リアル会場とリモート観客との調和再現技術

NTT人間情報研究所では，ライブ配信イベントをリモートで自宅からオンラインで楽しむ観客（リモート観客）の様子をイベント開催しているメイン会場（リアル会場）で違和感なく合成・提示するために，リモート観客の様子をリアル会場の状況に合わせて調和させつつ再現する技術の研究開発に取り組んでいます．2022年3月21日に開催された「第34回 マイナビ 東京ガールズコレクション 2022 SPRING/SUMMER」において，コロナ禍のため歓声を上げられないリアル会場の観客とリモート観客のために，低遅延映像通信とクロスモーダル検索を使って疑似歓声音をリアル会場で再生して盛り上がりをサポートする実証実験を実施しました．本稿では，この実証実験の取り組みについて紹介します．

黒住（くろずみ） 隆行（たかゆき）　長谷川（はせがわ） 馨亮（けいすけ）
松本（まつもと） 英一郎（えいいちろう）　江浦（えうら） 俊彦（としひこ）
深津（ふかつ） 真二（しんじ）
NTT 人間情報研究所

映像・音響の調和再現の必要性

NTTは，これまで高臨場感の映像を届けることに主眼を置いて，複数の遠隔の視聴環境を相互に接続し，高精細な映像を低遅延に届ける双方向映像通信の研究開発に取り組んできました．マラソン競技におけるリアルタイムリモート応援プロジェクト[(1)]では，非圧縮伝送による超低遅延通信技術と低遅延メディア処理技術により，札幌のマラソンコースと東京の応援会場をリアルタイムにつなぎました．これにより，遠隔地から観客の応援を選手に届け，沿道応援さながらの臨場感と，選手観客の一体感をつくり出し，新たな競技観戦のかたちを実現しました．

NTTは，この取り組みを家庭ユーザ向けに発展させるため，ライブ配信イベントの会場と家庭環境との間を双方向に映像と音響を通信する研究に着手しました．しかし，家庭環境とリアル会場との間で双方向に高臨場な映像通信を実現しようとすると不都合が生じることがあります．例えば，リモートワークをしている家庭環境と職場との間で映像と音を利用するWeb会議をするような場面では，家庭環境側のカメラ映像の背景に生活感のある部屋の様子が映り込んだり，マイク音声に家族の声が混入したりして，気まずいと思うことがあります．また，スポーツやエンタテインメントのライブ配信イベントのような場面で，リモートから参加し自身の存在も映像で届けて会場と一緒に盛り上がりたいと思う反面，このような家庭のあまり見られたくない，聞かれたくない情報がライブ会場に届いてしまったり，配信されてしまったりすることは避けたいと思う観客は多いでしょう．そこで，不要な情報は抑制し，会場で再現して欲しい情報は臨場感を高く，会場と調和のとれた状態で映像や音響を再現することが求められます．

一方，会場側でも調和のとれた再現というものが必要になる場合があります．新型コロナウイルス感染症拡大の影響により，人々の行動が制限され，ライブ会場でも感染防止のために観客はマスクを着用していても歓声を上げることが禁止されている状況にあります．音楽ライブにおいて，新型コロナウイルス感染拡大前はとても盛り上がった楽曲が，会場に観客がたくさん来場しているにもかかわらず歓声を上げられないという状況は，とても寂しくもあり，もの足りなさも感じてしまいます．また，演者の方にとっても，歓声を聞くことができないと観客の反応が分かりにくくなるため，観客とのインタラクションが難しくなります．そこで，会場側においても，新型コロナウイルス感染拡大前と同じように，違和感のない歓声による盛り上がりを提供できないかという問題意識で，映像・音響の調和再現について検討を進めました．

家庭向け双方向映像通信の実現に向けた共同実験

IMAGICA EEX，NTTコミュニケーションズ，NTTは，2022年3月21日に開催されました「第34回 マイナビ東京ガールズコレクション 2022

SPRING/SUMMER」[2]において，家庭ユーザを対象とした双方向高臨場映像視聴の実現に向けた共同実験を実施しました．実験では，コロナ禍のために歓声を上げられないリアル会場の観客の盛り上がりのサポートをすることと，リモートからのオンラインの視聴であっても会場とのインタラクションにより参加感を醸成することをめざして，低遅延映像通信技術とクロスモーダル検索技術[3]を使った疑似歓声音の再生を実現するシステムを構築し，リアル会場の観客とリモート観客の盛り上がりの様子から歓声を再現するアプローチで調和再現の検証を行いました．**図1**は，実証実験の全体像を表す概念図です．家庭環境でライブ配信を視聴している参加者（リモート観客）が，カメラ付きのPCを利用してNTTコミュニケーションズの双方向低遅延通信システム（Smart vLive®[4]およびSkyWay®[5]）を介してリモートから参加し，ステージ前面の左右の大画面ディスプレイに登場します．ここで，リモート観客の左右各々の映像から後述する歓声音の推定を行って，対応する左右のスピーカーから盛り上がりに応じて歓声音を再生します．一方，会場の観客についても，観客席をねらったカメラで撮影した映像から歓声音を推定し，会場スピーカーから歓声音を再生します．歓声音は，観客がペンライトを速く振ると大きくなり，ゆっくり振ると小さくなるよう，歓声音を観客がコントロールできるように調整しました．また，盛り上がりの大きさは，歓声音だけでなく，ライブ配信される映像のXR（Extended Reality）表現にも反映しました．IMAGICA EEXにより配信映像中のリアル会場の観客やリモート観客からの盛り上がりの大きさに応じて光の粒の量が変化する演出を行い，ライブ配信の視聴者は，観客の盛り上がりを音と映像で楽しむことができました．

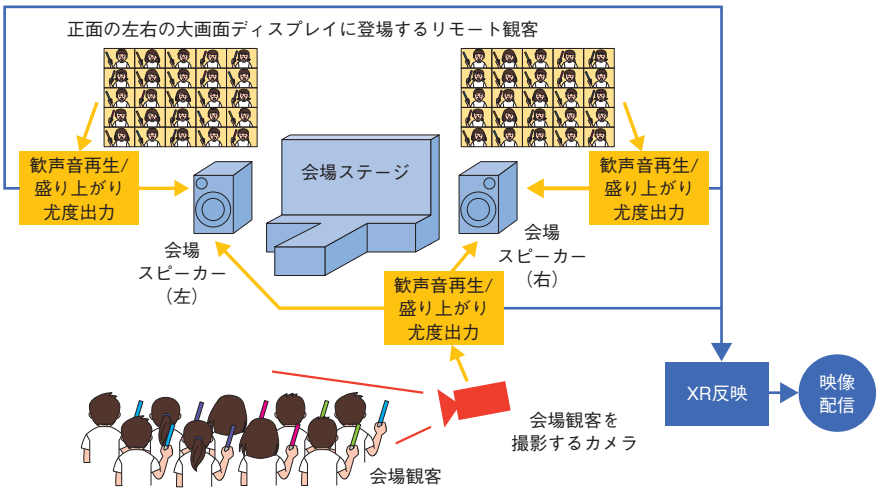

図1　実証実験の概念図

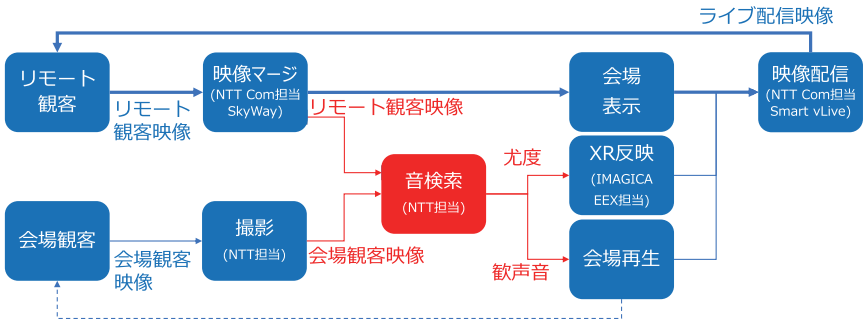

図2　システム構成と処理・情報の流れ

■システム構成

全体のシステム構成を**図2**に示します．実験システムは，リモート観客の映像をタイル状にレイアウトする機能，会場に大型ディスプレイ上に表示する機能，タイル映像と会場の観客映像から歓声音を検索し音を再生する機能，検索結果をXRに表現する機能，これらの結果を反映した会場の様子を映像配信する機能から構成されます．

今回使用した音検索システムは，機械学習に基づく技術を用いています．NTTコミュニケーション科学基礎研究所のクロスモーダル音検索技術[3]を用いて，ペンライトを振る観客の様子を映した映像から歓声音を推定しました．ペンライトを振る映像と歓声音を対応付けるために，あらかじめペンライトを振る会場観客，リモート観客の映像と歓声の音をペアにした学習データを準備し，映像から音を推定するモデルを学習しておきます．本番では，リモート観客については，**図3**に示すように5×5のタイル状にレイアウトして集約した映像を入力して，ペンライトを振る様子から対応する歓声音を検索し再生しました．一方，会場観客については，**図4**に示すように会場観客を撮影するためのカメラを会場内に設置して観客を撮影し，こちらも同様にペンライトを振る様子からその映像に対応する歓声音を検索し再生しました．再生する音源は，あらかじめ録音して用意した歓声音を使用しました．

ここでクロスモーダル音検索技術を使用した歓声音の検索に加えて，歓声

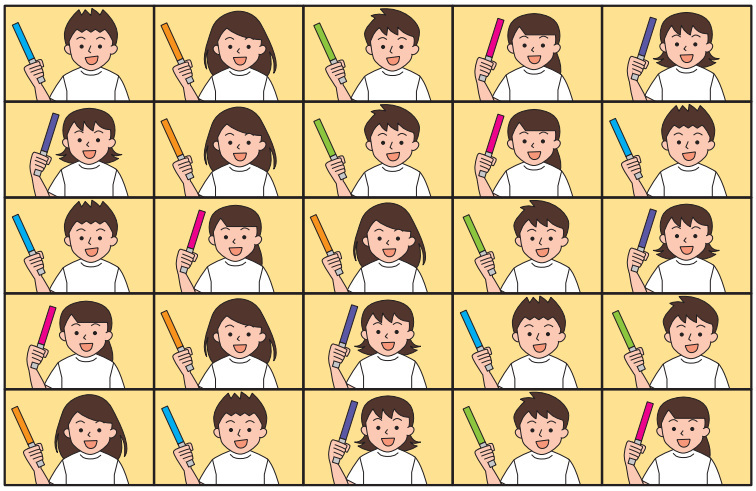

図3　ペンライトを振るリモート観客映像をグリッド状に配置し集約した映像

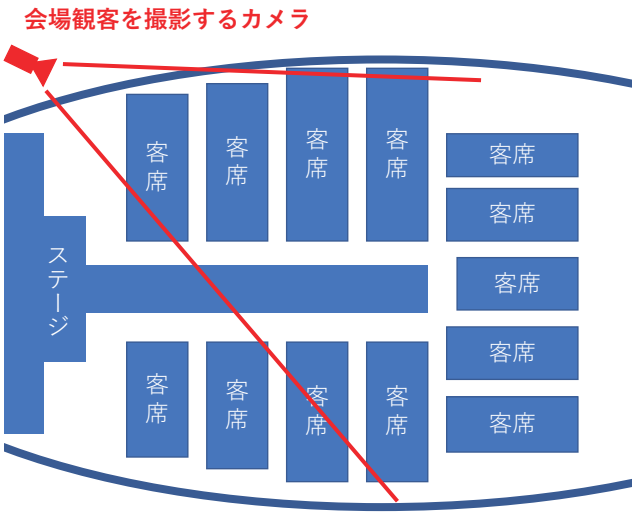

図4　会場観客を撮影するカメラの配置

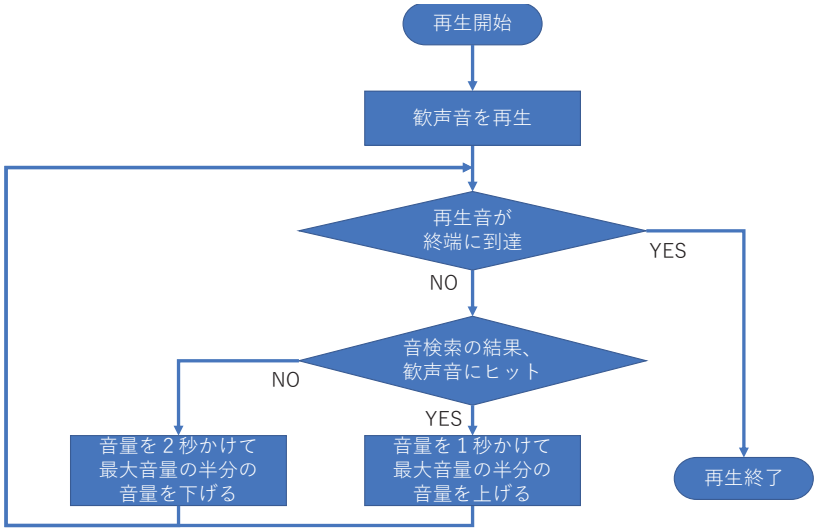

図5　音量決定のフローチャート

音の音量を滑らかに変化させるために，**図5**に示すフローチャートにより音量を決定する手法を実装しました．歓声音の検索がヒットの個所とそれに対応する音量変化の例を**図6**に示します．この仕組みにより，ペンライトの振りを継続すると音量が大きく，ペンライトの振りを止めると音量が小さくなるように制御でき，直感的な音量の変化を実現させることができます．

■参加者へのアンケートによる評価

今回の実証実験では，観客映像から歓声音を推定して会場で再生する調和再現の有効性を確認するために，実験期間中にリモート観客として参加いただいた被検者の方々へのアンケートを実施しました．「通常の視聴をするだけの配信と比べると，歓声が出る仕組みはあったほうが良いと感じた．」という一文に対して，「当てはまる」から「当てはまらない」までの5段階評価で回答する質問に対し，86.6%の方が「当てはまる」と回答しました（**図7**）．参加していただいた方の多くが調和再現する仕組みに対して好意的にとらえており，観客の反応を会場へ届けることの有効性を確認することができました．

今後の展開

今回，リモート観客には，自宅から1人ずつ接続いただくという形式で参加いただきましたが，どのような視聴環境でリモートから参加することが望ましいかについての調査もアンケートで実施しました．「ライブ配信を遠隔から鑑賞する状況として，次のいずれの視聴状況を望みますか」という複数回答可能な質問に対して，68.8%と半分以上の方が「自宅・友人宅に集まり1台のスマートフォン・PC・モニターから友人と一緒に参加」と回答しました（**図8**）．この結果は，リモートの

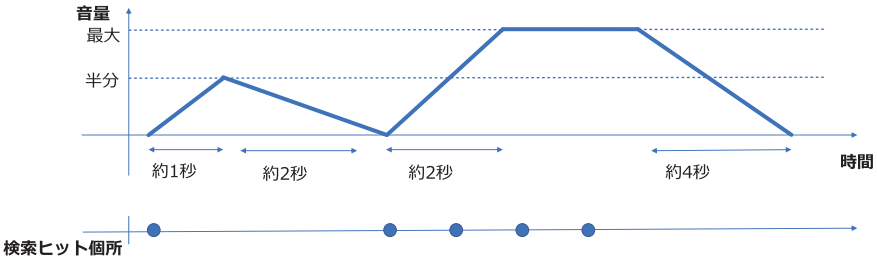

図 6　検索ヒット個所と音量の変化の例

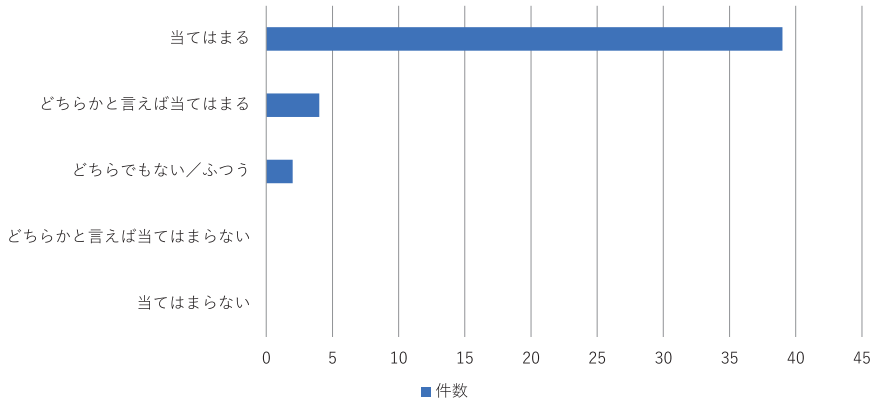

図 7　通常の視聴するだけの配信と比べると，歓声が出る仕組みはあった方が良いと感じた（45人回答）

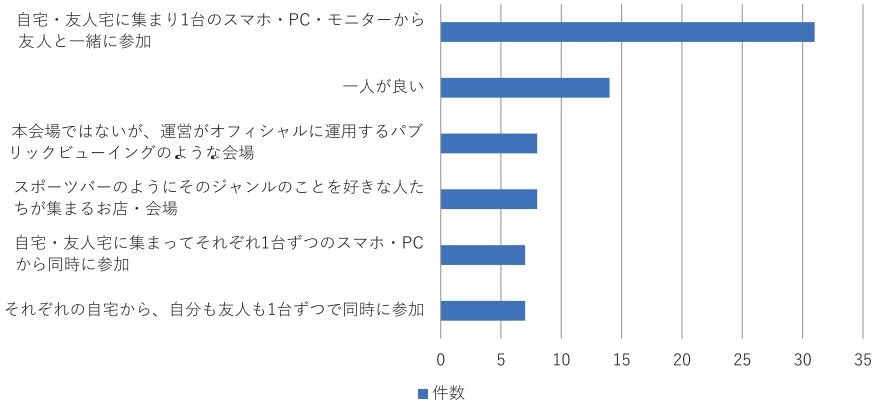

図 8　ライブ配信を遠隔から鑑賞する状況として，次のいずれの視聴状況を望みますか（45人回答，複数回答可）

視聴環境に仲の良い友人どうしが集まって参加し，そのような視聴環境が多数，会場に接続されて，一緒に視聴するという視聴スタイルが，今後，求められるということを示唆しているかもしれません．パブリックビューイングのような整備された環境であれば，会場の雰囲気も含めて高い臨場感を追求してリッチに伝送しても問題ないかもしれません．しかし，家庭向けの場合は，先述したように，家庭環境をカメラで撮影したそのままの映像やマイクで収音したそのままの音を伝送して会場で再生したのでは，演出上の問題が発生することが予想されます．このようなことから，NTTは，リアリティへの追求だけでなく，より強調したい情報，抑制したい情報があることを考慮して情報を選択し，ライブ配信に支障がないように調和した映像や音響を再現することができる双方向映像通信の研究開発を進めていきます．

■参考文献

(1) 薄井・深津・松本・井元・白井・木下：“マラソン × 超低遅延通信技術，” NTT技術ジャーナル，Vol. 33, No. 10, pp. 30-34, 2021.
(2) https://tgc.girlswalker.com/22ss/
(3) M. Yasuda, Y. Ohishi, Y. Koizumi, and N. Harada：“Crossmodal sound retrieval based on specific target co-occurrence denoted with weak labels,” in Proc. of Interspeech, 2020, pp.1446-1450, Oct. 2020.
(4) https://www.ntt.com/business/services/voice-visual-communication/business-support/smartvlive.html
(5) https://webrtc.ecl.ntt.com/

（上段左から）黒住 隆行/ 長谷川 馨亮/ 松本 英一郎
（下段左から）江浦 俊彦/ 深津 真二

ライブ配信イベントをリモートで自宅からオンラインで楽しむ観客に対して，リアル会場での参加では体験できなかった新しい体験を提供できるような双方向映像通信の実現をめざして研究開発を進めています．

◆問い合わせ先
NTT 人間情報研究所
サイバー世界研究プロジェクト
E-mail　dsg-contact-p @ hco.ntt.co.jp