5L-01

市販音楽CDを用いたユーザ歌唱に伴奏音が自動追従するスマートカラオケシステム

和田 雄介 † 中村 栄太 ‡ 糸山 克寿 ‡ 吉井 和佳 ‡

† 京都大学 工学部情報学科 ‡ 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

音楽の楽しみ方には歌を歌うことが含まれており，この楽しみ方の形態の一つにカラオケがある．カラオケ産業では，伴奏音の作成に，市販CD音源を専門家が耳で聴きながら手動で楽譜 (MIDIファイル) を書き起こすという方法が採られている．この方法では，1曲の伴奏音の生成に莫大な労力が必要であるだけでなく，MIDIを用いて合成した伴奏音の品質は元の市販CD音源に劣ってしまう．また，ある曲のテンポを自分の好みにアレンジして歌いたいと思ったときに，カラオケでは手動で伴奏音のテンポを変更しなければならない．このとき，ユーザは伴奏音のテンポを聴きながら自分の好みに合うかどうか判定し，テンポ設定を微調整する必要があり，手間がかかる．したがって，市販CD音源のみから伴奏音を生成し，伴奏音のテンポを自動でユーザ歌唱に追従させることができれば，ユーザは快適に歌うことができると考えられる．

これまで，市販CD音源からカラオケの伴奏音を生成し，それをユーザ歌唱に追従させるシステムがいくつか提案されている．例えば，楽譜及び歌詞情報が既知であるという前提のもとで，ユーザが歌唱のピッチを変更した際に，それに合ったピッチの伴奏音をMIDIを用いて合成するカラオケシステム [1] が提案されている．また，楽譜等の事前情報を用いず，市販CD音源からカラオケの伴奏音を生成し，そのピッチとテンポを自由に変更できるカラオケシステム [2] も提案されている．このシステムでは，市販CD音源中のボーカル音を抑制することでカラオケの伴奏音が生成される．伴奏音のピッチとテンポの変更は，ユーザが手動で行う．

本稿では，市販CD音源から伴奏音を生成し，伴奏のテンポをユーザ歌唱に自動で追従させて再生するカラオケシステムを提案する．本稿で提案するシステムでは，まず市販CD音源から音源分離手法を用いて歌声と伴奏を分離する．次に，分離された歌声を用いて，伴奏をユーザ歌唱に自動で追従させて再生する．このようなカラオケシステムを用いることで，ユーザはCD音源のみを用意すれば，自分の好きな曲を歌うことができる．また，事前の設定なしに，歌唱のテンポを自由にアレンジして歌うことができる．

2. ユーザインターフェース

本システムのGUIを図1に示す．CD音源の音源分離の進捗状況確認や，伴奏音の再生・停止，音量調整が可能である．

2.1 音楽ファイルの選択

図1の橙枠で囲まれたボタンを押すと，カラオケで歌いたい楽曲の音源ファイルを指定できる．現時点では，

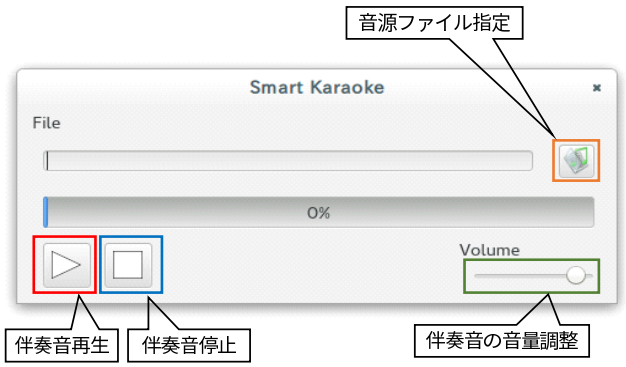

図 1: ユーザインターフェース

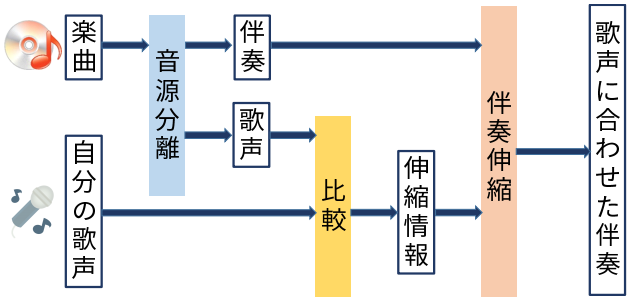

図 2: システムの全体像

モノラルWAVファイルのみをサポートしている．音源ファイルが指定されると，すぐに音源分離を開始する．分離の進捗状況は，ファイル選択部分の下部にあるプログレスバーに表示される．

2.2 音楽の再生・停止

図1の赤枠で囲まれたボタンを押すと，伴奏音の再生が開始する．マイクから取得された歌唱と分離された歌声信号を用いて伴奏の伸縮情報を計算し，それに従ってユーザ歌唱に追従するように伸縮された伴奏音が再生される．また，図1の青枠で囲まれたボタンを押すと，伴奏音の再生が停止する．再生を停止してからもう一度再生ボタンを押すと，伴奏音が最初から再生される．

2.3 伴奏音の音量調節

図1の緑枠で囲まれたスライダーにより，伴奏音の音量を調整する．ユーザは，このスライダーを用いていつでも伴奏音の音量を設定できる．

3. システム実装

本研究で実現するシステムの全体像を図2に示す．以下で各部分の動作について説明する．

3.1 音源分離

まず，ユーザが入力したCD音源を，歌声と伴奏に分離する必要がある．本システムでは，ユーザの待ち時間を低減させるために，処理時間を隠蔽することが重要である．そこで，本システムでは，坂東らによるロバスト

A Smart Karaoke System Based on Automatic Accompaniment for Singing Voices of Users : *Yusuke Wada, Eita Nakamura, Katsutoshi Itoyama, and Kazuyoshi Yoshii (Kyoto Univ.)*

非負値行列分解 [3] をオンライン化したものを用いて，オンラインでの音源分離を再生と独立したスレッドで行い，分離結果を保存することでこの課題を解決する．

3.2 特徴量抽出・アラインメント

音源分離によって得た伴奏をユーザ歌唱に追従させるため，伴奏をどのように伸縮すればよいかを計算する．本システムでは，ユーザ歌声と分離した歌声それぞれの音色情報を，動的時間伸縮法 (DTW) によって対応づけることで伸縮率を求める．音色情報として，人間の声道特性を表す特徴量であり，人間の母音知覚の特徴を考慮した量であるメル周波数ケプストラム係数 (MFCC) を用いる．

通常の DTW はバッチで動作するが，ユーザの待ち時間低減のため，DTW をオンラインで動作するように改変したものを用いる．以下に，オンライン DTW の処理の流れを記述する．

1. ユーザ歌唱と，音源のある一部分から分離された歌声それぞれの MFCC を計算する．入力の長さはどちらも 4096 サンプルである．それぞれの MFCC を $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}, \boldsymbol{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_n\}$ と書く．

2. $\boldsymbol{X}, \boldsymbol{Y}$ を入力として，下式に従って DTW 行列を計算する．

$$d_{0,0} = 0, d_{s,0} = d_{0,t} = \infty \quad (s = 1, 2, \cdots, m; t = 1, 2, \cdots, n) \tag{1}$$

$$d_{i,j} = ||\boldsymbol{x}_i - \boldsymbol{y}_j|| + \min \begin{cases} d_{i,j-1} \\ d_{i-1,j} \\ d_{i-1,j-1} \end{cases} \tag{2}$$

ただし，$||\boldsymbol{x}_i - \boldsymbol{y}_j||$ は，$\boldsymbol{x}_i$ と $\boldsymbol{y}_j$ の距離を表し，本研究では平均二乗誤差を用いる．

3. 通常の DTW と同様に DTW 行列をバックトラックし，対応する点対の系列 $L = \{\boldsymbol{o}_0, \boldsymbol{o}_1, \cdots, \boldsymbol{o}_l\}, \boldsymbol{o}_i = (o_{xi}, o_{yi}) (i = 0, 1, \cdots, l; 0 \leq o_{xi} \leq o_{xi+1} \leq m; 0 \leq o_{yi} \leq o_{yi+1} \leq n)$ を計算する．L をワーピングパスと呼ぶ．ここで，$L_X = \{o_{x0}, o_{x1}, \cdots, o_{xl}\}, L_Y = \{o_{y0}, o_{y1}, \cdots, o_{yl}\}$ とする．

4. ワーピングパス L から伸縮率の系列 $R = r_0, r_1, \cdots, r_n$ を計算する．r_i は，分離された伴奏の i 番目のフレームを何倍に伸縮するかを表し，下式に従って計算される．

$$r_i = \frac{L_Y \text{中の} i \text{の個数}}{L_X \text{中の} i \text{の個数}} \tag{3}$$

3.3 伴奏伸縮

2.3 節で求めた伸縮率に従って伴奏信号を，フェーズボコーダ [4] を用いて伸縮する．フェーズボコーダを用いることで，音声の音高を保ちつつ長さを伸縮することができる．その際，位相は隣接するフレーム間の整合性がとれるよう線形補間によって与える．

4. 評価

オンライン DTW の性能を評価するため，バッチ DTW によって得られるワーピングパスとの比較を行う．評価には，井上陽水「少年時代」のサビ部分の冒頭 2 小節分 (4.8 秒) と，その部分を歌った音源を用いた．また，

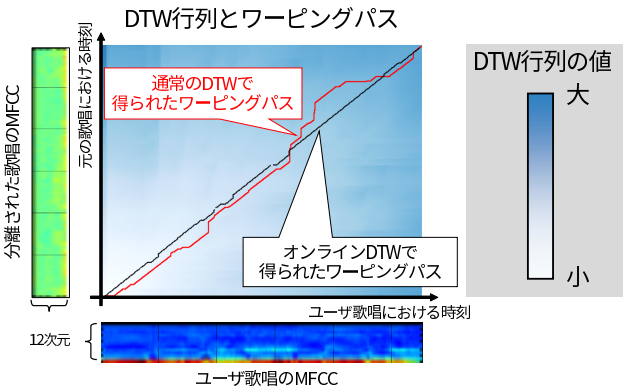

図 3: ワーピングパスの比較と DTW 行列

MFCC の次元は 12 次元とした．DTW 行列を濃淡で表したものと，通常の DTW とオンライン DTW のそれぞれによって得られるワーピングパスを，図 3 に示す．赤線が通常の DTW によって得られるワーピングパスを，黒線がオンライン DTW によって得られるワーピングパスを表す．オンライン DTW のワーピングパスは，通常の DTW に比べてあまり逸脱しておらず，妥当な伸縮率が得られると考えられる．

5. おわりに

本稿では，入力された市販 CD 音源から伴奏音を生成し，伴奏のテンポをユーザ歌唱に自動で追従させて再生するカラオケシステムを提案した．実験の結果，オンライン DTW でユーザ歌唱と元の楽曲中の歌唱が概ね正しくアラインメントされていることが確認できた．一方，オンライン DTW の定量評価やシステムに対する評価が不十分なため，ワーピングパスの一致率によるオンライン DTW の性能評価と，被験者実験によるシステムの評価を行う予定である．

インタフェース面では，ユーザに対するフィードバックが存在しないので，ユーザ歌唱と分離された歌声それぞれの F0 軌跡を表示させる機能を追加したい．また，ユーザの歌唱を補助するため，演奏位置に同期して歌詞が表示されるような機能の拡張を考えている．

一方，アラインメントにも精度向上の余地があると考えられる．例えば，ピッチに関する情報を同時に推定・利用することで精度向上が期待できる．また，DTW とは別のパーティクルフィルタを用いてアラインメントを行う手法 [5] も提案されている．これらの手法を現在のシステムで利用できるかどうかを検討する予定である．

謝辞 本研究の一部は，JSPS 科研費 24220006, 26700020, 26280089, 16H01744, 16J05486, 15K16054， JST CREST，JST ACCEL の支援を受けた．

参考文献

[1] Wataru Inoue, Shunji Hashimoto, and Sadamu Ohteru. Adaptive karaoke system: Human singing accompaniment based on speech recognition. *ICMA*, 1994.

[2] Hideyuki Tachibana, Yu Mizuno, Nobutaka Ono, and Shigeki Sagayama. A real-time audio-to-audio karaoke generation system for monaural recordings based on singing voice suppression and key conversion techniques. *IPSJ*, 24(3):470–482, 2016.

[3] 坂東 宜昭ら. 変分ベイズ多チャネル RNMF に基づく柔軟索状レスキューロボットのための音声強調. 日本ロボット学会, 2016.

[4] J.L. Flanagan and R.M. Golden. Phase vocoder. *Bell System Technical Journal*, 45:1493–1509, 1966.

[5] Nicola Montecchio and Arshia Cont. A unified approach to real time audio-to-score and audio-to-audio alignment using sequential montecarlo inference techniques. *ICASSP*, 2011.