

二者対話中の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

防音ガラスの向こう側で幾人かが雑談をしている時、それを客観的に観察している第三者が存在したならば、その第三者は現在どの人物が発話をしているのか、（あるいは沈黙しているのか）という話者推定や沈黙推定を行う事はそう難しいことではないのではなかろうか。

一般的に、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、人は視聴覚情報を統合して話者推定を行っているとい考えられる。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、話者推定タスクの難易度はそう高くはならないと予想できる。

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（Wakabayashi¹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、対話中の話者の横顔という口唇情報量の比較的少ないシーン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頭部動作の特徴を利用する方針で進めた。



Figure 1. 頭部領域図例（匿名化の為、画像処理済）

なお、報告者がこのような着想に至った経緯は、まず一般的な日本人話者の雑談等において散見されるのは、口を明瞭に開かずに、閉じたまま息を鼻に抜き、同意を示す相槌の“うん”等の発話の存在が少なくない点にある。即ち、話者ダイアライゼーションの解決法として、発話推定を口唇情報だけに頼れば、この閉口相槌を見落としてしまう事になる。そこで、閉口状態であっても開口状態であっても、非対面状態（例：電話などの音声通信）ですら自然と発生する話者の頭部情報に着目すれば、解決に至るのではないかと、というのが元となっている。

また、沈黙推定が可能となった場合に音声認識と組み合わせると、沈黙状態ならば音声認識をオフとし、転じて沈黙状態で得られた音響情報はノイズ、あるいはターゲット人物以外の音響情報であるため、音声認識の必要が無いというメリットを享受する事も可能となる。

2 演算手法

まず、話者を三人（話者 A/話者 B/話者 C）を準備した。次に、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）の対話情報が得られた。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。

対話状態を側面から撮影し、得られた動画像からそれぞれの話者の横顔画像として、話者一人当たり高さ 100 ピクセル・幅 150 ピクセル（Figure 1. 参照）を切り出し、そこから前後の他フレームとの画像差分から動作特徴量として、異なる単位時間間で動作が生じている二次元一組の中心座標 16 組/話者で 64 次元、その動作量 15 次元/話者で 30 次元、動作量比率や自己相関動作等で 15 次元とし、合計 109 次元の入力値を算出した。また、動作モーメント等の特徴量を算出するためのフレーム間画像処理には、OpenCV を用いた。

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”

であり、話者 A における実際の発話の有無を教師データとして用いた。(発話フレームを 1, t_i 沈黙フレームを 0 として学習.)

推定時には音響情報を用いず、前述解像度を持つ合計 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 Chainer²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 (Graves³)を利用した。活性化関数 ReLU の条件下、前述 109 次元の入力値を用いた。

3 結果

話中横顔の頭部動作から、発話、あるいは沈黙に関する推定に関し、発話時はそれぞれ適合率 0.40, 再現率 0.23, F 値 0.29 となり、また同様に沈黙時はそれぞれ適合率 0.82, 再現率 0.91, F 値 0.87 となった。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。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その可能性は少なくともあると考えられる。

4 考察

人は楽しい時には軽い笑みをたたえたまま、口を半開きに声を出さないこともあれば、反対に閉口状態であっても有声で意思表示をする事もある。これらのことから口唇状態の情報のみでは発話の有無を高精度に窺い知る事は困難である。さらには相槌を打つとき、それが有声であっても無声であっても、表出される動作は一般的に類似している。この僅かの差を RNN を用いて正確に分類することは、今回用いた特徴量だけではやや困難であることが判明した。

今後は、より推定を容易にする特徴量の探索が必要となるだろう。

参考文献

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

pp. 326-336, 2016.

- [2] 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.
- [3] Graves, A., "Generating Sequences With Recurrent Neural Networks", *arXiv*: 1308.0850., 2013.

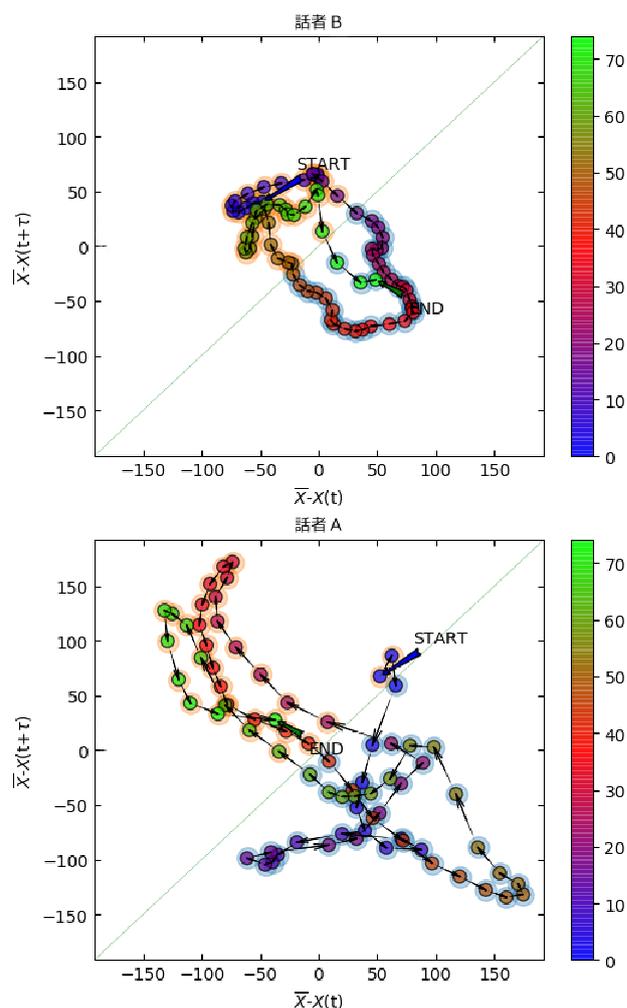


Figure 2. ある時刻における動作自己相関例