5B-06

# 二者対話中の頭部動作を用いた沈黙推定

善本淳†

情報通信研究機構†

## 1 はじめに

一般的に人は，話者が発する音源位置推定，話者の声質の他，動作に関する情報を用いるなど，視聴覚情報を統合して話者推定を行っていると考えられる．仮に，低品質な映像，並びにモノラルマイク録音による動画の再生であったとしても，人にとって話者推定タスクの難易度はそう高くはならないのではなかろうか．

カメラ・マイクロホンアレイ・深度センサ等を用い，話者ダイアライゼーション（複数の話者が存在する場にて，いつ・誰が発話を行ったのか，の同定）は従来から既に，幅広く研究されており(Wakabayashi[1]等），例えば一般論として発話時には口を開くことから，口唇変化の映像情報を取り入れた話者ダイアライゼーション等も，また平行して行われてきた．

ここで報告者は，話者ダイアライゼーションの一環として，対話中の話者の正面画像を録画した低解像度の動画を用い，計算機を利用して，発話の有無の推定を行う事を検討した．その推定の際，推定材料として，人が対話中に表出する非言語的な各身体パーツ位置の特徴を利用する方針で進めた．

## 2 実験

まず，話者三人(話者 A / 話者 B /話者 C)を準備した．話者 AB 間，次に話者 AC 間で，対話を着座状態で行わせた．そこからそれぞれ 554 秒(16,602 フレーム)，並びに 473 秒(14,176 フレーム)の対話動画が得られた．またその間，話者 A は 89.3 秒間，並びに 118.0 秒間の発話を行っていた．

動画は話者の正面やや下方から撮像され，左右の腕を含む，腰から上の画像が得られた．（図 1 参照）なお，撮像された画像は被験者一人当たり幅 360 ピクセル×高さ 240 ピクセルであり，頭頂から腰までが入る構図で撮像されているために人物像は比較的小さく，例えば左右耳介間距離は 35～60 ピクセル程度であった．そのため発話等による口唇の開閉は画像上では 1～数ピクセル以内の幅で判断する必要があり，仮に切り出された画像を人が見て判断したところで，明確に口が開いているとも閉じているとも言い難い画像が少なからず含まれていた．

## 3 演算

得られた動画像をフレーム毎に切り出し，それぞれの話者の 13 箇所の身体位置（右目・左目・右耳・左耳・鼻・首・腰・右肩・右肘・右手首・左肩・左肘・左手首）の二次元座標(以下，身体情報)を算出した．また同様に，70 箇所の顔ランドマーク位置の二次元座標（以下，顔情報）も算出した．これらの算出には OpenPose[2, 3] を利用した．これにより，1 フレームあたり身体情報 26 次元，及び顔情報 140 次元の変数が得られた．

今回対象とした推定項目は，話者 B/話者 C の両者と対話を行った話者 A 自身の発話・沈黙の

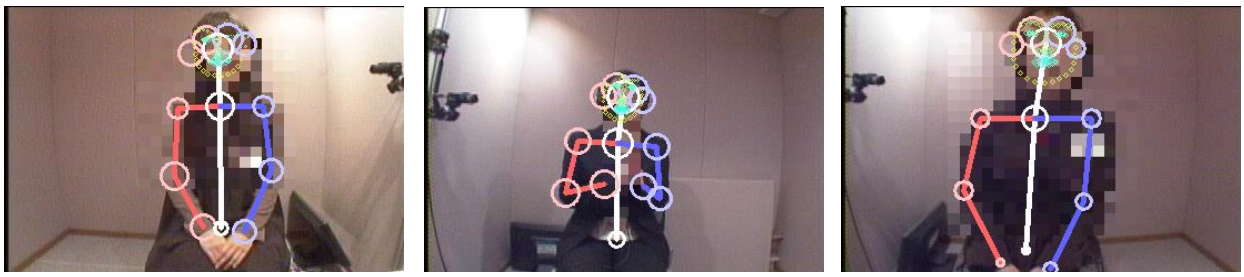

**図 1．被験者正面画像，並びに身体と顔の認識結果の表示図**
**左から被験者 A，B，C の撮像図．身体情報 13 箇所/人，顔情報 70 箇所/人の位置を推定し，元画像上に合わせて示した．**（被験者のプライバシー保護のため，ここでは人物にモザイク処理を施している．）

Silence Detection Based on Speaker's Head Movements
†Jun YOSHIMOTO (NICT)

ため，話者 A における実際の発話の有無を教師信号として用いた．(発話フレームを 1，沈黙フレームを 0 として学習．)

推定時には話者 B， C の音響情報を用いず，合計 1,027 秒の動画情報に対し，各フレームにおける話者 A の発話の有無を推定させた．

また今回，計算機を用いた教師付き学習による，発話あるいは沈黙に関する推定を行うにあたり，RNN (LSTM)を用いた．フレームワークには Chainer[4]を用い，四層・全結線・中間層ノード数 2048 とし，最適化アルゴリズムには Graves's RMS prop (Graves[5]) を利用した．活性化関数 ReLU の条件下，話者一人当たり前述 166 次元(26 次元 + 140 次元)等の入力値を用いた．学習には全情報の 2/3 を用い，推定には残り 1/3 の情報を用い，交差検定によって推定結果を得た．

## 4 結果

話中正面顔の身体情報や顔情報から，発話，あるいは沈黙に関する推定に関して表１に示す結果が得られた．26 次元の身体情報よりも，166 次元の身体並びに顔情報を併用した方が，より高い推定精度が得られた．

従来手法による対話中側面動画像からの動作量情報 109 次元[6]を用いた推定と比較しても，顔情報の有無に関わらず全体として同等程度,あるいは精度が向上した推定結果となった．

さらに，教師信号を時間的にスライドさせ，現在よりも 0.5 秒後の被験者 A の発話状態（発話/沈黙）の学習を行い予測推定させたところ，通常推定に比べ精度低下は認められるものの，予測不可能では無い事が分かった．

## 5 考察

発話推定の精度が，沈黙推定の精度よりも著しく低い理由は，沈黙を保ったまま無声で相槌を打つ動作と，閉口状態で発話しながら相槌を打つ動作が，類似している事が原因かと考えられる．また声を出さずとも，口を開いたままの状態もあるため，推定は容易ではなかったと考えられる．

そのため発話推定に用いるならば現状では利用困難だが，沈黙推定に用いるならば本報告の手法を用いることも可能かと考えられる．

例えば，人が機械相手に用いる音声対話用インターフェイスにおいて，利用者が沈黙維持の兆候(0.5 秒先予測)を示し続けているならば音声案内を続行し，反対に発話の兆し(0.5 秒先予測)を見せ始めたらならば音声案内を早急に切り上げるなど，より人に近い，相手話者の非言語情報を用いた，自然な発話のターン交代の実装に役立てることも，難しくはないと考えられる．

参考文献


[1] Wakabayashi, Y., Inoue, K., Nakayama, M., Nishiura, T., Yamashita, Y., Yoshimoto, H., and Kawahara, T., "Speaker Diarization and Source Number Estimation Based on Audio-Visual Integration. *IEICE*, Vol. J99-D, No. 3, pp.326-336, 2016.

[2] Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y., "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields", arXiv preprint arXiv:1812.08008, 2018.

[3] Simon, T., Joo, H., Matthews, I., and Sheikh, Y. "Hand Keypoint Detection in Single Images using Multiview Bootstrapping", arXiv:1704.07809[cs.CV], 2017.

[4] Tokui, S., Oono, K., Hido, S., and Clayton, J., "Chainer: a Next-Generation Open Source Framework for Deep Learning", Workshop on Machine Learning System in 29th conference, Neural Information Processing Systems, 2015.

[5] Graves, A., "Generating Sequences With Recurrent Neural Networks", *arXiv*: 1308.0850., 2013.

[6] 善本淳, "二者対話中の動作を用いた沈黙推定",情報処理学会第 81 回全国大会, 2019.


表１．異なる推定手法による発話状態（発話／沈黙）推定精度の比較

| 発話推定 | 適合率 | F値 |
|---|---|---|
| 側面身体動作量[6] | 0.40 | 0.29 |
| 正面身体 | 0.48 | 0.41 |
| 正面身体＋顔情報 | 0.64 | 0.53 |
| 正面身体＋顔情報(0.5秒先予測) | 0.69 | 0.37 |

| 沈黙推定 | 適合率 | F値 |
|---|---|---|
| 側面身体動作量[6] | 0.82 | 0.87 |
| 正面身体 | 0.85 | 0.87 |
| 正面身体＋顔情報 | 0.87 | 0.90 |
| 正面身体＋顔情報(0.5秒先予測) | 0.84 | 0.90 |