Vol.2015-MUS-109 No.11
2015/11/8

DTWに基づく歌唱音声からの歌唱誤り検出の検討

宮川 功[1]　中村 友佑[1]　能勢 隆[1]　伊藤 彰則[1]

概要：カラオケ歌唱音声中から，歌唱した歌詞の誤り箇所を検出する方法を検討する．提案法では，あらかじめ用意した標準歌唱と入力歌唱とをDTWによって対応付け，フレームごとの距離を観測することによって誤り箇所を発見する．しかし，この方法によるDTW距離の絶対値は標準歌唱と入力歌唱の話者性の影響を受けるため，線形変換によって話者性の正規化を試みた．実験の結果，歌詞を大きく誤っている場合には高い精度で検出が可能であった．

Singing Error Detection from the Singing Voice Based on Dynamic Time Warping

Isao Miyagawa[1]　Yusuke Nakamura[1]　Takashi Nose[1]　Akinori Ito[1]

Abstract: We investigate a method of detecting wrong lyrics from singing voice for karaoke. In the proposed method, we compare input singing voice and reference singing voice using dynamic time warping, and then observe the frame-by-frame distance to find the error location. However, the absolute value of the distance is affected by the speaker individuality of the reference and input singing voices. Thus, we attempted to normalize the speaker individuality by linear transformation. The results of the experiment showed that we could detect the wrong lyrics with high accuracy when the different part of the lyrics was long.

1. はじめに

近年，カラオケ文化は老若男女を問わずに流行しており，多くの人がカラオケを楽しむようになってきている．みんなで行って盛り上がる人も居れば，一人で歌ってストレスを発散するという人もいるだろう．特に最近では，採点機能も充実してきており，テレビ番組などでもよく取り上げられるようになってきている．

その中でも，一人でカラオケに行く場合には，歌詞を間違えていたときに気づけないことが多くあるだろう．一人で歌っているときに歌い間違えている部分を指摘してもらうことで，歌詞を修正し, 正しい歌詞で歌うことができるようになる．カラオケ教室や歌唱練習などにも使うことができる．また歌い間違いの検出ができれば，採点ゲームとは別の新しいゲームの開発にもつながり，採点ゲームの採点基準の一要素ともなりえる．

カラオケ音声の評価についてはいくつかの研究があるが，多くは音楽的な特徴か声質の評価を行っている．例えば，歌唱の音高・音長の正確さ [1]，歌唱テクニックを含めた歌唱のうまさ [2]，熱唱度 [3] などである．しかし，これらの研究では歌詞を全く扱っていない．そもそも歌唱音声中から歌詞を認識することは非常に難しく [4]，歌詞を歌唱評価のために使う試みはこれまで行われていなかった．

歌唱音声の中から歌詞の誤りを検出する方法としては，入力歌唱と歌詞のテキスト情報から誤りを検出する方法が考えられる．例えばHMMを使って正しい歌詞のモデルを作り，入力音声の確率を計算する [5] ことで歌詞誤りの検出が可能と考えられる．しかし，HMMを使って高精度に歌唱音声を評価するためには話者適応を行う必要があり，カラオケボックスなどのように不特定の歌唱者に対応するのが難しい．一方，一部のカラオケ機には，肉声や合成音声によるボーカルパートを伴奏と同時に演奏するガイドボーカル機能 [6] があり，近年ではその台数・曲数も増えつつある．この場合には，正しい歌詞で歌唱した音声が利用できるので，これを正解として利用することにより，歌詞テキストを用意せずに歌詞誤りが検出できるのではない

[1] 東北大学
Tohoku University

かと考えられる．ガイドボーカルを利用する方法には，誤り検出システムが言語に依存しないという利点もある．したがって本研究では，ガイドボーカルを標準歌唱音声として，これと入力歌唱音声とをマッチングすることにより歌い間違いを高精度で検出することを目指す．

2. 提案手法

2.1 手法の概要

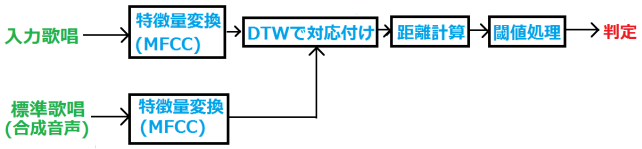

図 1 提案手法

提案手法の概略を図 1 に示す．特徴パターンの似ている程度を測る最も基本的な方法は，ベクトル間距離を用いる方法である．入力歌唱音声のあるフレームの特徴ベクトル $\boldsymbol{y}$=$(y_1 \cdots y_n)$ と，標準歌唱音声の対応するフレームの特徴ベクトル $\boldsymbol{r}$=$(r_1 \cdots r_n)$ のユークリッド距離は

$$d(y,r) = \sqrt{\sum_{i=1}^{n}(y_i - r_i)^2} \tag{1}$$

で求まる．これを利用し，距離が大きい場合はその部分に歌詞誤りがあると判定する．特徴量としては，音声処理に一般に用いられる MFCC を利用する．しかし、標準音声と入力音声の系列長が違う場合には，2 つの系列の特徴ベクトル間の対応を求める必要がある．そこで，非線形マッチング手法である DTW で 2 つの系列を対応付けし，距離を測る．

よって，今回は，標準歌唱と入力歌唱を MFCC とパワーの系列に変換し，特徴量間のユークリッド距離に基づく DTW によって，2 つの音声を対応付け，対応付けられたフレームごとにフレーム間の距離を計算し，距離が大きい場合に歌唱誤りとして検出を行うという方法で実験を行う．今回は曲全体を DTW で対応付けた．

2.2 DTW

DTW(Dynamic Time Warping) は，系列の非線形伸縮によって 2 つの系列間を対応付ける手法である．入力音声と標準音声の特徴量系列を x_1...,x_N および y_1...,y_N とする．このとき，入力音声 1~i フレーム，標準音声 1~j フレームの累積距離 g(i,j) を次のように求める．

$$d(i,j) = ||x_i - y_i||^2 \tag{2}$$

$$g(1,1) = d(1,1) \tag{3}$$

$$g(i,j) = d(i,j) + \min \begin{Bmatrix} d(i,j-1) + g(i-1,j-2) \\ g(i-1,j-1) \\ d(i-1,j) + g(i-2,j-1) \end{Bmatrix} \tag{4}$$

また，この時に最小値を与える対応を動的計画法によってバックトレースし，入力音声とガイドボーカルを対応付けることができる．

3. 特徴量変換と平滑化

標準歌唱と入力歌唱を DTW で対応付けてフレーム間距離を求めると，異なる歌詞を歌っている部分では距離が大きくなるが，声質が異なる場合は，正しい歌詞の箇所も距離が大きくなる．図 2~ 図 4 は標準歌唱と 3 つの誤り歌唱 (いずれも合成音声) をそれぞれ DTW で対応付けてフレーム間距離を求めたグラフである．縦軸がフレーム間の 2 乗距離で，横軸がフレーム数である．図 2 は合成音声の同一歌唱者の場合，図 3 は性別が異なる歌唱者でキーが同じ場合，図 4 は性別が異なる歌唱者でキーが 1 オクターブ下の場合である．誤り箇所のフレームは色を変えている．図 2 では同一歌唱者の場合は 5500~7000 フレームあたりの誤り箇所でのみフレーム間距離が大きくなり，正解箇所とはっきりと違いが現れていた．しかし，図 3 では誤り箇所以外のフレーム間距離がやや大きくなり識別がしにくい．図 4 では誤り箇所以外のフレーム間距離がさらに大きくなって識別ができない．したがって，声質が異なる歌唱者にも対応できるようにするため，特徴量の線形変換に基づく方法 [7] を検討した．入力歌唱の特徴ベクトル系列を x_1...,x_N，標準歌唱の特徴ベクトル系列を y_1...,y_N とし，これらの間に DTW で 1 対 1 の対応がついているとする．このとき，$y_i = Ax_i + e_i$ として，誤差ベクトル e_i の 2 乗和が最小になるように変換行列 A を定めると

$$Z = \sum_i ||e_i||^2 = \sum_i ||y_i - Ax_i||^2 = \sum_i (y_i - Ax_i)^T (y_i - Ax_i) \tag{5}$$

として Z を最小化する．

$$\frac{\partial}{\partial A} \sum_i (y_i^T y_i - 2x_i^T A^T y_i + x_i^T A^T x_i) = 0 \tag{6}$$

これを計算すると

$$\sum_i (-2y_i x_i^T + 2Ax_i x_i^T) = 0 \tag{7}$$

したがって

$$C_{xx} = \sum_i x_i x_i^T, C_{yx} = \sum_i y_i x_i^T \tag{8}$$

とおくと

$$A = C_{yx} C_{xx}^{-1} \tag{9}$$

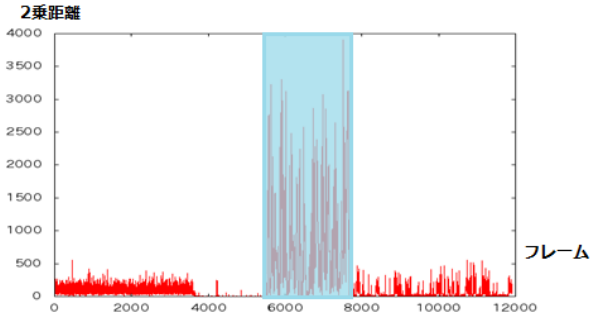

図 2　2 乗距離 (異邦人，女性-女性誤り，同キー)

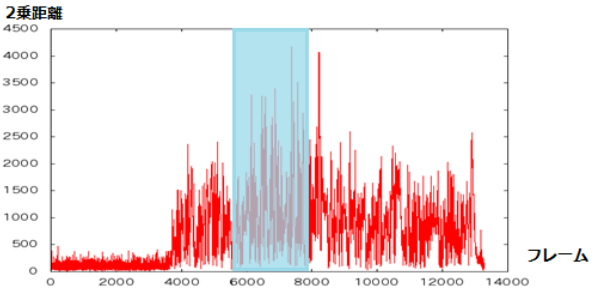

図 3　2 乗距離 (異邦人，女性-男性誤り，同キー)

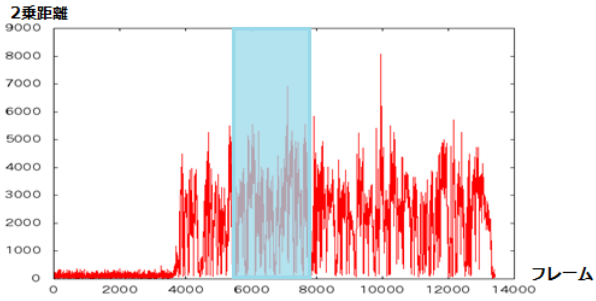

図 4　2 乗距離 (異邦人，女性-男性誤り，1 オクターブ下)

と求まる．これは x と y が完全に同じ歌詞を歌唱していることを前提としているが，実際の歌唱には歌詞誤りが含まれる．特徴量変換の際に, 誤り歌唱部分を用いるのは，無理やり異なる音素同士を対応付けてしまうために不適切である．そこで，同じ歌詞を歌っている部分だけを特徴量変換に用いるために，フレーム間距離が閾値以上のデータを計算から除外することで，誤り歌唱を無理やり対応付けることを抑える．

さらに，特徴量変換を繰り返し用いることとした．1 回目の特徴量変換で, 元々完全に判別不能である状態からある程度だけ誤り部分と正しい部分が分かれる．そこから閾値によって正解箇所だけを変換に用いて誤り箇所をはっきりさせていくことで，性能がより向上するのではないかと考えた．

特徴量変換を行った後の 2 乗距離の例を図 5, 6 に示す．図 5 は図 3 に対応し，図 6 は図 4 に対応している．どちらも DTW と特徴量変換を 3 回繰り返してあり，閾値には 3000 を用いている．図 5, 6 より図 3, 4 と比べて，5500〜7000 フレームあたりの誤り箇所の距離が大きく現れてきているのが分かる．この例からも，DTW と特徴量変換の繰り返しには効果があることが分かる．一方，誤りがない部分にも距離のピークが多数出現していることがわかる．そこで，実際に検出に用いる際には，誤り箇所以外における距離ピークの影響を小さくするために，移動平均フィルタによる平滑化を行う．オクターブ違いの場合のグラフ (図 6) から 200 フレームの移動平均フィルタで平滑化した結果を図 7 に示す．

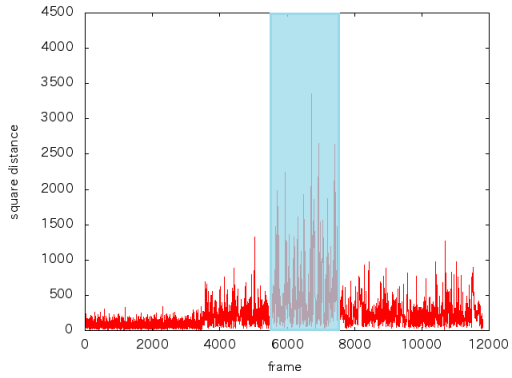

図 5　特徴量変換後の 2 乗距離 (異邦人，女性-男性誤り，同キー)

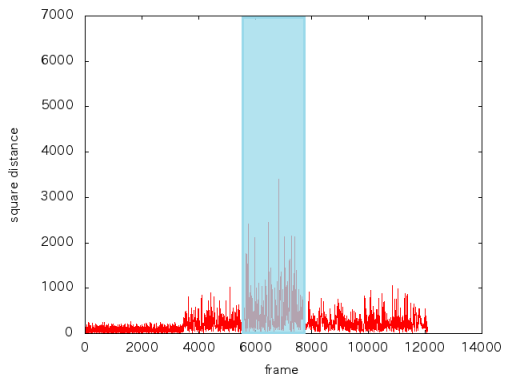

図 6　特徴量変換後の 2 乗距離 (異邦人，女性-男性誤り，1 オクターブ下)

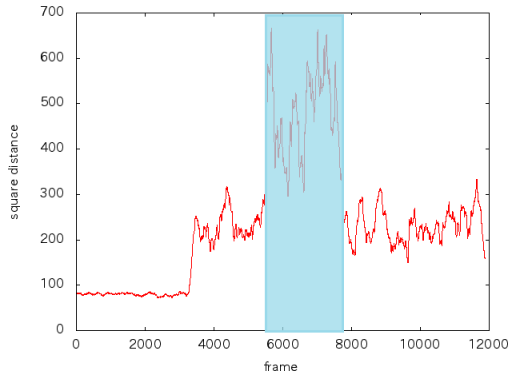

図 7　特徴量変換後の 2 乗距離 (図 6) の平滑化距離 (200 フレーム)

4. 最適な閾値の検討

特徴量変換を行う際に使用する最適な閾値を求める実験を行った．実験に使用した楽曲を表 1 に示す．これらの合成音声の歌詞の誤り方は「本来は 1 番の歌詞で歌うべきと

ころを 2 番の歌詞で歌った」というものである．また、「異邦人」と「粉雪」は誤り箇所が大きくまとまっているのに対し，「雪の華」と「やさしくなりたい」は細かく誤っている．これらの曲を，男性 1 名が合成音声と同じ歌詞で歌唱した．距離がほぼ 0 に近くなる伴奏区間をカットし，閾値として距離全体の平均の α 倍を用いた．DTW と特徴量変換は 3 回繰り返した．倍率 α を 0.3〜3.0 まで 0.1 刻みで変えて行い，距離の変化を調べた．音声分析条件は，サンプリング周波数 16kHz，MFCC 算出の窓長 25ms，フレームシフト 10ms，MFCC 次数は 13(パワー項も含む) である．ΔMFCC は利用していない．

表 1 閾値を求める実験で使った楽曲

曲名	標準歌唱の性別
異邦人	女性
粉雪	男性
雪の華	女性
やさしくなりたい	男性

倍率 α を 0.3~3.0 にした時の標準歌唱と入力歌唱のフレーム間の距離の変化を調べた．その結果を表 2 に示す．倍率を変えた場合の距離の変化を観察し，結果を「正解箇所の距離が大きくて、誤りの識別ができない」,「正解箇所の距離のみが小さくなって誤りの識別ができる」,「誤り箇所の距離も小さくなって誤りの識別ができない」の 3 つに分類した．「異邦人」の場合は，閾値が平均の 0.3~0.9 倍の時，正解箇所の距離が減少せずに誤りの識別ができなかった．倍率が 1.0~2.0 の時は正解箇所の距離のみが減少し誤りの識別ができた．倍率が 2.1~3.0 の時は誤り箇所の距離も減少し，誤りの識別ができなかった．表 2 の中には「該当なし」があるが，これは倍率が 3.0 の時でも誤り箇所の距離が減少しなかったことを示す．倍率をさらに大きくすれば，誤り箇所の距離も減少すると思われる．閾値が小さい場合には特徴量変換に使えるデータが少ないために正解箇所の距離も大きくなっていると考えられる．ある程度閾値が大きくなったところで，特徴量変換に十分なデータを使えるようになるため，正解箇所の距離は倍率を前後させても変化しなくなったと考えられる．閾値が大きすぎる場合には，誤り箇所も特徴量変換に用いられたため，誤り箇所の距離も小さくなってしまったと考えられる．

よって，正解箇所の距離が小さくなり，誤り箇所の距離が大きくなる閾値が適切であると考えられる．倍率 α を 1.5 程度にすればすべての曲についてこの領域になるため，以後は距離平均の 1.5 倍程度が閾値として適切であることが分かった．ただし，今回は男性 1 名の歌唱音声だけで行っているため，他の歌唱者について適した倍率に違いがあるのかは検討の余地がある．

表 2 閾値を求める際の倍率を変化させた時の距離の変化

	正解箇所の距離が減少せず	正解箇所の距離のみが減少	誤り箇所の距離も減少
異邦人	0.3〜0.9	1.0〜2.0	2.1〜3.0
異邦人キー下	0.3〜1.4	1.5〜3.0	該当なし
粉雪	0.3〜0.9	1.0〜1.7	1.8〜3.0
雪の華	0.3〜0.9	1.0〜3.0	該当なし
雪の華キー下	0.3〜1.4	1.5	1.6〜3.0
やさしくなりたい	0.3〜0.8	0.9〜3.0	該当なし

5. 実験

5.1 歌唱誤り検出実験

次に，実際に提案手法で歌詞が誤っている箇所を正確に検出できるかどうか調べるために，歌唱誤りを検出する実験を行った．入力歌唱として，男性 (大学生・大学院生)11 人に歌唱をしてもらった. 楽曲としては，JOYSOUND の男性が歌いやすい曲ランキング上位の曲から，「世界にひとつだけの花」と「空も飛べるはず」の 2 曲を選択した．対象者には 2 曲を正しい歌詞で歌唱してもらった．

標準歌唱には，男性の合成音声を用いた．標準歌唱には「正しい歌唱」,「文章・文単位の誤りを含む歌唱」,「単語・一文字単位の誤りを含む歌唱」の 3 通りを用意した．本来は入力歌唱に誤りが含まれるが，今回は標準歌唱を変えることで歌唱誤りのシミュレーションをしている．誤り歌唱にはそれぞれ複数個の誤り箇所が設けられている．誤りは全楽曲分あわせて，文章・文単位の誤り 8 箇所と，単語・一文字単位の誤り 9 箇所，合計で 17 箇所の誤りがある．「世界に一つだけの花」には，文章単位の誤り 2 箇所，短い文単位の誤り 4 箇所が含まれる．「空も飛べるはず」には，短い文単位の誤り 2 箇所，単語・一文字単位の誤り 9 箇所が含まれる．11 人分全てについて行ったとき，187 個の誤りが検出されれば正しいということになる．

これらのデータを用いて，検出実験を行った．DTW と特徴量変換の繰り返しは 3 回とした．その後，あらかじめ設定された平滑化フレーム数で平滑化を行う．平滑化した後，平滑化距離全体の平均の α 倍を閾値とし，その閾値よりも距離が大きな箇所を誤り箇所として検出する．誤り検出箇所はフレーム区間として与えられるので，検出されたフレーム区間と実際に誤っているフレーム区間に重なりがあれば正しく検出されているとみなした．1 つの誤りフレーム区間に対して 2 つの区間が検出された場合には，1 つの正解検出としてカウントする．実際の誤り箇所以外を検出した場合には誤検出とする．平滑化フレーム数と倍率 α を変化させて，検出の再現率と適合率を求め，その調和平均である F 値を調べた．

平滑化フレームと倍率を変化させた時の F 値の変化を図 8 に示す．横軸が倍率 α，縦軸が平滑化フレーム数である．F 値が最も高いのは α=1.9 倍，平滑化が 210 フレームの

時であり，F 値が 0.61 とあまり芳しい結果ではなかった．このような結果となった原因には，1 文字の誤りがうまく検出できていなかったことが考えられる．そこで，誤りが「文章・文単位」の大きいものと「単語・一文字単位」の小さいものに分けて実験を行った．大きい誤りの検出結果を図 9，小さい誤りの検出結果を図 10 に示す．大きな誤りを含む歌唱での F 値の最大値は 0.9，小さな誤りを含む歌唱での F 値の最大値は 0.23 となった．1 文字の誤りによって生じる距離の増大よりも，3 章のグラフにも生じていた極大点のほうが大きくなってしまうことにより，閾値が低いとどちらも検出され，閾値が大きいとどちらも検出されなくなってしまい，その結果 F 値は小さくなってしまったと考えられる．細かすぎる誤りをこの方法で検出するのは難しいと考えられる．

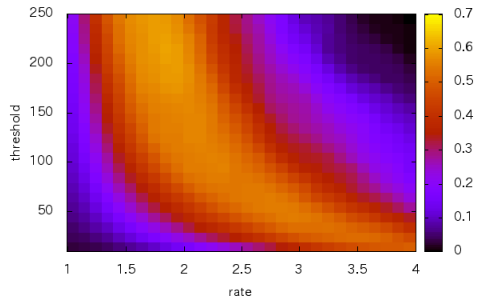

図 8 検出結果の F 値の結果

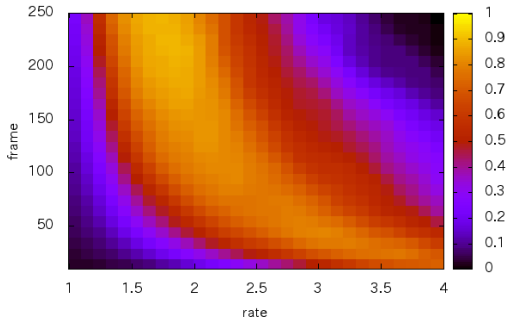

図 9 大きな誤りの曲の F 値の変化

5.2 曲全体での歌唱誤り識別実験

誤り箇所の検出は難しいことが分かったので，曲全体での誤りの有無を識別する実験を行った．前節の検出実験は，検出結果と実際の歌詞が誤っている箇所とを比較して，F 値を算出したが，本節では，1 曲の中で誤りが検出されるかどうか調べ、本来の識別結果を比較して，F 値を算出した．楽曲データは前章と同じである．66 通りある合成音声標準歌唱と人間歌唱音声の組み合わせのうち，正解歌唱 22 個, 誤り歌唱 44 個と識別されれば正しいということとなる．

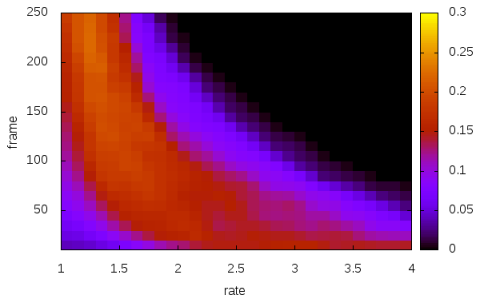

図 10 小さな誤りの曲の F 値の変化

標準歌唱と入力歌唱の距離を 2 章と同様に求め，あらかじめ決められた平滑化フレーム数で平滑化し，平滑化距離全体の平均の α 倍を閾値とする．閾値と平滑化距離の最大値を比較して，最大値のほうが大きければ誤りあり，最大値のほうが小さければ誤りなしと識別する．これを前章同様，平滑化フレーム数と閾値を変化させて F 値がどのようになるか調べた．

結果を図 11 に示す．最も高いところで F 値は 0.95 と高精度で識別できていた．最大値を用いることは，歌唱音声中の最もひどい箇所に注目していることになり，全ての歌唱には 1 文字の誤りと一緒に 3 文字以上の大きさの誤りが含まれているため，それが閾値を超えていたことが高精度の理由であると考えられる．3 文字以上の大きさの誤りに関してはかなりの高精度で検出できていると考えられる．しかし，歌唱音声中に 1 文字の誤りしか含まれていない場合にはこのようにうまく識別できるとは限らない．したがって，短い誤りを検出するためには，別の方法が必要であると考えられる．

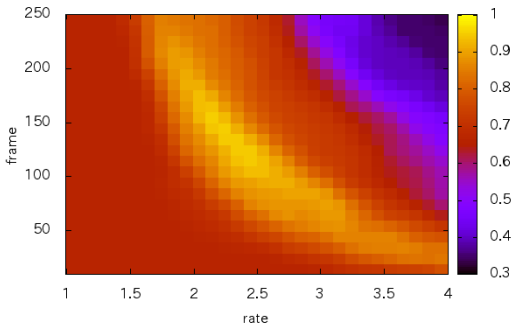

図 11 識別結果の F 値の結果

5.3 人による誤り検出の違い

人によって誤り検出に違いがあるかどうか調べた．その結果を表 3 に示す．検出方法は 5.1 節と同じである．

表 3 より F 値の最大値は人によって多少の差があるが，大きくは違わないことが分かる．

表 3　人それぞれの誤り検出実験の F 値の最大値

	F 値の最大値
歌唱者 1	0.69
歌唱者 2	0.67
歌唱者 3	0.62
歌唱者 4	0.62
歌唱者 5	0.57
歌唱者 6	0.60
歌唱者 7	0.65
歌唱者 8	0.67
歌唱者 9	0.63
歌唱者 10	0.62
歌唱者 11	0.62

5.4　特徴量変換の回数

これまでの実験では特徴量変換を 3 回繰り返していたが，変換回数として 3 回が適当なのかどうかは明らかではない．そこで，回数を変えて回数を変えて検出実験を行った．その結果を図 12 に示す．

図 12 より，異なる曲や異なる誤りの種類に対しても，F 値の最大値は回数によってさほど変化しないことが分かる．したがって，特徴量変換は 1 回で十分であることが分かった．ただし，今回の実験は入力歌唱・標準歌唱とも男性音声であったため，これらの性別が異なった場合については改めて実験が必要である．

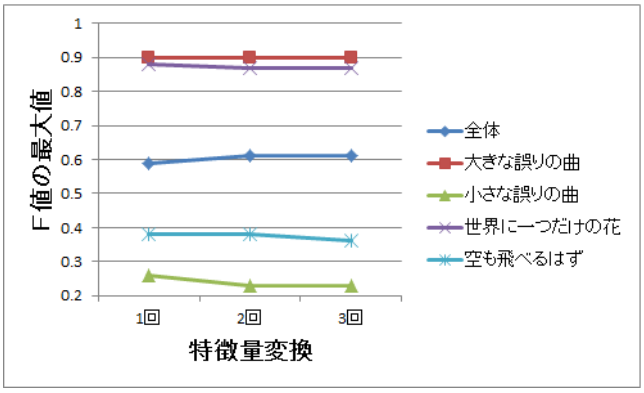

図 12　特徴量変換の回数を変えた時のそれぞれの F 値の最大値

6.　まとめ

歌唱音声中の誤りを検出する方法を提案した．DTW によって入力歌唱音声と標準歌唱音声を対応付け，特徴量変換で声質の違いによる影響を抑える．さらに平滑化で正解箇所に生じる極大点の影響の影響を抑えることで, ある程度まとまった大きさの歌詞の誤りについては高精度で検出できることがわかった．しかし，短い誤りに対しては, 高精度で検出することは難しそうであることもわかった．また、人によって誤り検出に違いがあるかどうか調べた結果，人による差は小さいことがわかった．また，特徴量変換の回数は 1 回で十分であることが分かった．

今後は，検出結果の分析を進めるとともに，様々な条件での性能を調べていきたい．

参考文献

[1] 竹内英世，保黒政大，梅崎太造：人の主観評価に近いカラオケ採点法，電気学会論文誌 C，Vol130，No.6，pp. 1042-1053，2010.
[2] 中野倫靖，後藤真孝，平賀譲：楽譜情報を用いない歌唱力自動評価手法，情報処理学会論文誌，Vol48，No.1，pp. 227-236，2007.
[3] Daido, R., Ito, M., Makino, S., and Ito, A. : Automatic evaluation of singing enthusiasm for karaoke, Compurter Speech&Language, Vol28，No.2，pp．501-517，2014.
[4] Mesaros, A. and Virtanen, T. : Automatic recognition of lyrics in singing, EURASIP Journal of Audio, Speech and Music Processing, Vol2010，article No.4，2014.
[5] Suzuki, M., Hosoya, T., Ito, A and Makino, S. : Music Information Retrieval from a Singing Voice Using Lyrics and Melody Information, EURASIP Journal on Advances in Signal Pricessing, Vol2007, doi:10.1155/2007/38727.
[6] 松下電器産業株式会社：カラオケ装置，特開 2001-42879，2001.
[7] Matsumoto, H. and Inoue, H. : A piecewise linear spectral mapping for supervised speaker adaptation, Proc. ICASSP, Vol.1，pp.449-452，1992.