音楽コンサート動画における観客の掛け声のタイミングに合わせた音声絵文字インタフェースの開発

松島 圭佑 * 阿部 優樹 * 坂本 大介 * 小野 哲雄 †

概要. 本研究は，YouTube 等の動画プラットフォームで他の視聴者と一緒に音楽コンサート動画を楽しむための音声絵文字インタフェースを提案する．現在は，ライブコメントにより他の視聴者のリアクションを知ることができるが，これらはテキストや絵文字などの視覚情報に依存するため，視覚障がい者や，ランニング中や作業中であるために画面を見ない聴取者の参加は困難である．これに対し，既存研究はライブコメント中の文字を音声提示する手法を提案したが，コメントの 70%を占める絵文字の提示は検討されていない．さらに我々の初期テストでは，音声絵文字提示が音楽体験を過度に妨げることが判明した．そこで本研究では，音楽コンサートにおいて，観客が声で参加する「コールアンドレスポンス」に着目した．この「コール」文化を応用することで，音楽体験を妨げない絵文字コメントの音声合成が可能になると考えた．具体的には，システムが自動でコールを検出し，そのタイミングで音声絵文字を挿入する音声絵文字インタフェースを提案する．

1 はじめに

Human-Computer Interaction の分野では，Social TV，Watch Party，Public Viewing など，複数の視聴者が動画体験を一緒に楽しむ動画インタフェースが研究されてきた [15, 16, 20]．その中で，YouTube，Twitch，ニコニコなどの動画共有プラットフォームが，そのような「共同視聴体験」の場となっている [2]．これらのプラットフォームでは，主に動画に紐づいたライブコメント機能が視聴者間の主要な交流媒体となり，感情共有や一体感を向上させている [3]．しかし，現在のライブコメント機能はテキストや絵文字が中心であり，その体験は視覚に大きく依存している．そのため，視覚障がい者や，運転や作業をしながら音楽を聴く聴取者は，ライブコメントにアクセスできないという課題があった [1]．

この課題に対し，コメントを音声で提示するアプローチが提案されている [1]．これは，テキストコメントを音声合成で読み上げ，画面を見なくても共同視聴体験を可能にした．しかし，ライブコメントの文字情報の 70%を占める [14] 絵文字の音声提示については検討されておらず，これは感情体験の共有を制限している．そこで本研究は [1] の考え方を絵文字へと応用し，以下の 2 点を実装した．

(1) 絵文字の音声表現である音声絵文字

(2) 楽曲の妨げとならないよう，コールアンドレスポンスを応用した音声絵文字インタフェース

* 北海道大学
† 京都橘大学

2 音声絵文字

絵文字の音声表現である「音声絵文字」の作成を行った．

2.1 音声絵文字の作成

音声絵文字は，先行研究における音声顔文字 [12, 17, 11] や音声アイコン [13, 11]，イヤコン [4, 11] の概念を参考に設計した．音源には先行研究 [17] やインターネットで公開されているもの [8, 6, 18, 9, 10, 7] を利用した．対応付けの基準は以下の通りである．

- 感情系絵文字 : 人の感情表現に伴う音 (音声顔文字)
- 物理的イベントに関連する絵文字：関連する物理音（音声アイコン)
- 上記以外の絵文字：抽象的な音（イヤコン）

2.2 ユーザスタディ

音声絵文字が共同視聴体験に与える影響を調査するため，12 名の参加者を対象に，音声絵文字を提示する条件としない条件で比較実験を行った．その結果，音声絵文字を提示すると「つながり」が有意に向上する一方で，音楽への「妨げ」も有意に増加することがリッカート尺度アンケートの結果から確認された．インタビューから,「妨げ」の原因は提示タイミングと楽曲のビートのズレであることが示唆された．

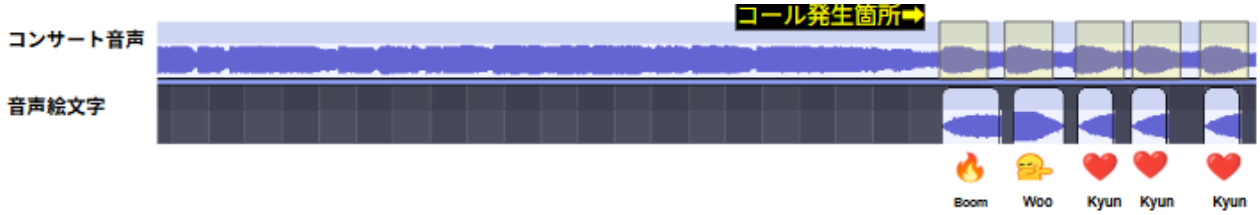

図 1. コールタイミングに合わせた音声絵文字提示

3 “コール” を用いた音声絵文字インタフェース

ユーザスタディから「妨げの原因は提示タイミングと楽曲のビートのズレである」との知見が得られた．この問題の解決策として，観客がビートに合わせて掛け声を出す「コール」のタイミングで音声絵文字を提示する手法を提案する (図 1)．コールは楽曲のビートに同期し，かつ観客の感情が自然に表出されるため，音声絵文字の提示タイミングとして適切であると考えた．

3.1 コール検出システムの実装

コール検出と，音声絵文字の挿入は以下の手順で実施される．まず比較対象の「コンサート音源」と「スタジオ音源」をボーカルと楽器に分離し，スタジオ音源のボーカルを基準に動的時間伸縮法を用いてコンサート音源の歌唱区間と間奏区間を特定する．コール検出は間奏区間のみを対象として行い，madmom ライブラリ [5] で検出した 4 拍子のうち，裏拍（2・4 拍目）で前後の拍より音量が大きい箇所をコールと判定する．最終的に，検出されたコールのタイミングに投稿された絵文字を音声絵文字としてコンサート音源に挿入する．

3.2 技術的評価

コール検出システムの精度を技術的に評価した．本システムの技術的評価は，多様なジャンルのコンサート動画から抽出したコールを含む 5 つのデータセットを用いて実施した．比較基準となる正解データ（歌唱区間，コールタイミング）は，2 名の参加者による手動アノテーションにより作成した．評価指標には，歌唱区間中挿入割合（提案システムがコールと検出したのタイミングのうち，ボーカルの歌唱中にコールと判定した割合），偽陽性割合（コール非存在区間にもかかわらず，提案システムがコールとして誤って検出した割合），偽陰性割合（実際にコールが存在するにもかかわらず，提案システムが検出しなかった割合），および平均誤差（提案システムが検出したコールと，正解のコールタイミングとの時間的なズレの平均値）の 4 つを用いた．評価結果を表 1 に示す．

歌唱区間中挿入割合は平均 5.942%であり，歌唱区間検出に関する先行研究 [19] が示す誤判定割合 3%を上回る結果となった．この差異は，先行研究がスタジオ音源を対象としているのに対し，本研究ではコンサート音源を用いている点に起因すると考えられる．コンサート音源に観客の歓声が含まれるため，観客の歓声と歌唱を識別する必要があるのだが，その際に，歌唱を観客の歓声と誤認してしまったことが歌唱区間の判定精度を低下させた主因と推察される．今後の改善策として，楽曲の歌詞情報とコンサート音源の文字起こし結果を照合し，観客の歓声と歌唱の識別精度を高めるアプローチが有効であると考えられる．偽陽性割合と偽陰性割合は，それぞれ平均 11.36%，28.84%であった．特にデータセット 3 および 4 においてこれらの値が他のデータセットより高い傾向が見られたが，これは両音源のボーカルが男性であることに起因する可能性がある．本研究の音源分離プロセスにおいて，女性ボーカルに比べ男性ボーカルの分離精度が低かったため，後のコール検出に悪影響を及ぼしたと推察される．この課題に対しては，より高性能な音源分離モデルの採用が対策として考えられる．一方で，コールタイミングの平均誤差は最大でも 0.119 秒に留まった．これは，評価に用いたデータセットの 0.5 拍長の最小値 0.170 秒よりも十分に短い値である．したがって，本手法によるコール挿入のタイミング精度は，音楽的な許容範囲内に収まっていると結論付けられる．

表 1. 評価結果

データセット	歌唱区間中挿入割合:%	偽陽性割合: %	偽陰性割合: %	平均誤差:秒
1	6.45	6.45	14.71	0.119
2	0	8.11	26.53	0.103
3	3.03	24.24	35.9	0.039
4	15.79	15.79	34.21	0.087
5	4.44	2.22	32.84	0.089
平均	5.942	11.36	28.84	0.087

4 まとめと今後の展望

本研究では，音楽コンサート動画におけるライブコメント中の絵文字をコールのタイミングに合わせて挿入する音声絵文字インタフェースを開発した．今後ユーザテストを実施し，その有用性と改良を探索したい．

謝辞

本研究は，立石科学技術振興財団，JSPS 科研費（25KJ0518），JST 創発的研究支援事業の支援を受けたものである．

参考文献

[1] Y. Abe, D. Sakamoto, and T. Ono. "I feel lonely when they stop chatting": Exploring Auditory Comment Display for Eyes-Free Social-Viewing Experience in Online Music Videos. In *Proceedings of the 2025 ACM/IEEE International Conference on Computer-Supported Cooperative Work Social Computing (CSCW '25)*, Bergen, Norway, October 2025. ACM.

[2] A. Bartolome and S. Niu. A Literature Review of Video-Sharing Platform Research in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, p. 1–20. ACM, Apr. 2023.

[3] S. Benford, P. Mansfield, and J. Spence. Producing Liveness: The Trials of Moving Folk Clubs Online During the Global Pandemic. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, p. 1–16. ACM, May 2021.

[4] M. Blattner, D. Sumikawa, and R. Greenberg. Earcons and Icons: Their Structure and Common Design Principles. *Human-Computer Interaction*, 4(1):11–44, Mar. 1989.

[5] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer. madmom: A New Python Audio and Music Signal Processing Library. In *Proceedings of the 24th ACM international conference on Multimedia*, MM '16, p. 1174–1178. ACM, Oct. 2016.

[6] F. W. Buhl. EmoGator: A New Open Source Vocal Burst Dataset with Baseline Machine Learning Classification Methodologies, 2023.

[7] Taira Komori 制作/著作 小森平. 無料効果音で遊ぼう！, 2025. Accessed: 01-29-2025.

[8] C. © 2012-2025 OtoLogic. BGM・ジングル・効果音　フリー素材サイト | OtoLogic, 2025. Accessed: 01-29-2025.

[9] 2013-2023 効果音ラボ All Rights Reserved. 効果音ラボ, 2025. Accessed: 01-29-2025.

[10] © 2025 くちぶえ音楽院 All rights reserved. くちぶえ音楽院, 2025. Accessed: 01-29-2025.

[11] c. Csapó and G. Wersényi. Overview of auditory representations in human-machine interfaces. *ACM Computing Surveys*, 46(2):1–23, Nov. 2013.

[12] P. Fröhlich and F. Hammer. Expressive Text-to-Speech: A user-centred approach to sound design in voice-enabled mobile applications. 10 2004.

[13] W. Gaver. Auditory Icons: Using Sound in Computer Interfaces. *Human-Computer Interaction*, 2(2):167–177, June 1986.

[14] J. Guo and S. R. Fussell. A Preliminary Study of Emotional Contagion in Live Streaming. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '20 Companion, p. 263–268, New York, NY, USA, 2020. Association for Computing Machinery.

[15] Z. Lu, H. Xia, S. Heo, and D. Wigdor. You Watch, You Give, and You Engage: A Study of Live Streaming Practices in China. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–13. ACM, Apr. 2018.

[16] S. Schirra, H. Sun, and F. Bentley. Together alone: motivations for live-tweeting a television series. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14. ACM, Apr. 2014.

[17] G. Wersenyi. Evaluation of auditory representations for selected applications of a graphical user interface. In M. Aramaki, R. Kronland-Martinet, S. Ystad, and K. Jensen eds., *Proceedings of the 15th International Conference on Auditory Display (ICAD2009)*, p. N/A, Copenhagen, Denmark, May 2009. International Community for Auditory Display, Georgia Institute of Technology.

[18] G. Wersenyi. Evaluation of auditory representations for selected applications of a graphical user interface, 2009. Accessed: 01-29-2025.

[19] X. Zhang, Y. Yu, Y. Gao, X. Chen, and W. Li. Research on Singing Voice Detection Based on a Long-Term Recurrent Convolutional Network with Vocal Separation and Temporal Smoothing. *Electronics*, 2020.

[20] 高野 祐太郎, 大島 浩太, 田島 公治, 高田 理, 寺田 松昭. 投稿型動画視聴におけるユーザ間リアルタイムコミュニケーション 支援システム. 電子情報通信学会論文誌 D, J93-D(10):2302–2316, 2010.