は得られなかった．以上より，いずれのネットワークも単純にモデルやテストセットに適用するだけでは効果が得られないことが分かった．

表 5 音節 3 グラム言語モデルを用いて歌詞認識をした結果 - 音節認識精度 [%]

発音辞書	Base			MAP40		
	CI	$\text{CI}_{\text{r2s}_{\text{d}}}$	$\text{CI}_{\text{r2s}_{\text{m}}}$	CI	$\text{CI}_{\text{r2s}_{\text{d}}}$	$\text{CI}_{\text{r2s}_{\text{m}}}$
なし	20.4	19.8	19.4	36.2	35.7	35.2
あり	28.8	27.9	28.1	**46.1**	45.8	45.2

表 6 単語 3 グラム言語モデルを用いて大語彙歌詞認識をした結果 - 単語認識精度 [%]

発音拡張	LM	Base			MAP40		
		CI	$\text{CI}_{\text{r2s}_{\text{d}}}$	$\text{CI}_{\text{r2s}_{\text{m}}}$	CI	$\text{CI}_{\text{r2s}_{\text{d}}}$	$\text{CI}_{\text{r2s}_{\text{m}}}$
なし	新聞	4.4	3.9	4.2	11.0	11.4	11.3
	歌詞	9.1	7.1	7.2	22.9	22.6	23.8
あり	新聞	5.6	4.7	5.4	12.5	12.1	11.3
	歌詞	11.4	10.5	10.2	25.0	**25.9**	25.4

4.2.4 MAP 適応や発音辞書の拡張による音節認識実験結果

CI, $\text{CI}_{\text{r2s}_{\text{d}}}, \text{CI}_{\text{r2s}_{\text{m}}}$ に対して MAP 適応や発音拡張をした場合の音節認識精度を**表 5** に示す．ベースモデルに対して MAP 適応をした CI とニューラルネットワークによる朗読 → 歌声変換をしたモデルに対して MAP 適応をした CI_{r2s} を比較すると，ベースモデルの方が認識率がより向上した．発音拡張辞書を用いるといずれの場合も認識精度が向上しており，MAP 適応をした CI が音節認識精度で 46.1% (音素認識精度 59.0%) となり，最も良い性能を示した．

4.2.5 大語彙歌詞認識実験結果

大語彙歌詞認識の実験結果を**表 6** 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．新聞言語モデルを使用した結果より歌詞言語モデルを使用した結果のほうが全体的に単語認識精度が 4-13%向上している．朗読音響モデルを歌声音響モデルに変更することで単語認識精度が更に 5-14%向上している．発音辞書の拡張前後では，いずれの場合でも単語認識精度が 1-3%向上している．この理由を調査するため，言語重みとワードペナルティを揃えて発音辞書の拡張前後の音響スコアと言語スコアを比較した．この結果，音響スコアはほとんど同じだったのに対して言語スコアに大きな差があることがわかった．このことから音響的に正しい音節列を推定できても歌声に頻出する「音韻の引き延ばし」の影響で誤った単語が推定されやすくなっていることが分かる．最もよい性能を示したのは歌詞単語 3 グラム言語モデル，MAP 適応した $\text{CI}_{\text{r2s}_{\text{d}}}$，そして発音拡張辞書を用いたシステムで単語認識精度 25.9%となった．

4.2.6 話者適応による認識実験結果

テストセットと同じボーカルで異なる楽曲を使って話者適応をした場合の音節-単語認識精度を**表 7** に示す．Base_{SPK} はベースモデル Base に対して 1 曲の話者適応データで MAP 適応している．$\text{MAP40}_{\text{SPK}}$ は歌声適応モデ

ル MAP40 に対して 1 曲の話者適応データで MAP 適応している．曲 1 は話者適応データが見つからなかったため話者適応学習はやっていない．いずれの楽曲でも話者適応モデル Base_{SPK} より歌声適応モデル MAP40 の方が高い認識精度を示した．一方で歌声適応モデルに対して話者適応したモデル $\text{MAP40}_{\text{SPK}}$ は歌声適応モデル MAP40 より平均的に高い認識精度を示した．しかし歌声適応モデルに対して話者適応をした場合，曲 2 では大きく認識精度が低下している．この理由は事前分布 (MAP40) で既に高い認識精度を示していたため，MAP 適応によって逆にモデルが表現できる空間を狭めてしまったためだと考えられる．

表 7 話者適応によるテストセットの認識結果 [%] 括弧内は 1-7 曲の平均

(a) 音節認識精度

曲	Base	Base_{SPK}	MAP40	$\text{MAP40}_{\text{SPK}}$
1	18.4	-	-	-
2	25.4	40.3	64.1	59.3
3	21.8	26.8	34.7	36.1
4	17.0	26.0	35.2	37.5
5	18.3	26.0	27.5	31.4
6	19.0	28.8	36.5	36.5
7	24.6	34.9	38.0	37.8
平均	20.8(20.4)	31.5	41.3	41.3

(b) 単語認識精度

曲	Base	Base_{SPK}	MAP40	$\text{MAP40}_{\text{SPK}}$
1	4.4	-	-	-
2	20.9	18.9	37.3	36.8
3	8.8	12.2	22.4	22.9
4	8.2	10.7	24.3	24.7
5	3.6	9.0	9.7	10.0
6	8.0	20.7	19.8	22.4
7	12.4	9.6	14.7	18.6
平均	9.8(9.1)	13.4	20.9	22.0

5. 関連研究との比較

類似研究の文献 [5] では，歌声適応や歌詞言語モデルを使ったシステムを提案している．MLLR による歌声適応をした GMM-HMM と歌詞テキストで学習した音素 2 グラム-単語 2 グラム言語モデルを用いた場合，英語の歌詞認識における伴奏なし男性歌唱の音素認識精度は 34.9%，伴奏なし男女の歌唱の単語認識精度は 12.4%であった．また文献 [6] では，歌声適応やフレーズの繰り返し部分を統合して出力するシステムを提案している．英語の歌詞認識における伴奏なし男女歌唱の音素認識精度は 26.99%，単語認識精度は 9.48%であった．いずれの研究とも実験条件が異なるため単純な比較はできないが，提案システムがより高い認識精度を示している．

6. おわりに

本稿で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．歌詞コーパスを用いた言語モデルは単語未知語率が 1.8%，パープレキシティが 107 となり歌詞言語モデルの使用が有効であることが分かった．音響モデルは歌声 MAP 適応，ニューラルネットワークによる朗読 → 歌声変換が有効であることが分かった．ニューラルネットワークによる歌声変換は歌声変換した朗読音声による音響モデルの再学習よりも朗読モデルの歌声変換の方が良かった．但し，歌声 MAP 適応と併用することで初めて効果が発揮されるという結果になり，今後はこの理由を調査する必要がある．歌声適応と併用して話者適応をすることによって認識性能を改善できた．最終的な結果としては，音節認識精度 46.1%（音素認識精度 59.0%），単語認識精度が 25.9%となり類似研究 [5], [6] より高い認識結果が得られた．

今後は，朗読-歌声音声のペアの増加，朗読-歌声変換が単体では上手くいかない理由の調査，DNN-HMM を用いたシステムの性能改善を図りたい．

参考文献

[1] kai Wang, C., Lyu, R.-Y. and Chiang, Y.-C.: An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker., *INTERSPEECH*, ISCA, (online), (2003).
[2] 佐宗晃 他: ARHMM に基づいた音声分析手法と歌声認識による評価, 電子情報通信学会 技術研究報告, SP2005-45, vol.105, no.199, pp.19-24 (2005).
[3] 佐宗晃, 後藤真孝: 音声認識方法および音声認識装置, 特許第 4576612 号．(2007.03.29)．
[4] Hosoya, T. et al.: Lyrics Recognition from a Singing Voice Based on Finite State Automaton for Music Information Retrieval, *in Proc. ISMIR, 2005*, pp. 532–535 (2005).
[5] Mesaros, A. and Virtanen, T.: Automatic Recognition of Lyrics in Singing, *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2010, No. 1, p. (2010).
[6] 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* (2014).
[7] Young, S. J. et al.: *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK (2006).
[8] Boersma, P. and Weenink, D.: Praat: doing phonetics by computer. (accessed 11th Nov 2014).
[9] Fujii, Y., Yamamoto, K. and Nakagawa, S.: Large Vocabulary Speech Recognition System: SPOJUS++, MUSP, pp.110-118 (2011).
[10] Maekawa, K. et al.: Spontaneous speech corpus of japanese., in *LREC*. 2000, European Language Resources Association.
[11] Naptali, W., Tsuchiya, M. and Nakagawa, S.: Topic-Dependent-Class-Based n -Gram Language Model, *Audio, Speech, and Language Processing, IEEE Transactions on*, Vol. 20, No. 5, pp. 1513–1525 (online), (2012).