Vol.2016-MUS-110 No.6
2016/2/29

対話音声の韻律情報を用いたメロディ生成の試み

大野拳汰† 山本樹† 佐々並賢志† 梅村祥之†

概要：人と人の対話では感情を伝えるのに韻律情報が使われる．感情表現が豊かな映画作品での対話音声を基にして，音声波形からピッチを抽出し，その周波数と継続時間から音符の音高と長さを求めてメロディラインを生成するシステムを開発した．洋画の対話音声から，8 小節程度の長さを生成できる数秒間の区間を 1,000 箇所切り出し，そのシステムを用いて 1,000 曲のメロディを生成した．この中に，良いと感じられる箇所を含む曲がどれだけ含まれるかを主観評価したところ，5%程度含まれていた．

Melody generation based on intonation of spoken dialogs

KENTA OHNO† TATSUKI YAMAMOTO†
SATOSHI SASANAMI† YOSHIYUKI UMEMURA†

Abstract：We use intonation in dialogs to convey our emotions. We developed a melody generation system that generate melodies made of notes which pitches and durations are determined by speech waveforms. We extracted 1,000 portions of waveforms which durations are enough to generate melodies with about 8 measures and generated 1,000 melodies. We subjectively evaluated whether these melodies are good. As the result the proportion of good melodies is about 5%.

1.はじめに

これまで，自動作曲技術の研究開発の中で，自然界に存在する信号や，自然界の形状を模擬した生成モデルによる信号を基にして曲を生成する方法，詞や声のような人の発する情報を基に曲を生成する方法などが研究開発されている．自然界に存在する信号を用いた生成法としては，例えば，カメラで撮影した映像における物体の色情報と距離の情報を基に曲を生成する方法が提案されている[1]．自然界の形状を模擬した生成モデルによって生成した信号を基にした生成法としては，フラクタル図形によって生成された樹木の形状や雪の結晶の形状を基に曲を生成する方法などがある[2].人の発する情報を基にする方法としては，生体信号を用いるものとして，脳波から wavelet 解析によって得られた特徴量を基に曲を生成する方法がある[3]．以上は，元々楽曲とは関係のない信号から曲を生成する手法である．それに対し，楽曲に関連の深い情報に基づくものとして，作詞を基に，言語解析によって詞の韻律情報を抽出して曲を生成する自動作曲システム Orpheus が開発されている[4].また，鼻歌の音響信号を分析して曲を生成するソフトウェア Songsmith が開発されている[5].人と人とのコミュニケーションにおける音楽の役割を扱った書籍「音楽的コミュニケーション」[6]において，「音楽は、コミュニケーションの基本的チャネルの一つである。それは人々が共有することの出来る感情、意図、意味の手段を与える。」(同書 p1 より引用)と述べられている．この考え方から，音声による人と人とのコミュニケーションである音声対話を素材にして音楽に変換することを考える．対話音声の素材として，映画の中で主人公が相手役と感情をあらわにして言葉をやりとりする対話音声が望ましいと考え，それを素材に採用する．音声から抽出して利用する情報は韻律情報であり，発話内容に関する情報ではないことと，日本人にとって，英語の音声の方が日本語の音声よりもリズミカルに感じる人が多いと考え，今回，英語の映画作品を対象とする．以上の考え方に基づいて，音声信号から楽曲に変換するソフトウェアを開発する．映画の対話音声は大量に存在するため，そのソフトウェアを用いれば，短時間のうちに大量の楽曲を生成することが出来る．その中には，楽曲として良い曲から悪い曲まで様々な感性品質の曲が含まれるであろう．我々の研究室では，以前より，楽曲の客観評価法の研究を継続して行っている[7-8]．そこで，開発してきた客観評価法を利用して，生成曲の機械選別ができないかどうかについても，フィジビリティスタディを行う．

2. 楽曲生成法

2.1 基本処理

対話音声の wav ファイルから基本周波数(以下，F0 と表現する)を抽出するために，ソフトウェア The Snack Sound Toolkit[6] の F0 抽出機能を用いる．このソフトウェアの場合，分析窓の移動量が標準で 10msec となっている．10msec 単位で音符を生成したのでは音長が非常に短いため，いくつかの F0 データをまとめて音符として違和感の

† 広島工業大学 情報学部 情報工学科
Hiroshima Institute of Technology

ない長さにする．そこで，まとめる単位をおよそ単語単位とすべく，F0 の時系列波形から周波数の似通っている部分をまとめる方法を試みる．図 1 は「What are you doing．」という音声対話データを，１英単語毎を１つの音と見なした完成予想図である．

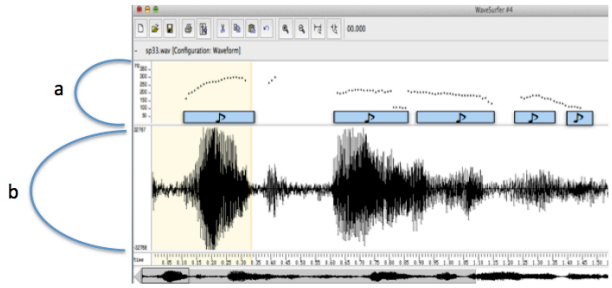

図１　１英単語を１つの音符に対応された完成予想図
Fig. 1 Conceptual figure assigning one English word to one note.

図 1 の(a)の点の集合が F0 データ，(b)が音声波形を示している．求めた F0 データの平均値を，その平均値に近いピアノ音の全音階における音高で音を決定する．無音区間は休符とする．図の中には，無音区間に挟まれた１つの F0 データの集合に，複数の単語が含まれる場合もある．そのため，平均値を求める前に，無音区間に挟まれた F0 データ数１０個以上の１つの集合を半分に分割する．前半の平均と後半の平均を求め，小さい値を大きい値で割って比を求める．比が 1.2 より大きければ図２のように，前半の F0 データと後半の F0 データは違う単語とし，F0 データの集合が１０個以上なら同じ処理を繰り返しする．比が 1.2 より小さければ，図３のように，F0 データの集合を１単語とし，次の F0 データの集合で同じ処理をし，すべての F0 データの集合の分割処理をする．

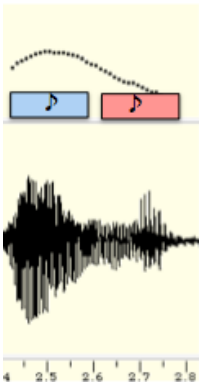

図２　比＞１.２
Fig. 2 ratio > 1.2

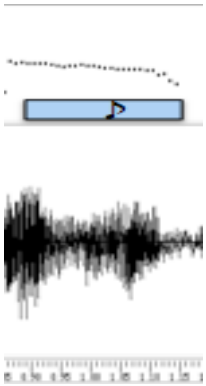

図３　比＜１.２
Fig.3 ratio < 1.2

2.2 リズム制御およびドラム伴奏付け

2.1 で生成された音符系列は小節単位にまとまってはいない．そのため，音符の継続時間中に小節線が現れたり，小節の始まりが休符になることが無秩序に発生する．通常，強起の曲であれば小節の先頭に長い音長の音符が生じやすいといった秩序を持ち，それによって曲を聴く人はリズムを知覚する．そこで，音符の位置と音長を調整することによって，小節に同期して音符が現れるようにする．以下，図 4 を参照しながら具体的なアルゴリズムを 3 ステップで説明する．以下，4/4 拍子，テンポ 120 に限定して説明する．すると，4 分音符の音長は 120 となる(要確認)．

<step 1>
各小節線の位置を基準に，-120 から+360 の範囲に音符(休符は除く)の中心（オンセットとオフセットの中間点）が存在すれば，その中で最も音長の長い音符のオンセット位置を小節線の位置に移動する．その区間に音符の中心が存在しない場合，休符が存在すれば，休符の中で最も音長の長い休符のオンセット位置を小節線の位置に移動する．

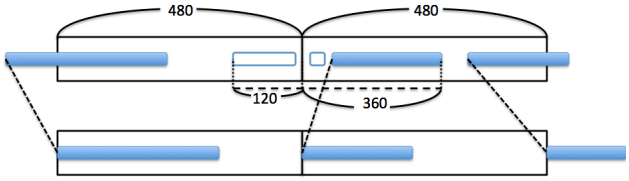

図４－１　リズム制御方法に関する説明図(step1)
Fig. 4-1 Illustration of Rhythm control(step 1).

<step 2>
step 1 で選定された複数の音符のオンセット位置を，該当する小節線の位置に移動させるのに合わせて，各小節単位で，元の楽譜と相似形になるように，オンセットとオフセット位置を比例配分させて，全音符及び休符の位置を設定する．

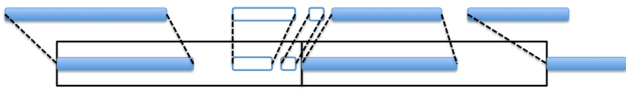

図４－２　リズム制御方法に関する説明図(step2)
Fig. 4-2 Illustration of Rhythm control(step 2).

<step 3>
音符と休符の音長を予め設定された長さに量子化する．設定は，8 分音符，4 分音符，付点 4 分音符，2 分音符，付点 2 分音符，全音符の 6 種類とする．量子化に際して音符および休符が小節線をまたがないという条件を課す．

図４－３　リズム制御方法に関する説明図(step3)
Fig. 4-3 Illustration of Rhythm control(step 3).

以上の処理により小節区間に同期して音符が現れるようになり，拍を感じ取りやすくなる．さらに，リズム感を高めるため，リズム楽器を伴奏に加える．
リズム楽器として，MIDI ノート番号 47 の Low-Mid Tom, MIDI ノート番号 43 の High Floor Tom, MIDI ノート番号 38 の Acoustic Snare の 3 楽器を用いる．演奏には，我々の研究室で開発された演奏ソフトを使用する．そのソフトが使用する音源として，Apple 社製コンピュータ Macintosh 付属ソフト QuickTime Player 7 によって MIDI ファイルを再生して出力される波形ファイルを用いる．メロディにリズム伴奏をつけて演奏する．そのリズムパターンを図 5 に示す．

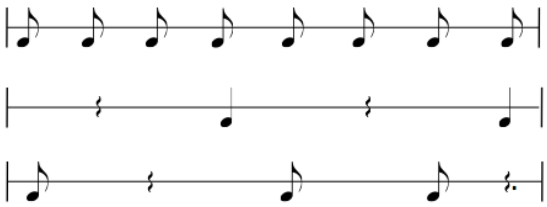

図 5　リズム伴奏パターン
Fig. 5 Rhythm accompaniment patterns.

3. メロディの生成及び主観評価結果

3.1 素材の選定

素材は洋画の中のシーン(約 5 秒)を選定する．メロディを生成する上で日本語より英語の方が対話のテンポが良く，良いメロディが生成できると考えたからである．用いた洋画のタイトルは，「ローマの休日」「パリの恋人」「風と共に去りぬ」「12 人の怒れる男たち」「紳士は金髪がお好き」「マイフェアレディ」の 6 作品である．この中から最終的に 1,000 個の素材を選定した．

3.2 楽曲生成

選定された 1,000 個の対話音声から上記の方法で 1,000 曲の楽曲を生成した. 映画タイトル別に曲数をまとめると，「ローマの休日」17 曲，「パリの恋人」63 曲，「風と共に去りぬ」271 曲，「12 人の怒れる男たち」419 曲，「紳士は金髪がお好き」30 曲，「マイフェアレディ」200 曲である．1,000 曲の生成は全て自動処理によって行った．

生成曲 1,000 曲の概略を数量的に示すと，1 曲あたり小節数の平均が 6.3 小節，音符数と休符数の合計の平均が 25.9 音符，音符数の平均(休符は除く)が 15.5 音符であった．生成した楽曲の例を図 6 に示す．

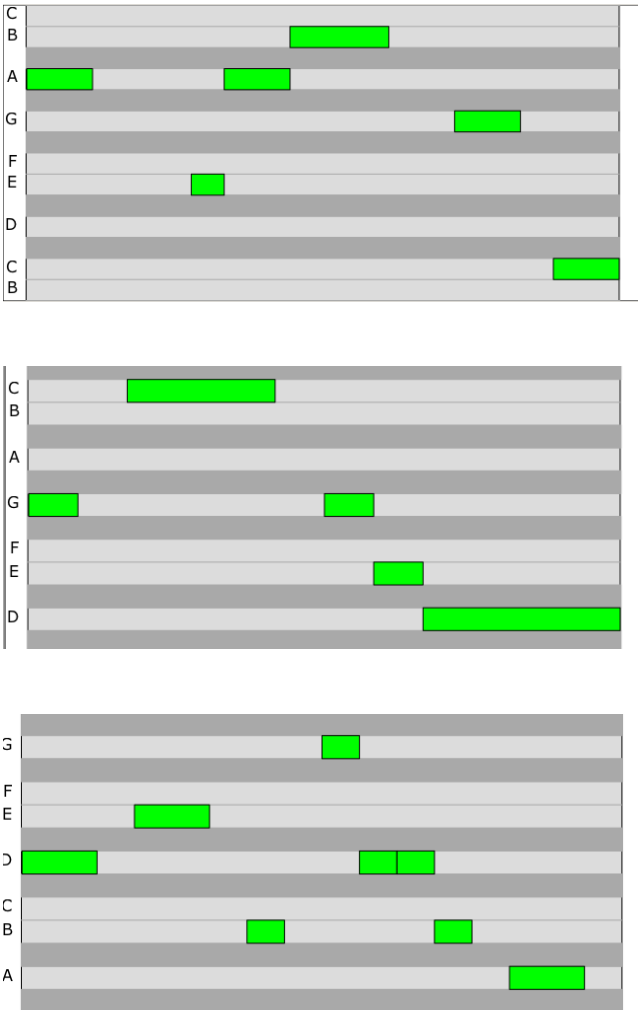

図 6　3 曲の楽曲生成例
Fig.6 Example of 3 generated pieces.

3.3 主観評価

生成した楽曲 1,000 曲に対して 2 名で主観評価を行う．評価の方法としては，予備検討の中で比較的良いと感じられた曲(以下，比較対象曲と呼ぶ)を選定し，主観判断の対象とした．と比較をし，同等または、それ以上と判断できる曲を探すというものである．

また，感性は人それぞれ違うので，評価者 2 名の評価の一致度を調べるために 1,000 曲のうち 100 曲を別々に評価し，評価結果を照らし合わせたところ，87 曲に対する評価が一致した．このことから，2 名の評価の一致度が高いことが分かった．以降，残りの 900 曲に対する評価を分担して評価した．

1,000 曲に対する最終的な評価の結果は，比較対象曲と同等，またはそれ以上と判断されたのが 67 曲であった．

4. 曲の良し悪しに関する客観評価法の開発

著者らの研究室で楽曲の心地よさに関する客観評価法の研究を行っている[3 文献]. 音高系列が出現しやすいパター

ン(聴き慣れたパターン)かどうかの特徴と，音楽理論に基づく調性等の特徴量を用いている．客観評価法の技術開発において，これまで，全て同じ音長の音符 8 音符で構成される 1 フレーズという短い単位を評価対象としてきた．ところで，曲の心地よさに関する主観評価は，個人個人の感じ方の違いにより，評価者によるばらつきが大きい．しかし，評価者 1 名毎の評価値を教師データとして機械学習，機械判定を行うと，評価者によってはかなり高い正解率で機械判定できることを示してきた．このようにして開発してきた客観評価法を，対話音声に基づく曲生成で得られた生成曲の中から良い曲を機械選別する技術として利用できないか検討する．

客観評価法開発において，判定タスクを，心地よい，心地よくないの 2 カテゴリで判定するタスクとしていた．そのため，客観評価法の判定出力の心地よい，心地よくないのカテゴリであった．今回の応用においては，基礎検討段階のため，カテゴリの前段階である連続値を用い，詳細を探る．客観評価法において線形判別分析を用いているため，判定結果を得る前のスコアをこの連続値に採用する．極性は，数値が小さいほど，心地よいことを示している．

先の客観評価法の研究では，個人毎の評価値を教師データとして機械学習して判別関数を得て，それを用いて，同じ評価者のテストデータを判定した．本応用においては，先の研究で収集した 5 名(今回の評価者とは異なる)の主観評価値を学習データとして構築した判別関数を使用する．

客観評価法の開発段階で 8 音符からなる 1 フレーズを対象にしたのに対し，本応用では，音符数が平均 15.5 音符で，フレーズを 1 個ないし 2 個程度含む長さである．そこで，図 6 に示すように，1 曲を 2 小節単位の複数の区間に分けて，各々の区間に客観評価法を適用する．

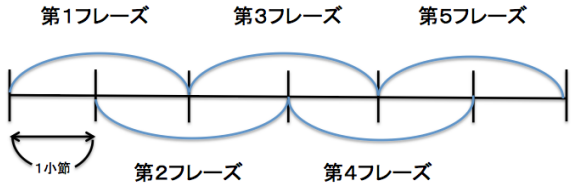

図 7　2 小節単位での客観評価指標算出
Fig. 7 Computing objective evaluation indexes per 2 measures.

なお，客観評価法は，本研究室ないで継続的に改良を重ねているため，先の文献[7]で述べた特徴量に対し，多少の改良が加えられている．

5. 客観評価による生成曲の自動選別実験及び結果

前述で述べた客観評価法を用いてメロディ生成法で生成された 1000 曲を客観評価値で比較して良い曲と悪い曲に選別できないかについて，フィジビリティスタディを行う．

生成された 1000 曲のうち，主観評価で選定されたフレーズについて，客観評価値のヒストグラムを描き，それに重ねて，選定されなかったフレーズについて，同様のヒストグラムを描いた（図 8）．その結果，選定されたフレーズは，おおよそ客観評価値が良いフレーズであることを示している．

両分布の平均値の差を t 検定する．R 言語付属の関数 t.test を用い，welch の方法により有意水準 95%で片側検定した．結果 $p = 4.6\times10^{-6} < 0.05$ となり有意差が認められた．

今回，6 タイトルの映画を用いたので，その各々について，同様のヒストグラムを描く．ただし，3 タイトルの中に選定されたフレーズがなかったため除外し残るタイトルについて描画する．結果を図 9，10，11 に示す．なお，t 検定の結果は各々，$p = 2.3\times10^{-5} < 0.05$ で有意差有り，$p = 1.3\times10^{-6} < 0.05$ で有意差有り，$p = 0.3\times10^{-2} < 0.05$ で有意差有りとなった．以上 6 タイトル中，図 9 の「マイフェアレディ」が最も分布が分離していた．赤のグラフが主観評価で良い曲と判断したもので，青いグラフが悪い曲である．また赤のグラフと青のグラフは 2 小節ごとのフレーズで区切られており、すべてのパターンを割り出すためにフレーズの始まりを 1 小節ごとにシフトしている．

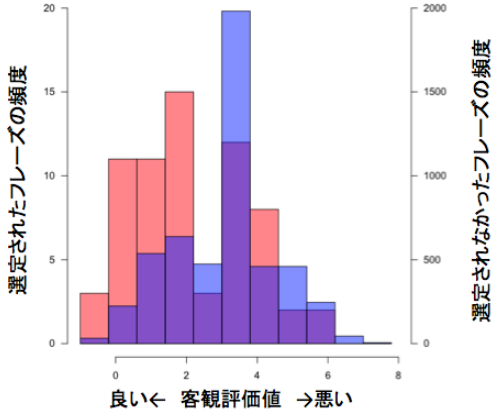

図 8　良いと選定されたフレーズおよびそれ以外のフレーズの客観評価値の分布(全体タイトル)
Fig. 8 Histogram of objective evaluation values of comfortable phrases and others (all movie titles).

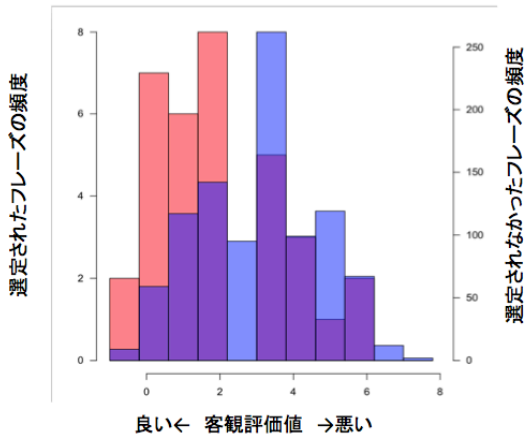

図９　良いと選定されたフレーズおよびそれ以外のフレーズの客観評価値の分布(「マイフェアレディ」のみ)
Fig. 9 Histogram of objective evaluation values of comfortable phrases and others (“My Fair Lady”).

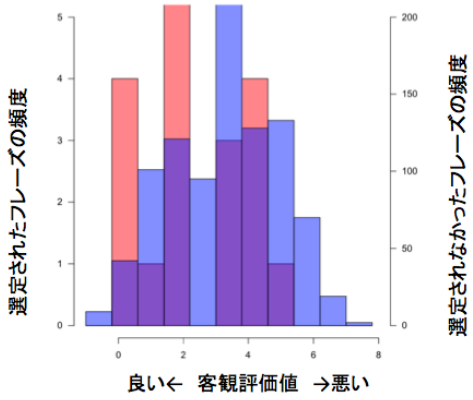

図１０　良いと選定されたフレーズおよびそれ以外のフレーズの客観評価値の分布(「風と共に去りぬ」のみ)
Fig. 10 Histogram of objective evaluation values of comfortable phrases and others (“Gone With the Wind”).

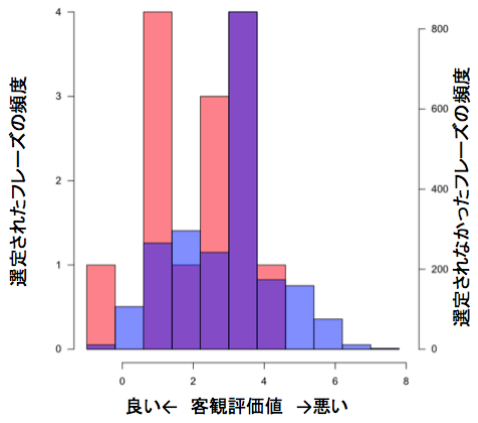

図１１　良いと選定されたフレーズおよびそれ以外のフレーズの客観評価値の分布(「12 人の怒れる男たち」のみ)
Fig. 11 Histogram of objective evaluation values of comfortable phrases and others (“12 Angry Men”).

6. 考察

洋画の対話音声から, 8 小節程度の長さを生成できる数秒間の区間を 1,000 箇所切り出し，そのシステムを用いて 1,000 曲のメロディを生成した．この中に，良いと感じられる箇所を含む曲がどれだけ含まれるかを音楽経験のない実験参加者 2 名で分担して主観評価したところ 5%程度含まれていた．

内省報告より,「マイフェアレディ」が一番良い結果になった理由を考察すると，6 タイトルの映画の中で会話の抑揚がはっきりしており，なおかつリズミカルな口調で会話していたからだと考えられる．

良いと感じられる箇所について，どの程度良いのかを詳細に評価していない．しかし，音楽未経験者が心地よい箇所として選定したうちの 13 箇所について作曲経験のある実験参加者 1 名に作曲用の素材として使いたいか評価して貰った結果，半数程度が使いたいと評価された．このことから，本システムで生成された曲の中に，人間が作曲する際のモチーフを想起する際の手助けとなるものが含まれていると考えられる．

今回生成した 1,000 曲中，良い箇所が 50 箇所あった．本技術を作曲支援システムに応用する場合, このままでは, 生成曲中の 5%程度しか良い箇所が含まれていないということは，適合率が低いために効率のよい作曲支援とはならない．

今回得られた良い曲の，映画タイトルによる分布を調べた結果，映画タイトルによる偏りがかなり大きいことが判明した．今後，さらに分析を重ねることにより，良い曲が生成されやすい映画の特徴が解明されれば，映画タイトル選定の際のスクリーニング機能として実装することにより, 生成曲中の良い曲の割合を向上させることができるかもし

れない.

生成曲中の良い曲の割合を向上させる方法の検討として,客観評価指標による自動選別の可能性を検討した.すなわち,主観評価により良い曲とされた曲断片の客観評価値とそれ以外の曲断片の客観評価値に関する平均値の差に有意差が認められた.しかし,両分布の分離の程度から,現状の客観評価紙票を使えば,良い曲とそれ以外を機械選別できるレベルには至っていない.また,現状では,客観評価指標を使って機械選別すると,その中の曲はほとんど良い曲であるといったレベルにも至っていない.

7. まとめ

感情表現が豊かな映画作品での対話音声などの素材を基に,音声波形からピッチを抽出し,その周波数と継続時間から音符の高さと長さを定めて楽曲を生成するシステムを開発した.

洋画の対話音声から,8 小節程度の長さを生成できる数秒間の区間を 1,000 箇所切り出し,そのシステムを用いて 1,000 曲のメロディを生成した.この中に,良いと感じられる箇所を含む曲がどれだけ含まれるかを主観評価したところ 5%程度含まれていた.# 音楽経験のない実験参加者 2 名で分担して主観評価したところ 5%程度含まれていた.また,良いと感じられたうちの 13 箇所を作曲経験のある実験参加者 1 名に評価してもらったところ,作曲用の素材として使いたいと感じるものが半数程度含まれていた.以上から,本技術は良い曲を生成する技術として有用であると考える.

今後の検討課題としては,生成曲の良さを詳細に主観評価すること,対話音声素材の選定指針やスクリーニング方法を検討すること,生成曲の自動選定に向けた客観評価技術の改良などが挙げられる.

謝辞

本研究は,主観評価実験等,広島工業大学の多くの方々にご協力頂いた.関係各位に深く感謝する.

参考文献

[1] 高橋弦太, 笹岡 久行: 実空間の情報を用いた背景音楽の自動生成,情報処理学会第 74 回全国大会, pp.2-377 - 2-378(2012).
[2] G. Nierhaus: 6 Chaos and Self-Similarity In Algorithmic Composition, Springer, pp.131-156(2009).
[3] W. Dan, L. Chaoyi, Y. Yu, Z. Changzheng and Y. Dezhong: Music Composition from the Brain Signal: Representing the Mental State by Music, Computational Intelligence and Neuroscience, Vol. 2010, Article ID 267671(2010).
[4] 深山覚, 中妻啓, 酒向慎司, 西本卓也, 小野順貴, 嵯峨山茂樹:音楽要素の分解再構成に基づく日本語歌詞からの旋律自動作曲, 情報処理学会論文誌, Vol.54, No.5, pp.1709-1720(2013).
[5]http://research.microsoft.com/en-us/um/redmond/projects/songsmith/
[6] D. Miell(ed), R. Macdonald(ed) and D. J. Hargreaves(ed):Musical Communication, Oxford University Press(2005), s 星野悦子(監訳):音楽的コミュニケーション, 誠信書房(2012).
[7] http://www.speech.kth.se/snack/index.html