2D-05

発話ターンレベルでの音響同調は精神科面接中の信頼を予測する

横谷　謙次†　　高木　源‡　若島　孔文‡

徳島大学大学院社会産業理工学研究部†　　東北大学大学院教育学研究科‡

1. 問題と目的

音声対話では、話し手のコミュニケーション行動は、対談者の行動に似るようになる（Giles、2016）。この現象は、コンピューター科学では同調(Levitan et al., 2015; Reichel et al., 2018; Weise et al., 2019)および心理学では同期(Ramseyer & Tschacher, 2014; Yokotani et al., 2019)と呼ばれる。話者の同調は、効率的なコミュニケーションを維持し、対話者との信頼を築くために不可欠である（Shepard、2001）。特に、信頼はコンピューター科学者と心理学者の両方にとって重要な概念である。ロボットと人間の信頼の前提条件は、両者の同調（Kopp、2010; Sadoughi et al。、2017）である。また、信頼を構築することで、さまざまな社会環境でのロボットの配置（Lala et al。、2017; Yokotani et al。、2018）および最も効果的な治療の提供（Nasir et al。、2015; Rusiewicz et al。、2014 ）が可能になる。実際、コンピューター科学者と心理学者の間のいくつかの共同研究により、同調と信頼の関係が明らかにされてきた（Kawahara et al。、2015; Xiao et al。、2015）。しかし、音響同調の単位レベルが十分に比較されなかったため（De Looze et al。、2014）、音響同調の最適な単位レベルは不明のままである。

本研究では、休止間単位(IPU)と発話単位(LUU)での音響同調を比較し、2 つの単位のどちらが精神科面接中の信頼をより良く予測するのかを明らかにする(Yokotani et al., 2018, 2019)。

2. 音声対話の分割

2.1.コーパス

コーパスは、日本での構造化臨床インタビューデータ（Yokotani et al., 2018、2019）を使用した。 データには、精神障害の診断および統計マニュアル、第 4 版、テキスト改訂 Axis I 障害、非患者版（First et al。、1997）の日本語版（First et al。、2010）の構造化臨床インタビューが含まれている。

音声を明確に分離するため（Ditter & Gerkmann、2019）、異なる性別のペアのみを使用した。 最終的に 1 名の男性面接者による 29 人の女性調査協力者のインタビューデータが使用された。インタビューの長さは個々で異なるため、最初の 903 秒の記録データを利用した。記録されたデータの最初の 3 秒はノイズプロファイルに使用され、最後の 900 秒は音声対話として分析された（Ramseyer＆Tschacher、2014; Yokotani et al。、2019）。

また、コーパスには、信頼に関する 5 段階評定の質問紙データも含まれていた（柿井、1997）。この質問紙が分析に使用された。

更に、調査協力者は、音声対話システムの設定を介して 16,000 ヘルツで自分の声を録音した（Yokotani et al、2018）。録音データには調査協力者の声のみが含まれていたため、これらは協力者の音源として参照し得る。セラピストも、日本の典型的な音を含む 2 つのひらがな表を読み、16,000 ヘルツで録音した。この読み上げの声もセラピスト自身の声の音源として参照し得る。

2.2. IPU 分割

ノイズプロファイルに従って音声対話からノイズが除去され、その後、音声が正規化された。データは 100 ミリ秒より長い無音で分割された。この分割を IPU とし、その集まりを IPUs とした（Den et al。、2010; Koiso&Den、2000）。

2.3. 記録されたデータをセラピストと調査協力者の声に分離

IPUs にはセラピストと調査協力者の両方の声が含まれているため、先行研究（Seungwon、2019/2019）と同じニューラルネットワークモデル（Wang et al。、2018）を用いて話者分離を行った。 訓練データは LibriSpeech データセット（Panayotov et al。、2015）で、460 時間のクリーンな音声である。 モデルは混合音声データを 644,000 回にわたって音声分離を学習し、7.69 の音声妨害比（SDR）を示した。 このモデルを用いて、調査協力者とセラピストの個々の声の音源を参照させ、先ほどの IPUs をセラピストの音声 IPUs と調査協力者の音声 IPUs に分離した。 調査協力者の声を含む IPUs は調査協力者 IPUs と呼び、セラピストの声を含む IPUs はセラピスト IPUs と呼ぶ。

2.3. ピッチ、ビート、RMS、クロマ特徴、及び MFCC の推定

ピッチ（f0）を推定するために、デフォルト設定の Parselmouth を使用した。ビート、RMS、クロマ機能、および MFCC の推定のために、デフォルト設定の LibROSA も使用した。どちらも音声分析用の python ライブラリである。

2.4. モーラ推定

セラピストと調査協力者の IPU のモーラ数を次のように推定した。まず、MFCC のパワーとそのパワーのデルタを計算した。 次に、そのパワーとデルタを乗算して、単一の曲線を作成した。曲線を滑らかにするために、ウィンドウサイズを 4 に設定したガウスフィルターでフィルタリングした。

$$F = G(p \circ \Delta p, 4)$$

なお、式の○はアダマール積で、 G（ , 4）は、ウィンドウサイズ 4 のガウスフィルター設定を示している。

また、極大値を見つけるために、平滑化された曲線の値を前回、現在、および次回の間で比較した。 現在のスコアが前のスコアと次のスコアよりも大きい場合、現在の時刻にはモーラがあると判定した。

2.5. 発話ターン分割

話者のターンを決定するために、セラピストと調査協力者の IPUs の間でモーラ数を比較した。現在の時間（t）に 1 人のスピーカー（例セラピスト）のモーラ数が他のスピーカーのモーラ数よりも大きい場合、現在の IPU はスピーカー（セラピスト）のターンと見なされた。 モーラ数が現在の時間（t）で等しい場合、前の時間（t-1）と次の時間（t + 1）のモーラ数が追加され、比較された。前 5 回（t-5）から後 5 回（t + 5）までのモ-ラ数を合計した後でもモ-ラ数が等しい場合、現在の IPU（t）ではターンなしと割り当てた。 この計算により、IPUs はすべて、セラピストのターン、調査協力者のターン、またはターンなしのいずれかに紐づけられた。前後の IPUs でターンが等しい場合は結合され、異なる場合は分割された。この分割を LUU（発話ターン）とし、その集まりを LUUs とした。

3. 音響同調の測定

ピッチ、ビート、RMS、およびモーラの同調を評価するために、Jensen- Shannon Divergence（JSD）を使用して 2 つの音声間の非同調を測定した。

クロマ特徴と 12 個の MFCC の同調を評価するために、2 つの音声間の非同調の尺度として Wasserstein Distances（WD）を使用した。

また、真の同調[純正ペア間の同調]が、疑似同調[疑似ペア間の同調]と異なるかどうかも検査した（Gatewood&Rosenwein、1981）。29 のインタビューデータのうち 17,022 個の調査協力者 IPUs が 17,022 個のセラピスト IPUs とランダムにペアリングされた。同様に、4,005 個の調査協力者 LUUs は、4,005 個のセラピスト LUUs とランダムに組み合わせられた。これらのペアはそれぞれ IPUs と LUUs の疑似ペアとした。

4. 結果

4.1. 純正ペアと擬似ペアの音響同調の比較

ピッチ、ビート、RMS、モーラ、クロマ特徴及び 12 個の MFCC の IPU レベルでの JSD と WD は、純正ペアの方が疑似ペアよりも有意に少なかった。同様に、LUU レベルでのこれらの特徴に関する JSD と WD は純正ペアの方が疑似ペアよりも有意に少なかった。つまり、純正ペアの方が音響同調は有意に高いと言える。

4.2.調査協力者の音響的特徴、セラピストの音響的特徴、および IPU レベルでの調査協力者間の音響同調によるセラピストへの信頼の予測

RMS と mora の JSD は、個々の音響特性と IPUs の数が統制された後でも、信頼を負に予測した。同様に、MFCC の WD は信頼を負に予測した。つまり、調査協力者とセラピスト間の RMS、モーラ、MFCC の同調が高いほど、セラピストへの信頼も高かった。

4.3.調査協力者の音響的特徴、セラピストの音響的特徴、および LUU（発話ターン）レベルでの調査協力者間の音響同調によるセラピストへの信頼の予測

ピッチ、ビート、RMS、およびモーラの JSD は、セラピストへの信頼を負に予測しました。つまり、ピッチ、ビート、RMS、およびモーラの同調が高いほど、セラピストへの信頼も高かった。さらに、RMS と mora の JSD を含む回帰分析は、それぞれ信頼を 66%と 48%で説明した。また、LUU（発話ターン）レベルでの回帰分析の説明率は、IPU レベル（5/6）よりも高かった。

5. 考察

本研究では、ピッチ、ビート、RMS、モーラ、MFCC などの音響特性の LUU（発話ターン）レベルでの同調が、IPU レベルの同調よりも、セラピストへの信頼をより良く予測する因子であることを明らかにした。

6. 謝辞

本研究に関して、コメントを頂戴した佐藤裕教授（徳島大学）に感謝します。本研究の一部は、Speech Communication に投稿中の論文です。また、本研究は、科学技術融合財団（平成 27 年 10 月）および科研費（18K02141）からの助成金を受けています。

“Acoustic entrainment at the Turn Unit level predicts trust during psychiatric interviews”

Yokotani, Kenji†,Takagi, Gen‡,Wakashima Kobun‡,
†Graduate School of Technology, Industrial and Social Sciences, Tokushima University
‡Graduate School of Education, Tohoku University