音声収録における頭部運動の Kinect による測定

Measurement of head movements using Kinect in voice recording

5119E003-6　今枝 文彦　　指導教員　及川 靖広　教授

IMAEDA Fumihiko　　Prof. OIKAWA Yasuhiro

概要：音声の収録を行う場合，本来は発声およびマイクロホンの指向特性や明瞭度を考慮し，発声者とマイクロホンの位置と方向を不変とするのが理想的である．しかし実際の収録では発声者の位置や向きが変化してしまい，伝達関数の変化や，近接効果といった周波数特性に関する問題が生じる．発声者自身にマイクロホンを装着することで位置や方向を不変にできるが，マイクロホンの種類が限られ音作りの幅が狭まってしまう．したがって，音声収録においては発声者の身体，特に頭部の運動を同時に測定し，運動が与えた影響に対して処理を行うことが望ましい．そこで本研究では，Azure Kinect DK デバイスを用いて，音声収録時に発声者の頭部運動が計測できるシステムを提案した．実験によって，デバイスに対し発声者が正面 1m ほど離れた位置が適切であることを示した．また，歌唱中の動きによって収録された音声の周波数ごとのエネルギーに変化が生じることを確認した．

キーワード：Azure Kinect DK デバイス，body tracking，深度センサ，マイクロホンアレイ，アバター

Keywords: Azure Kinect DK device, body tracking, depth sensor, microphone array, avatar.

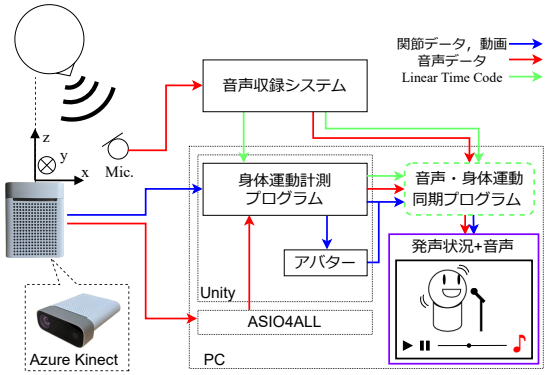
1. ま え が き

音声収録を行う場合には，発声者とマイクロホンの位置と方向を不変とするべきであるが，特に歌や長時間の収録の場合には位置と方向の変化が避けられない．またこの課題に対し，音作りの幅を狭めないためにマイクロホンの種類を限定しない方法が求められる．近年開発された Azure Kinect には，RGB カメラ，深度センサ，加速度計とジャイロスコープ (IMU)，マイクロホンアレイが搭載されており，動きや位置，音声の収録などを非接触でリアルタイムに行うことができる．本研究では，Azure Kinect DK デバイスを用いた音声・身体運動同期収録システムを提案し，発声者の位置・方向によるシステム性能および収録音声の変化を確認した．

2. 音声・身体運動同期収録システム

音声収録と発声者の身体運動の同期計測を行うには，音声収録システムと身体運動計測プログラムでの記録が同期されることが必要である．身体運動計測プログラムによって Azure Kinect からのデータを記録すると同時に，音声収録システムが生成するタイムコードをオーディオ信号として記録する．これにより音声と身体運動の同期収録を可能とする．

音声収録に向けて Azure Kinect で測定すべきデータについて説明する．頭部運動の記録のためには，まず頭部付近の特徴点における 3 次元の位置情報と角度情報が不可欠である．さらに計測において，実際の頭部の動きを直感的に理解できるように，収録状況を動画として記録すべきである．音声・身体運動同期収録システムの概要を図–1 に示す．メインの音声に加え以下の 4 種類のデータを記録する．



図–1　音声・身体運動同期収録システム

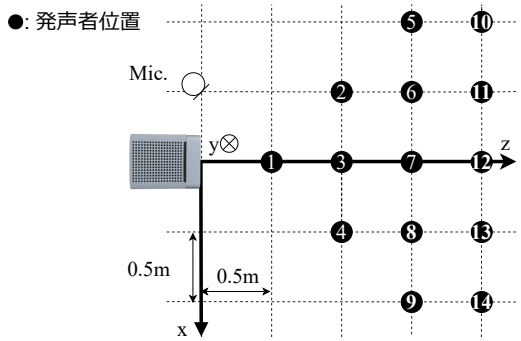
1) 深度カメラと IMU による 32 関節データ
2) 7 マイクロホンアレイによる音声データ
3) カラーカメラによる動画
4) 音声収録システムが生成したタイムコード

3. 実　　験

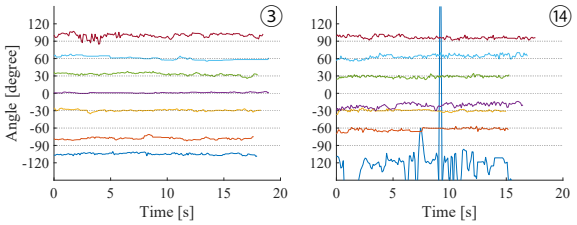
音声収録中における発声者の位置と向きの変化を本システムを用いて測定し確認するために，会議室にて実験を行った．実験における空間座標軸と位置関係を図–2 に示す．いずれの実験においても，マイクロホンと Azure Kinect の高さはともに 0.9 m，音声の標本化周波数は 48 kHz，Azure Kinect 深度モードは NFOV Unbinned とした．本実験においては実際の位置や回転角を確認しやすくするために椅子に座った状態で発声を行った．

3.1 発声者の位置と向きによるシステム性能の変化

Azure Kinect に対する発声者の位置と角度ごとに，本システムに性能差が生じるか確認するため測定実験を



図–2　測定実験での位置関係



図–3　発声者の位置と方向ごとに測定された角度の変化

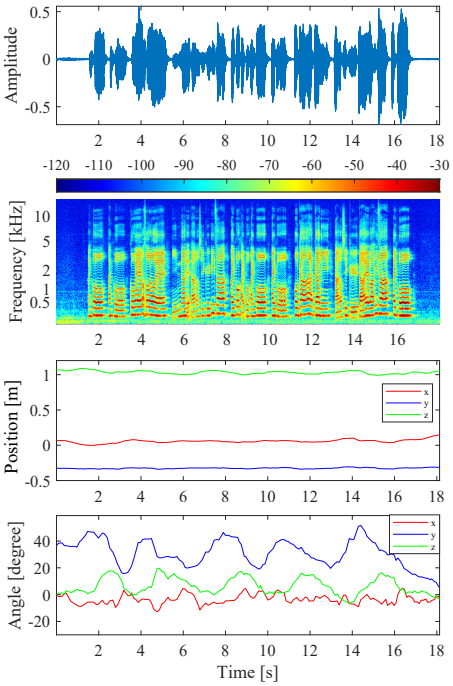
行った．図–2 に示す 14 箇所それぞれにおいて z 軸負の方向を 0 度とし，左から右向きにかけて-90 度から 90 度まで 30 度ごとに測定を行った．発声に用いるテキストには JSUT コーパス [1] を使用した．図–2 における③と⑭の結果を図–3 に示す．外側になるにつれ，特により遠い場所に位置する⑭においては設定方向と実測値の差が大きくなっている．要因として実際の発声者の初期方向がずれていたことに加え，Azure Kinect に対し遠い位置や横向きでは，複数の関節が重なり認識できない場合があると考えられる．したがって，収録の際には③ のように Azure Kinect 近く正面に位置するのが適切と考えられる．しかし，① においては近すぎて被写体が Azure Kinect の画角に収まらない場合も生じ，ほとんどの角度，時間において適切に測定ができなかった．現状では，z 軸方向 0.5 m を超えた場所に位置すべきと考えられる．

3.2　歌唱中の位置および角度変化

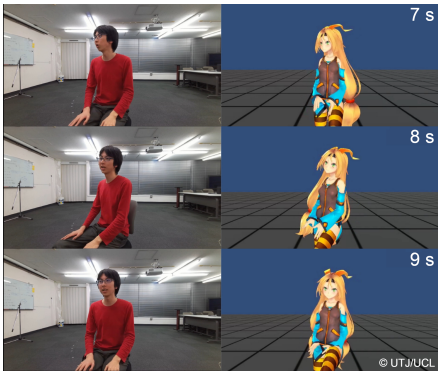
次に，図–2 の③において「ハイ・ホー」を自由に歌唱して，収録を行った．歌唱中は，身体ごと左右に向きを変え約 3 秒で往復を繰り返す動作をした．図–4 に収録した音声波形，スペクトログラム，各座標ごとの発声者の頭部位置，回転角度を示す．位置は最大 0.118 m，角度は最大 32.2 度の変化があった．約 3 秒間隔で動きが繰り返されていること，それら動作には微妙な動きの違いがあることが見て取れる．

3.3　アバターを用いた発声者の動作の再現

本システムにより計測した頭部位置や回転角度から発声者の動作をアバターを用いて再現することが可能である．これにより収録した音声が発声された状況をアバターによっても直感的に確認することが可能である．



図–4　音声・身体運動収録データ



図–5　アバターでの発声動作再現

Azure Kinect に搭載されているカメラで収録した映像と，収録したデータから動作を再現したアバターの様子を図–5 に示す．再現した様子からもデータ収録が行われていることを確認できる．

4.　む　す　び

本研究では，マイクロホンでの収録に加えて Azure Kinect を用いて頭部運動を記録するシステムを構築した．今後の研究では，7 マイクロホンアレイデータを利用した発声の指向特性推定等の検討を行う．また，頭部運動と収録音声のパラメータの関係性を明らかにするとともに，それに基づいて頭部の位置と方向の変化による収録音声の変化を補正するなどの処理につなげていく．

参 考 文 献

[1] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” Acoust. Sci. & Tech., vol. 41, pp. 761–768, 2020.