Vol.2020-MUS-127 No.30
Vol.2020-SLP-132 No.30
2020/6/7

ステレオ分散録音された対話音声に対するDNNを用いた発話区間検出

河内 秀人[1,a)]　若林 佑幸[1]　小野 順貴[1]　越智 景子[2]　大和田 啓峰[3]　児島 正樹[3]　嵯峨山 茂樹[3]
山末 英典[4]

概要：本研究では，自閉スペクトラム症の診断支援を目指し，自閉スペクトラム症の症状をもつ被験者の音声分析，特に韻律特徴量，対話特徴量の自動抽出を行うために，対話音声データから発話区間を検出する．実施者と被験者両方の襟元に設置した分散マイクで録音したステレオ音声データから，短時間フーリエ変換を用いて，実施者と被験者それぞれのスペクトログラムを音響特徴量として抽出し，ニューラルネットワークを用いて，発話区間を検出する手法を提案する．客観評価実験を行い，音声信号のパワーに対する閾値処理を用いたベースライン手法よりも正答率が高いことを示す．

Deep Neural Network Based Voice Activity Detection for Dialog Speech Recorded by Stereo Distributed Microphones

Abstract: In this study, aiming to assist the diagnosis of autism spectrum disorders (ASD), we report the voice activity detection from dialogue speech data to analyze the speech of participants with ASD, in particular, to automatically extract prosodic and dialog features. We collected the dialog speech of an administrator and a participant that was obtained as a stereo recording with wireless lavalier microphones attached to a collar or other clothing of each speaker. We propose a method of voice activity detection by using both speakers' spectral features and neural network. An objective evaluation showed that the proposed method achieved higher accuracy than a power-thresholding based baseline method.

1. はじめに

自閉スペクトラム症（Autism Spectrum Disorder; ASD）は，社会的コミュニケーションの障害等を特徴とする発達障害である [1]．対話における症状の例として，ピッチの変動幅が小さく，抑揚が平坦である，発話のタイミングが独特である，話速が遅い，全体の印象に違和感があるなどが挙げられる [2]．現状，その症状を定量的に診断する確立された方法としては，あらかじめ定められた質問を用いた実施者と被験者の対話を録画した動画に対し，訓練された実施者がスコアをつける Autism Diagnostic Observation Schedule（ADOS）[3] という方法があるが，スコアリングに手間がかかる，継時的な症状の変化をとらえにく，評価が臨床家の主観に依存しているなどの問題がある．そこで我々は，実施者と被験者の対話音声データを用いた自閉スペクトラム症の症状の定量化を目指し，被験者の音声分析，特に韻律特徴量，対話特徴量の自動抽出の研究 [4][5] を進めている．

被験者の音声から，音声のパワーやピッチのような韻律特徴量，発話長や話者交替に要する時間のような対話特徴量を抽出するためには，その前段階として，話者の発話区間を求めることが必要となる．これまでの研究 [4][5] では，特徴抽出を正確に行うために発話区間は人手で求めていた．本研究ではこの自動抽出を目指し，対話音声データからの発話区間検出について検討する．発話区間検出はこれまで様々な手法が検討されており，多くの先行研究がある [6]．本研究では発話区間検出のアルゴリズムだけでなく，音声データの取得法もあわせて検討している．具体的

[1] 東京都立大学
Tokyo metropolitan university
[2] 東京工科大学
Tokyo university of technology
[3] 東京大学
The university of Tokyo
[4] 浜松医科大学
Hamamatsu university school of medicine
[a)] kawauchi-hideto@ed.tmu.ac.jp

Vol.2020-MUS-127 No.30
Vol.2020-SLP-132 No.30
2020/6/7

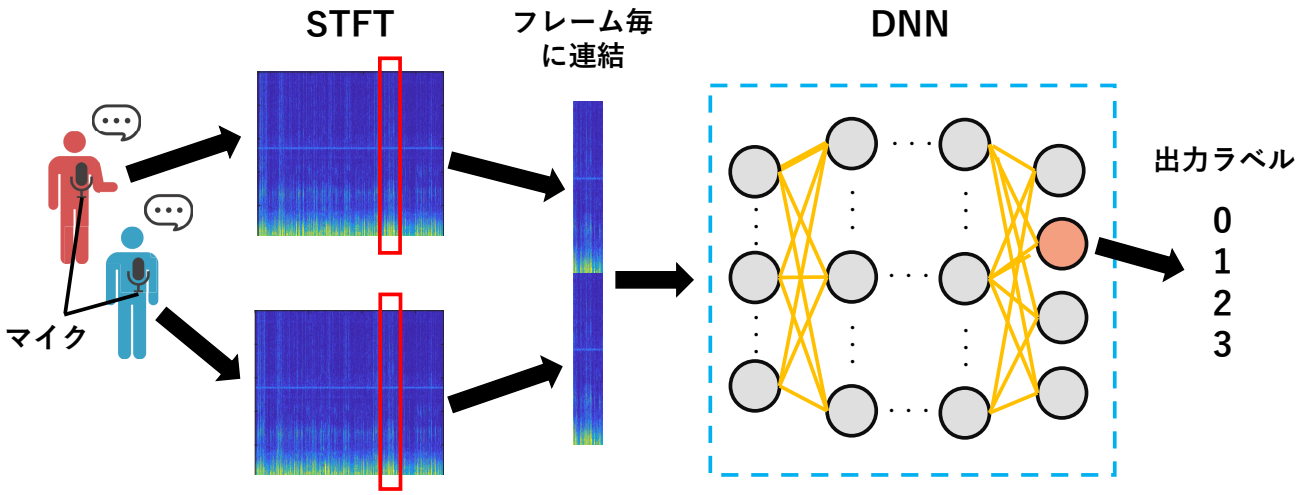

図 1　提案手法の処理の流れ
Fig. 1　Processing flow of the proposed method

には，実施者と被験者それぞれの襟元にワイヤレスマイクを固定し，分散マイクによるステレオ録音を行うことで，診察室で臨床診断として行われた ADOS 実施中の対話における雑音の影響を低減している．本研究ではこの分散録音を活用し，ディープニューラルネットワーク（Deep Neural Network; DNN）を用いて分散録音に含まれる音声特徴を学習することで，発話区間の推定を行なう．具体的には，実施者，被験者，両方の音声のスペクトルを音響特徴量として DNN に入力し，実施者，被験者それぞれの発話／非発話状態をフレーム毎に推定する．客観評価実験を行い，音声信号のパワーに対する閾値処理に基づくベースラインの発話区間検出手法 [7] との比較を行う．

2.　分析方法

図 1 に本研究で行った処理の流れを示す．まず，対話音声データを元に，学習するために必要である音響特徴量の抽出と正解ラベルの作成を行う．実施者と被験者の音声データに短時間フーリエ変換 [8][9]（Short-Time Fourier Transform; STFT）を行い，時間フレーム n における STFT の対数パワーをベクトルとして表したものをそれぞれ $\boldsymbol{x}_{\mathrm{D}}(n), \boldsymbol{x}_{\mathrm{T}}(n) \in \mathbb{R}^{K\times 1}$ で示す．ここで，K は周波数ビン数である．本研究では，フレーム毎に音声/非音声区間を推定するため，以下のように分散録音された音響特徴量を連結し，DNN の入力とする．

$$\boldsymbol{x}(n) = \begin{pmatrix} \boldsymbol{x}_{\mathrm{D}}(n) \\ \boldsymbol{x}_{\mathrm{T}}(n) \end{pmatrix} \tag{1}$$

次に，各対話音声に人手で付与された発話区間の時間データを基に，音響特徴量のフレーム毎に正解ラベルを作成した．1 フレームのうち，半分以上発話しているフレームを発話フレーム，発話が半分未満の場合を非発話フレームと定義し，実施者も被験者も発話していない区間を 0，実施者のみ発話している区間を 1，被験者のみ発話している区間を 2，実施者も被験者も発話している区間を 3，のように 4 つのクラスに分類した．こうして作成した音響特徴量，正解クラスを入出力データとし，DNN を学習した．

3.　評価実験

3.1　実験条件

実験に用いた音声データは，2014 年 11 月から 2016 年 6 月に実施した，成人男性被験者を対象とした多施設・並行群間比較・プラセボ対照・二重盲検によるオキシトシン経鼻剤投与臨床試験（UMIN000015264）の一環として収録されたものである．本研究では，東京大学医学部附属病院精神神経科外来において評価された投薬前被験者 65 名の，ADOS モジュール 4（流暢に話す被験者向け）の課題 7「感情」における会話音声を対象とした．なお，録音の不備等があったデータを除いたため，今回対象としたデータは被験者 61 名で，音声データの合計時間は，3 時間 58 分 02 秒であった．その内，被験者 42 人の 2 時間 48 分 33 秒の音声データを学習用に，被験者 19 人の 1 時間 9 分 29 秒の音声データを評価に使用した．これらの音声データは，サンプリング周波数 44100Hz，もしくは 48000Hz で録音されたものであるが，本研究では後者はダウンサンプリングし，サンプリング周波数 44100Hz に統一して実験を行った．

STFT は，分析フレーム長を 1024 点，フレームシフトを 512 点とし，窓関数にはハミング窓を使用した．DNN は，簡単のため全結合フィードフォワードネットワークを利用した．構造を 1026-513-256-128-64-32-4 とし，発話区間

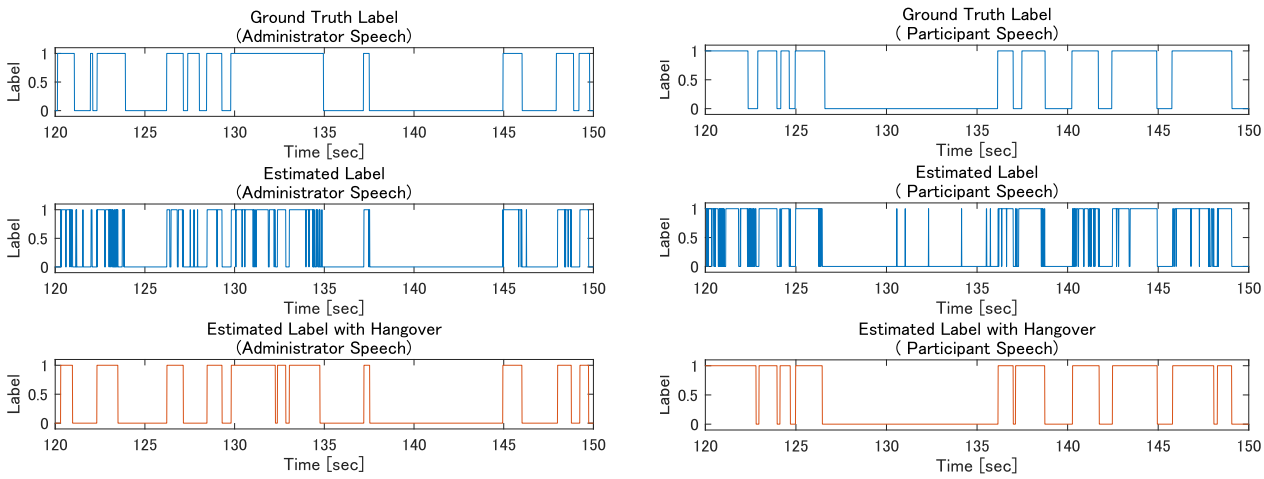

図 2 実施者（左）と被験者（右）の発話正解ラベル，推定ラベル（Hangover 有/無）の比較
Fig. 2 Comparison of the ground truth and estimation of VAD labels w / w.o. Hangover for an administrator (left) and a participant (right)

の推定を Chainer [10] を用いて実装した．エポックを 100 回，バッチサイズを 100 とした．活性化関数には ReLU 関数を，損失関数には交差エントロピー誤差を使用した．

発話は開始されるとしばらく継続するという仮定に基づき，検出された発話区間をひとまとまりにする処理 (Hangover 処理 [11][12]) を後処理として行った．具体的には，発話が終了してから 9 フレーム（約 0.1 秒）以内に次の発話が開始すると推定した場合，その間の非発話区間を発話区間とし，また，発話区間と推定した区間が 14 フレーム（約 0.15 秒）以下の場合にその範囲を非発話区間とする処理を行った．

3.2 実験結果と考察

発話／非発話の状態の推定結果を正解ラベルとフレーム単位で照合した．具体的には，ラベルが一致しているフレーム数の総和を全フレーム数で割り，正答率を求めた．推定ラベルの正答率は 84.1 %，Hangover 処理を行った推定ラベルの正答率は 88.5 %であった．また，発話区間検出のベースラインとして音声信号パワーの閾値処理に基づく方法を用い，最も正答率の高い閾値による結果と提案手法との比較を行った．ベースラインによる推定ラベルの正答率は 64.1 %で，Hangover 処理を行うと 66.5 %であった．Hangover 処理の有無に関わらず提案手法の正答率が高く，これは提案手法の有効性を示しているといえる．

図 2 に実施者と被験者の正解ラベル，推定ラベル（Hangover 有/無）を比較した結果を示す．左図が実施者の発話区間を推定した結果，右図が被験者の発話区間を推定した結果の例である．検出された推定ラベルは，おおよそ正解ラベルと合致しているが，断片的である．Hangover 処理を導入し，検出された発話区間をひとまとめにし，誤って短い範囲で局所的に発話と推定した区間を非発話区間とすることで，正答率が向上したことを示している．

図 3，**図 4** にそれぞれ提案手法，ベースラインによる推定の混同行列を示す．ただし，縦軸は正解ラベル，横軸が推定ラベルであり，両図とも Hangover 処理を行った結果である．各要素は非発話区間 (No Speech)，実施者と被験者の単一発話区間 (Only Administrator, Only Participant)，二人の同時発話区間 (Both) におけるそれぞれの正解，推定ラベルのフレーム総数を示している．また，混同行列の右にクラス毎の再現率，下にクラス毎の適合率をまとめて表示している．図 3 より，対角成分の総数が多く，多くの場合，正しく推定されたことを示している．また，図 3 と図 4 を比較すると，ベースラインによる推定では，実施者と被験者の単一発話区間を非発話区間，または二人の同時発話区間と誤推定していることを表す要素の数が多いのに対し，提案手法では，その数が抑えられている．このことから，非発話と誤推定されるほど小さい声量である，対話相手の音声の声量が大きく，単一録音を用いた場合には非発話を発話と誤推定するなど，ベースラインによる推定では正しく推定することが難しい状況でも，提案手法では実施者と被験者それぞれの単一発話区間と正しく推定することができることがわかる．

4. おわりに

本論文では，話者の襟元に装着した分散マイクで録音したステレオ対話音声データから，STFT により短時間スペクトルを抽出し，DNN を用いて，実施者と被験者の発話区間を検出する手法を提案した．客観評価実験を行い，音声信号のパワーに基づいたベースラインよりも正答率が高いことを示し，DNN を用いることにより，発話区間検出

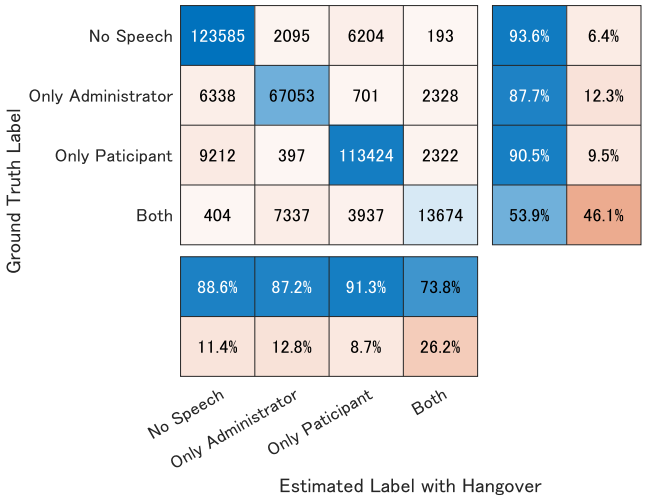

図 3 提案手法（Hangover 処理有り）による推定結果の混同行列
Fig. 3 Confusion matrix of VAD results estimated by the proposed method with Hangover

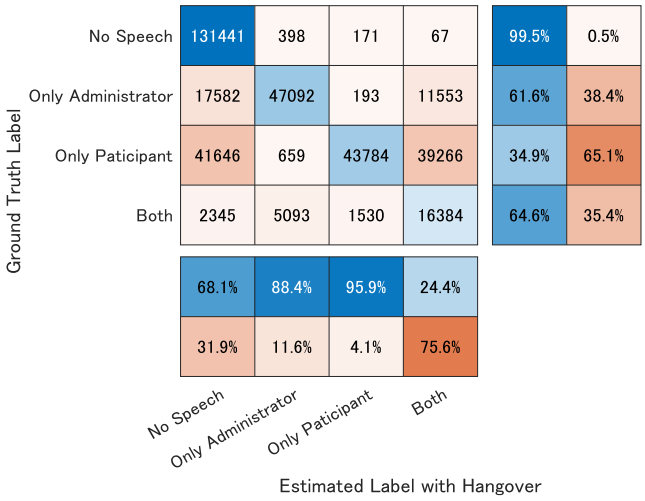

図 4 ベースライン（Hangover 処理有り）による推定結果の混同行列
Fig. 4 Confusion matrix of VAD results estimated by the baseline method with Hangover

法の性能が向上したことを示した．今後は，本研究を応用し，ASD の診断支援を行うため，被験者の音声分析，特に対話特徴量の自動抽出の研究を進めていく．

謝辞 本研究の一部は，JSPS 科研費基盤研究 A (課題番号 JP20H00613)，脳科学研究戦略推進プログラム (日本医療研究開発機構: JP18dm0107134) の助成を受けて行われたものである．

参考文献

[1] 越智 景子, 小野 順貴, 大須賀 智子, 大和田 啓峰, 児島 正樹, 黒田 美保, 山末 英典, 嵯峨山 茂樹, “自閉スペクトラム症者と定型発達者の音声特徴による識別の検討,” 日本音響学会秋季研究発表会講演論文集, pp. 631–632, 2016.

[2] 近藤 綾子, 出口 利定, “自閉スペクトラム障害児の発話におけるプロソディの特徴,” 日本聴覚言語障害学会, vol. 42, no. 1, pp. 23–30, 2013.

[3] C. Lord, M. Rutter, S. Goode, J. Heemsbergen, H. Jordan, L. Mawhood, and E. Schopler, “Austism diagnostic observation schedule: A standardized observation of communicative and social behavior,” Journal of Autism and Developmental Disorders, vol. 19, no. 2, pp. 185–212, 1989.

[4] 越智 景子, 小野 順貴, 大和田 啓峰, 黒田 美保, 山末 英典, 大須賀 智子, 嵯峨山 茂樹, “自閉スペクトラム症の ADOS スコアと対話音声の相関分析の検討,” 日本音響学会秋季研究発表会講演論文集, pp. 489–490, 2015.

[5] K. Ochi, N. Ono, K. Owada, M. Kojima, M. Kuroda, S. Sagayama, and H. Yamasue, “Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder,” PLOS ONE, vol. 14, no. 12, 2019: e0225377.

[6] 石塚 健太郎, 藤本 雅清, 中谷 智広, “音声区間検出技術の最近の研究動向,” 日本音響学会誌, vol. 65, no. 10, pp. 537–543, 2009.

[7] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints of isolated utterances,” The Bell System Technical Journal, vol. 54, pp. 297–315, 1975.

[8] J. Allen, “Short term spectral analysis, synthesis, and modification by discrete Fourier transform,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 25, no. 3, pp. 235–238, 1977.

[9] 小野 順貴, “短時間フーリエ変換の基礎と応用,” 日本音響学会誌, vol. 72, no. 12, pp. 764–769, 2016.

[10] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton, “Chainer: a Next-Generation Open Source Framework for Deep Learning,” Proc. NIPS, 6 pages, 2015.

[11] “Speech processing transmission and quality aspects (STQ) advanced distributed speech recognition (ADSR); frontend feature extraction algorithm; compression algorithms,” ETSI ES 202 050 v. 1.1.4, Nov. 2006.

[12] D. Vlaj, M. Kos, and Z. Kačič, “Quick and efficient definition of hangbefore and hangover criteria for voice activity detection,” Proc. IWSSIP, 4 pages, 2016.