ビートとコードをリアルタイムで認識しながら音楽に合わせて歌って踊るロボット

津牧美葉子[1]，大和勝宣[2]，和佐圭悟[2]，池田賢矢[3]
坂東宜昭[4]，大喜多美里[4]，糸山克寿[4]，吉井和佳[4*]

[1] 同志社女子高等学校
[2] 近畿大学附属和歌山高等学校
[3] 灘高等学校
[4] 京都大学大学院情報学研究科
* yoshii@kuis.kyoto-u.ac.jp

概要

本稿では，2015 年度に開講された専修コース「ロボット聴覚と音楽情報処理」の研究成果について報告する．本コースでは，高校生四名と大学院情報学研究科知能情報学専攻音声メディア分野の大学院生数名とが協力しながら，音楽を聴きながらビート時刻とコードをリアルタイムに認識し，音楽に合わせて即興で歌いながらダンスをするロボットの開発に取り組んだ．我々が開発したロボットは，主に音楽解析部（大和・池田が担当）・ダンス制御部（津牧が担当）・歌唱制御部（和佐が担当）の三つから構成されている．これらのモジュールは独立性が高くなるように設計することで，高校生は自らの担当部分に専念することができ，最後に統合実験まで計画通り行うことができた．研究成果は，音楽情報処理のトップカンファレンスである ISMIR 2015 の Late Breaking Demo セッションにて発表した．聴衆から高い評価を受け，多くの著名な研究者と有意義な議論・交流を行うことができた．

図 1. 専修コースの構成メンバー：後列左から大和・池田，前列左から和佐・津牧，その他は大学院生 TA.

歌って踊れるロボット

将来的に人間との共存が期待されるエンターテインメントロボットの中でも，ロボットダンサーやロボットシンガーなどの音楽ロボットはメディア解析技術の最も魅力的な応用先のひとつである．我々が開発したロボットは，主に音楽解析部・ダンス制御部・歌唱制御部の三つから構成されている．

- 音楽解析部では，音楽音響信号をマイクから取り込みながら，リアルタイムでビートとコードの認識を行う．具体的には，汎用的なロボット聴覚システム HARK が提供するビジュアルプログラミング環境において，ビート認識とコード認識のモジュールを実装することで実現される．
- ダンス制御部では，認識したコードの種類に応じて，あらかじめ定めておいた 24 種類の動作のうちのいずれかの動作を生成する．
- 歌唱制御部では，認識したコードのルート音と同じ音高の歌唱制御信号を生成し，歌声合成用の MIDI デバイス（Yamaha eVY1）に送信することで，歌唱信号をロボットのスピーカから再生する．

ビートやコードの情報は ROS と呼ばれるミドルウェア上で非同期に通信させることとし，各部の独立性を高めながら見通しの良い開発を行った．

開発の経緯

本コースは，4 月から 6 月の土曜日六回分（午後 2 時から 6 時）で構成されている．まず，第一回目に，高校生らとディスカッションを行い，研究内容を「音楽に合わせて歌って踊れるロボット」に決定し，班分けを行った．その後，プログラミングに必要となる Python 言語の演習を行った．第二回目に，各部を実装するうえで専門的な知識の習得を行った．コード認識班は，音楽音響信号処理や統計的機械学習の基礎について学んだ．第三回目では，大学院生の協力を得ながら各部の実装を進めた．HARK 上でのモジュールの実装方法や eVY1 の制御方法について学んだ．第四回目では，ダミーのビート・コード情報に対して，ダンス制御部や歌唱制御部を実際に動作させながら改良を行った．音楽解析部に関しては，いくつかのテスト音をロボットに聴かせてみて，ビートやコードをリアルタイムで正しく認識していることを確認した．第五回目には，三つの機能を接続してテストを行い，さらに各部に改良を施した．最終回では，実際のポピュラー音楽を用いた実験を行い，コード変化に対する反応遅れを小さくするため音楽解析部の高速化にも取り組んだ（図 1）．

国際会議での発表

研究成果を英語で 2 ページの論文にまとめ，音楽情報処理のトップカンファレンス（総参加者は 300 名以上）である

図 2. デモ発表会場：左から糸山（助教）・和佐・大和・津牧・中村（ポスドク）・吉井（講師）.

図 3. 英語で質疑に回答.

International Society for Music Information Retrieval Conference 2015（ISMIR 2015）に投稿した．ただし，メイントラックではなく，Late Breaking Demo（LBD）セッションと呼ばれる，最新の話題を気軽に発表できるセッションである（図 2）．会議の最終日に，メイントラックがすべて終了した午後に行われる本セッションでは，フランクな雰囲気のもと，すでに打ち解けた参加者どうしで活発な議論が展開される．初めての海外あるいは欧州という高校生も多かったが，多数訪れた外国人研究者に対して，どうにか英語で対応することができていた（図 3）．日本から人型ロボット（アルデバラン社製 Nao）を持参し，歌って踊るライブデモを披露したこともあり，注目度は圧倒的であった（図 4）．

ISMIR では，女性の学生や研究者を奨励する活動にも力を入れており，Women in MIR（WiMIR）というセッションが存在する．その中で，日本から女子高校生が発表のために参加していることが話題となり，音楽情報処理研究の裾野の拡がりが大変喜ばしく思われるという一幕があった．ISMIR はシングルセッション形式を採用しており，常に全ての研究者が一堂に会しているので，図らずも我々の研究成果の良いアピールになった．実際，バンケットや LBD セッションを通じて多くの外国人研究者が熱心に話を聞いてくれるきっかけとなり，高校生にとっても貴重な経験となった．

図 4. 外国人研究者に好評．中央は今年のプログラム委員長である Meinard Müller.

コース内容設計の工夫

本専修コースでは，時間の制約が非常に厳しい中で，高校生らが各自の強みやバックグラウンドを生かしてタスク分担したことが成功につながった．和佐一名を除いて高校生はプログラミング初心者であったので，徐々にプログラミングに慣れ親しんでいくよう配慮した．本専修コースは，前年度の 12 月から 2 月まで開講された基盤後期コース「ロボット聴覚」の上級版として位置付けられており，津牧・大和・和佐の三名は継続しての参加となっている．そのため，ロボット聴覚やプログラミングに関してある程度理解が進んでいる点で有利であった．一方，専修コースからの参加となった池田は，プログラミングの習得を進めつつ，本人の興味を考慮して，主に数理的なアルゴリズムの考案を担当することにした．プログラミング経験豊富な和佐（電子デバイスの制御経験もある）は，歌声合成デバイスである eVY1 の制御を担当することにした．

また，教員と大学院生らによる事前の綿密な計画・準備が重要であることは言うまでもない．基盤後期コースでは，大学の「高等教育」を体験することを主目的としていたのに対し，専修コースでは，大学での「学術研究」を体験することを主目的として，学術的に新規性・有効性のある内容となるよう配慮した．ただし，研究というものの性格上，成功するかどうかを完全に予測できるわけではなく，その分野に精通した教員の経験によるところが大きい．そこで，ミニマルサクセスを定め，成功のレベルを段階的に設定することにより，All or Nothing という博打を避けられるような工程を考案した．若くて感度の高い高校生らにとっては成功体験が重要であり，今後の人生において困難に挑戦していく態度を醸成するうえで，リスクと成功とのトレードオフは慎重に検討すべきである．

海外発表を行うという選択肢も，少なくとも情報系の場合，十分に検討に値する．確かに，限られた実習時間では，国際会議にフルペーパーを投稿することは極めて困難であるが，併催されるワークショップなどでは比較的発表しやすいという事情がある．技術の進歩が速く，トップカンファレンスが重視される情報系においては，著名な研究者の前で発表できる意義は大きく，高校生の英語に対する意識向上にもつながる．

本報告は、ELCAS 専修コース「ロボット聴覚と音楽情報処理」分野の研究成果であり、International Society for Music Information Retrieval Conference 2015（ISMIR2015、http://ismir2015.uma.es/）で発表されたものである。なお、99–100 ページに ISMIR2015 におけるプロシーディングスを転載した。

A HUMANOID ROBOT THAT CAN SING AND DANCE TO MUSIC BY RECOGNIZING BEATS AND CHORDS IN REAL TIME

Miyoko Tsumaki[1*] **Masanobu Yamato**[2*] **Keigo Wasa**[2*] **Kenya Ikeda**[3*]
Yoshiaki Bando[4] **Misato Ohkita**[4] **Katsutoshi Itoyama**[4] **Kazuyoshi Yoshii**[4]
[1]Doshisha Girls' Senior High School, Japan [3]Nada Senior High School, Japan
[2]Kinki University Wakayama Senior High School, Japan [4]Kyoto University, Japan
{yoshiaki,ohkita,itoyama,yoshii}@kuis.kyoto-u.ac.jp

ABSTRACT

This paper presents a humanoid robot capable of singing and dancing to a song in an improvisational manner while recognizing the beats and chords of the song in real time. Among various kinds of entertainment robots that are expected to live with humans in the future, music robots such as robot dancers and singers are considered as one of the most attractive applications of music analysis techniques. Our robot mainly consists of listening, dancing, and singing functions. The listening function captures music audio signals and recognizes the beats and chords in real time. The dancing function switches dancing movements according to the types and root notes of the estimated chords. The singing function, on the other hand, generates singing voices whose pitches change according to the root notes of the chords. The information on beats and chords are exchanged between the three functions. The preliminary experiment showed the great potential of the proposed dancing robot. We plan to improve the response of dancing and singing functions by predicting next chords.

1. INTRODUCTION

Development of entertainment robots that can interact with humans through music is an important research direction in the field of music information retrieval (MIR). Since various kinds of robots are expected to get into our daily lives in the future, not only task-oriented robots but also entertainment robots that people feel familiarity with have been developed, *e.g.*, a violinist robot [3] and a flutist robot that can play the flute in synchronization with a melody played by a human [5]. Since dancing is a universal form of expression seen in many cultures, in this paper we focus on music robots that can dance interactively with humans. Although many researchers have tackled music signal analysis for content-based music selection (retrieval and recommendation), we aim to enhance a user's experience of listening to selected musical pieces by leveraging the analyzed contents for music robots [4].

A robot that can dance synchronously with music needs to adaptively control its movements while recognizing the content of music. Several robot dancers have already been developed. Murata *et al.* [4], for example, enabled a bipedal humanoid robot to step and sing in synchronization with musical beats, Kosuge *et al.* [2] devised a dance partner robot that can predict the next step intended by a human dancer, and Kaneko *et al.* [1] developed a humanoid robot that can generate natural dancing movements by using a complicated human-like dynamical system. In contrast to these robots, our robot is capable of singing and dancing to *any* musical pieces in an improvisational manner by recognizing the beats and chords in real time instead of using musical score information.

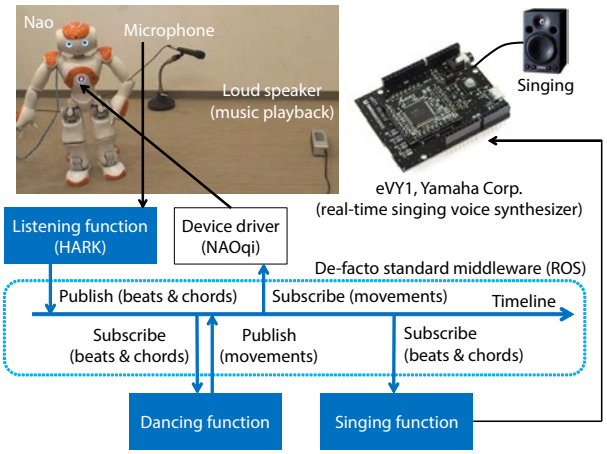

Figure 1. System architecture of a singing robot dancer.

2. PROPOSED SYSTEM

This section explains the internal architecture of the proposed singing robot dancer (Figure 1). Our system mainly consists of listening, dancing, and singing functions that are communicated with each other in an asynchronous manner through data streams managed by the Robot Operating System (ROS). The listening function, which is implemented on an open-source robot audition software called

ISMIR2015 におけるプロシーディングスを転載

*The four high school students contributed equally. This study was supported by the Global Science Campus project, JST, Japan.

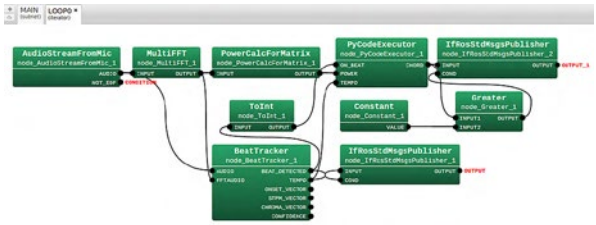

Figure 2. Visual programming interface of HARK.

HARK, takes music audio signals captured by a microphone and recognizes the beats and chords of those signals in real time. The dancing function then receives the recognition results and then determines dancing movements. The singing function also receives the recognition results, determines vocal pitches and onsets, and synthesize singing voices by using a singing-voice synthesizer called eVY1, Yamaha Corp. (MIDI device).

2.1 Listening function

The listening function mainly consists of two modules: beat tracking and chord estimation, which are performed sequentially on the dataflow-type visual programming interface of HARK (Figure 2).

Beat tracking This module, which is included in HARK, is based on an efficient beat tracking method called spectro-temporal pattern matching (STPM) [4]. This method extracts "edges" (spectral components with rapid power increase) from a music spectrogram by using Sobel filters. This is a standard technique of edge extraction in image processing. The tempo and beat times are estimated by calculating the autocorrelation of those components.

Chord estimation This module classifies 12-dimensional beat-synchronous chroma vectors extracted from the music spectrogram into 24 kinds of chords (12 root notes $\times$ 2 types (major/minor)) by using a template matching method based on the cosine distance.

2.2 Dancing function

The dancing function concatenate dancing movements according to the chord progression of the target musical piece. We defined 24 different dancing movements corresponding to the 24 kinds of chords (Figure 3). In fact, a proprietary device driver called NAOqi should be linked to the ROS for controlling the movements of the robot.

2.3 Singing function

The singing function controls the eVY1 device for generating beat-synchronous singing voices whose pitches match the root notes of the estimated chords. eVY1 can be controlled in real time as a standard MIDI device.

3. EXPERIMENT

We conducted an experiment using a sequence of simple chords (toy data) and a Japanese popular song (real data) in a reverberant room. Each of the audio signals were played back from a loudspeaker. The audio signals were captured by using a microphone behind the robot. The distance between the loudspeaker and the microphone was about 1 m.

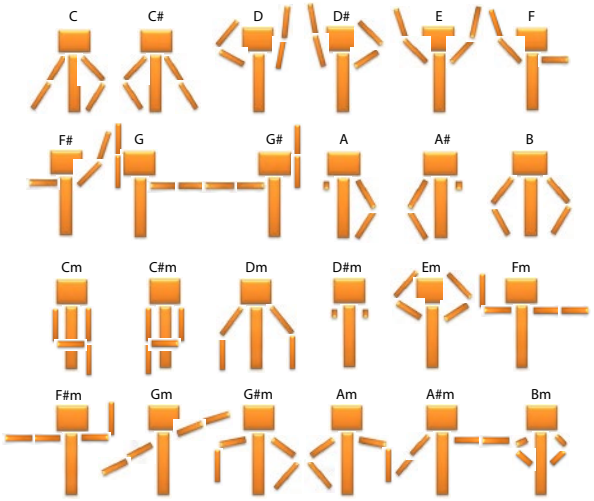

Figure 3. Predefined dancing movements.

The experimental results showed that our robot has great potential as an entertainment robot. It could recognize the chord progressions of both data to some extent and generate chord-aware beat-synchronous dancing movements. The response of dancing and singing, however, was delayed for two beats after new chords began because the robot has no function of chord prediction. Demos are at http://winnie.kuis.kyoto-u.ac.jp/members/yoshii/gsc2015/

4. CONCLUSION

This paper presented a singing robot dancer that can recognize the beats and chords of music audio signals in real time. Our humanoid robot was developed by combining an open-source robot audition software called HARK, signing-voice synthesis hardware called eVY1, and a robot-motion controller in an asynchronous manner through a standard middleware called ROS. Although the experimental results showed the potential of the proposed system, we found that the response of singing and dancing should be improved by implementing real-time prediction of next chords. To enable the robot to use its own ears embedded in the body to capture audio signals, we plan to improve the robustness of the listening function by using dereverberation and ego-noise cancellation techniques.

5. REFERENCES

[1] K. Kaneko *et al.* Cybernetic Human HRP-4C. *Humanoids*, 2009.

[2] K. Kosuge *et al.* Partner Ballroom Dance Robot — PBDR—. *SICE Journal of Control, Measurement, and System Integration*, 1(1):74–80, 2008.

[3] Y. Kusuda. Toyota's Violin-playing Robot. *Industrial Robot*, 35(6):504–506, 2008.

[4] K. Murata *et al.* A Beat-Tracking Robot for Human-Robot Interaction and Its Evaluation. *Humanoids*, 2008.

[5] K. Petersen *et al.* Development of a Aural Real-Time Rhythmical and Harmonic Tracking to Enable the Musical Interaction with the Waseda Flutist Robot. *IROS*, 2009.

ISMIR2015 におけるプロシーディングスを転載