単一話者特徴を利用したターンテイキング予測モデルの検証

大西　一誉† 　田中　宏季† 　中村　哲†

† 奈良先端科学技術大学院大学　〒 630–0192 奈良県生駒市高山町 8916 番地の 5
E-mail: †{onishi.kazuyo.oi5,s-nakamura,hiroki-tan}@is.naist.jp

あらまし　二者間会話の発話予測は、人間とバーチャルエージェントのターンテイキングを自然にするため重要である。多くのモデルは二者の特徴を必要とするが、バーチャルエージェントに組み込む際には、バーチャルエージェント自身の特徴量を取得することが困難といった問題が生じる。本研究では、単一話者特徴のみを用いた Transformer に基づく音声活動予測モデルを提案し、3 つのコーパスで性能の検証を行った。検証では、二者特徴を用いたモデルとの比較、言語や個人差の特性獲得、非言語特徴の効果を調査した。その結果、単一話者モデルの性能はやや低下するものの、言語差によるターンテイキング特性の違いを取り込むことや非言語特徴を追加することで性能が向上することが確認された。

キーワード　話者交替, 単一話者, 音声活動予測

Turn-Taking Prediction Model Using Single Speaker Features

Kazuyo ONISHI†, Hiroki TANAKA†, and Satochi NAKAMURA†

† Nara Institute of Science and Technology　Takayama–cho 8916–5, Ikoma–shi, Nara, 630–0192 Japan
E-mail: †{onishi.kazuyo.oi5,s-nakamura,hiroki-tan}@is.naist.jp

Abstract　Prediction of utterances in two-party conversations is important to make turn-taking between humans and virtual agents natural. Many models require double-speaker features, but when incorporating them into virtual agents, problems arise, such as the difficulty of obtaining the virtual agent's own features. In this study, we proposed a Transformer-based voice activity prediction model using only single-speaker features and tested its performance on three corpora. In the validation, we compared the model with a model using double-speaker features, acquired features by language and individual differences, and investigated the effect of non-verbal features. The results confirmed that although the performance of the single-speaker model was somewhat degraded, performance could be improved by incorporating differences in turn-taking characteristics due to language differences and by adding non-verbal features.

Key words　Turn-taking, single-speaker, voice activity prediction

1. はじめに

二者間対話において話し手と聞き手は内容や場面を共有し、交互に発話を行いながらコミュニケーションを進める。ターンテイキングとは、コミュニケーションの円滑化と相互理解の確保のために、参加者が交互に話す役割を切り替える行為である [1]。人間は平均 200 ミリ秒という高速で話し手と聞き手の役割を流暢に切り替え、このような役割分担を効率的に行う [2]。このようなターンシフトの中には、話し手の発話終了に被さるように聞き手が発話を始める場合がある。このようなオーバーラップを生じるターンシフトは会話中にしばしば出現する。さらに、対話のもう 1 つの重要な要素としてバックチャネルがある。バックチャネルは聞き手が話し手に対して送る短いメッセージのことであり、人間はバックチャネルの適切なタイミングを認識し、聞き手の話をよく理解していると示す [3, 4]。

しかし、バーチャルエージェントなどの音声対話システムには、まだ充分なそれらの機能が備わっていない [5]。現在の多くの音声対話システムは、頻繁にユーザの話を遮り、反応が遅れる傾向があり、迅速なフィードバックの欠如が会話の自然な流れを阻害する可能性が指摘されている。ターンテイキング性能の低下はコミュニケーションを妨げ、ユーザに意図しないメッセージを伝える可能性がある [6]。ユーザは対話システムを礼儀正しく、人格的であると認識することでより利用満足度が高まり、適切なターンテイキング性能はこのような満足度を向上させるため不可欠である。多くの対話システムは、Automatic Speech Recognition (ASR) を用いてユーザー発話の終了を検知して、沈黙時間の閾値によって発話を決定している。しきい値を小さく設定すると、応答時間が短縮されるが、ユーザの発

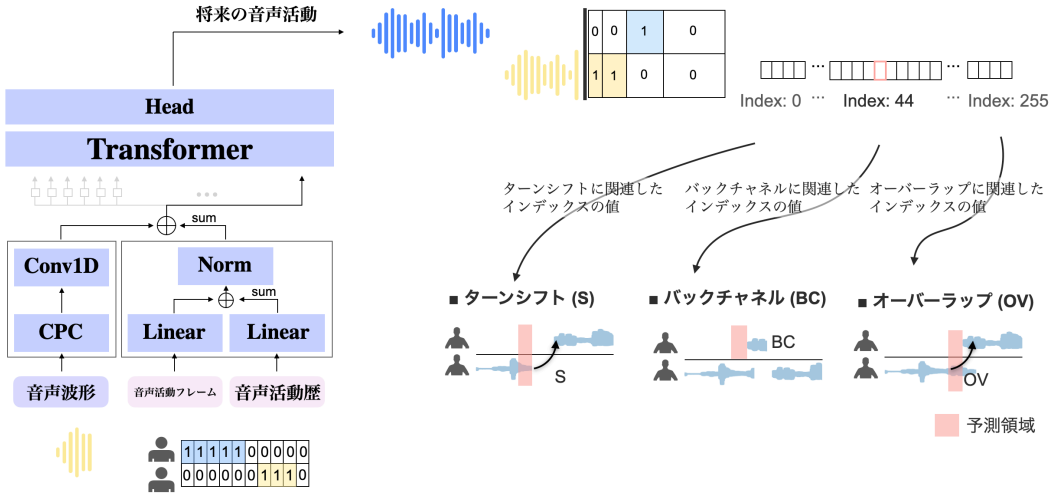

図 1　単一話者特徴を利用した音声活動予測からターンテイクイベントの決定までの流れ

話の中断を引き起こすリスクを伴う。一方で、しきい値を大きく設定すると、ユーザはシステムが応答していないと誤解する可能性がある。このように、現在の ASR を用いた閾値による決定はシステム的な限界があり、応答速度が制限され、様々な状況に適応できないといった課題に直面し、より高速にターンを予測するモデルが必要である [7,8]。

そこで、機械学習を用いてターンテイキングを予測し、人間のような自然な応答を目指すモデルの研究に取り組まれている [9–12]。中でも Skantze の提案した音声活動予測モデルは、ターンテイキングイベントを直接予測することなく、音声活動を予測しその出力を用いることでターンを決定する [13]。これまでターンシフトやオーバーラップ、バックチャネルといった異なるターンテイキングイベントを異なるモデルで解決していたのに対して、音声活動を予測することで全てのイベントを統括的に決定する手法を提案した。私たちはこれまでに、音声活動予測モデルが視線方向、Action Unit、頭部運動、関節点の 4 つの非言語特徴をどのように取り込むことができるかを検証し、ターンシフトの F1 スコアで 0.748 (+0.023)、バックチャネルで 0.687 (+0.018) の有意な改善が得られることを明らかにした [14]。

これまでのターンテイキングモデルは、二者の特徴量を入力に用いるが、バーチャルエージェントなどの音声対話システム自身の特徴量を得られないといった問題が生じる。仮に音声対話システム自身の音声を入力した場合、人間-人間対話のコーパスで評価した場合とは異なる性能になることが予想され十分な評価とは言えない。そこで私たちは、単一話者特徴を用いた Transformer に基づく音声活動予測モデルを提案する。本研究では、私たちの提案する単一話者特徴モデルが十分な機能を有することを確認するために以下の 3 点について検証を行った。

(1) 二者の特徴を用いたモデルと比較し、ターンテイキング性能にどのような影響を及ぼすか。

(2) 国や職業ごとにファインチューニングすることで異なる特性を獲得することは可能か。

(3) 単一話者モデルに非言語特徴を組み込むことで性能向上に寄与するか。

2. 単一話者特徴を利用したターンテイキングモデル

ここでは、単一話者特徴を利用した音声活動予測モデルによるターンテイキングについて説明する。図 1 に単一話者特徴を利用した音声活動予測からターンテイクイベントの決定までの流れを示す。このモデルは、特定のターンテイキングイベントに学習されたものではなく、将来の音声活動の予測に基づいてターンテイキングイベントを決定する。先行研究に基づいて音声波形と現在の音声活動を入力するが [15]。これまで、音声波形は二者の音声を用いていたが、ここでは単一話者の音声波形とする。Ekstedt らによって提案された離散的な出力窓を採用し、0.2 秒、0.4 秒、0.6 秒、0.8 秒の 4 つの不等間隔領域に分割された 8 つのビンを設定する。各ビンがアクティブの場合を 1、非アクティブの場合を 0 として、サイズ (2, 4) の離散的なワンホット表現を作成する。ベクトルをインデックスに対応付け、分類問題として扱い学習を行う。将来に渡る音声活動の出力を用いて、ゼロショットで検出されたターンシフト、バックチャネル、オーバーラップにマッピングを行う。ターンシフト、オーバーラップは、1 人の話者のみがアクティブになる領域を定義するプリオフセット （1 秒） とポストオンセット （1 秒） という 2 つのパラメータも定義し検出する。バックチャネルは、発話が 1 秒未満かつその発話を囲むその話者による発話がない領域 （プリサイレンス (1 秒) とポストサイレンス (2 秒)）で定義される。これらは、先行研究により公開されたソースコードを用いて実装を行った [16]。本研究では、以下の 3 つのターンテイキングイベント定義し評価を行う。

ターンシフト (S): 小さな沈黙を伴い、違う話者に発話が移り変わること

バックチャネル (BC): 相手が話しているときに、聞き手が短い言葉を返すこと

オーバーラップ (OV): 発言が一部重なるようにして、違う

表 1　学習データの分割

コーパス		Train [h]	Val. [h]	Test [h]
Sw		217.28	24.02	11.07
NoXi	イギリス	8.79	0.85	1.78
	ドイツ	2.77	1.38	1.49
	フランス	5.30	1.12	2.01
	全て	16.86	3.35	5.28
EALD	高齢者 - 介護士	3.77	1.39	1.72
	高齢者 - 臨床心理士	2.90	1.31	1.39
	高齢者 - 学生	3.68	1.37	1.69
	全て	10.35	4.06	4.79

表 2　テストデータに含まれるターンテイキングイベント数

コーパス		S	BC	OV
Sw		1054	2045	330
NoXi	イギリス	77	408	37
	ドイツ	111	254	31
	フランス	173	352	70
	全て	361	1014	138
EALD	高齢者 - 介護士	87	49	106
	高齢者 - 臨床心理士	82	30	107
	高齢者 - 学生	98	75	81
	全て	267	154	294

話者に発話が移り変わること

3. 性能評価

ここでは、1 章で設定した 3 つの研究課題について実験・評価を行った。

3.1 コーパス

モデル訓練に使用する異なる 3 つのコーパスについて説明する。

(1) Switchboard Corpus (Sw) [17] 543 人の参加者、2,400 の音声ファイルから構成され、特定のトピックについて電話越しによるフリーディスカッションが収録されている。約 259 時間と非常に多くの時間が収録されており、言語は英語のみである。活発な議論が収録されており、音声活動予測を行う観点からは難しい内容である。

(2) NoXi Database (NoXi) [18] 87 人の参加者、84 の音声と映像ファイルから構成され、特定のトピックについて専門家と初学者によるスクリーン越しの対話が収録されている。約 25 時間と時間は多くないものの、映像が収録されているため非言語特徴の利用が可能である。さらに、イギリス、ドイツ、フランスの 3 つの国で英語、ドイツ語、フランス語などが収録されている。特定のトピックについて議論している点では Sw と同じであるが、専門家と初学者が対話しているため、必然的に専門家の発話比率が高くなり、音声活動予測の観点からは Switchboard Corpus よりも容易な内容と予想される。

(3) Elderly Attentive Listening Dialogue (EALD) [19] 39 人の参加者、60 の音声と映像ファイルから構成され、介護士、臨床心理士、学生が、高齢者に対してさまざまな情報について共有している様子が収録されている。全て日本で収録されており、3 つのコーパスで唯一の日本語である。傾聴対話であるため、高齢者の発話比率が少なく、音声活動予測の観点からは最も容易な内容であると言える。

それぞれのコーパスは、表 1 に示すように Train、Validation、Test データを分割する。テストデータ内に含まれるターンシフト、バックチャネル、オーバーラップの数を表 2 に示す。Test データの時間によってばらつきはあるものの十分なサンプル数を得られた。

3.2 二者特徴モデルとの比較

ここでは、提案した単一話者特徴モデルと二者特徴を用いたモデルとを比較する。

表 3　単一話者特徴モデルと二者特徴モデルのターンテイキング性能の比較

コーパス	イベント	二者特徴	単一話者特徴	差分
Sw	S	0.712	0.663	-0.049
	BC	0.714	0.683	-0.032
	OV	0.711	0.694	-0.017
NoXi	S	0.731	0.702	-0.029
	BC	0.685	0.663	-0.022
	OV	0.604	0.591	-0.013
EALD	S	0.694	0.647	-0.047
	BC	0.568	0.555	-0.013
	OV	0.466	0.458	-0.008

3.2.1 実験方法

1 秒の入力窓を 0.5 秒ずつスライドしながら入力する。Transformer は隠れ層サイズ 256、2 層、4 ヘッド、ドロップアウト 0.4 で構成した。早期停止基準を 3 エポックとし、最適化 AdamW、学習率 3.63e-4、バッチサイズ 128 で学習を行った。seed を 0 から 9 まで変化させて学習した平均を用い、ターンシフト、バックチャネル、オーバーラップの観点から評価を行う。単一話者モデルは片方の話者を入力とするが、Sw は参加者 2、NoXi は初学者、EALD は介護者、心理学者、学生を用いてテストを行った。

3.2.2 実験結果

表 3 に単一話者特徴モデルと二者特徴モデルのターンテイキング性能の比較を示す。それぞれのコーパスに対する、ターンテイキングイベントの F1 スコアを示している。コーパスによりばらつきはあるものの、ターンシフトは 0.029〜0.049 ポイント減少、バックチャネルは 0.013〜0.032 ポイント減少、オーバーラップは 0.008〜0.017 ポイントの減少となった。

3.2.3 考察

単一話者特徴モデルと二者特徴モデルの性能差を見ると、ターンシフト予測では大きく低下する場合があり、最大 0.049 ポイントの低下となった。これは、ターンシフトが強い相互作用によるものであることを示唆しており、特徴量を単一話者にすることで予測が困難な場合が生じる。しかし、単一話者モデルのターンシフト予測性能を個別に確認すると、7 割弱ほどのスコアであり、先行研究と比較しても決して低いスコアではない [13,15]。一方で、オーバーラップとバックチャネルの予測で

表 4　国および職業でのファインチューニング前後の性能変化

	NoXi			EALD		
国	イギリス	ドイツ	フランス	日本	日本	日本
入力	初学者	初学者	初学者	介護士	臨床心理士	学生
S	+0.170	+0.156	+0.163	+0.106	+0.035	+0.083
BC	+0.094	+0.090	+0.143	+0.046	-0.086	+0.080
OV	+0.127	+0.080	+0.046	-0.044	-0.105	-0.075

はターンシフトほどの低下は見られなかった。ターンシフトが相互作用によるものであるのに対して、オーバーラップとバックチャネルは片方のユーザのみで判断できる要素が大きいことが示唆された。

3.3　国や職業ごとの特性の取り込み

ここでは、提案した単一話者特徴モデルが国や職業ごとにファインチューニングを行うことで、言語や個人差による特徴を取り込むことができるかを調査する。

3.3.1　実 験 方 法

Switchboard Corpus で学習した単一話者モデルを用いて、NoXi の国別、EALD の職業別に表 1 で分割したデータを用いてファインチューニングを行う。ファインチューニング前とファインチューニング後を比較し、どの程度 F1 スコアが上昇するかを確認する。国別にファインチューニングを行うことで、言語によるターンテイキング特性の差を取り込む。職業別にファインチューニングを行うことで、個人差によるターンテイキング特性の差を取り込む。

3.3.2　実 験 結 果

表 4 にファインチューニング後からファインチューニング前の F1 スコアの差分をとった結果を示す。国別でファインチューニングを実施すると、すべての項目で大幅にスコアが向上した。一方で、職業別でファインチューニングをした場合は、ターンシフトのみに共通して有意な向上が見られた。バックチャネルは、介護者と学生に有意な向上が生じ、特に学生が高い値となった。オーバーラップには有意な上昇を得ることはできなかった。

3.3.3　考　　察

国別でファインチューニングを行うことで、言語による差を得ることができる。言語間でターンテイキングの性質に差があることは先行研究で示唆されており、単一話者モデルが上手くそれらの特性を取り込めた可能性がある [1]。音声対話システムに適用する場合にも、言語によるファインチューニングが有効である可能性がある。職業別にファインチューニングした場合、共通して有意差が得られたのはターンシフトのみであった。介護士と臨床心理士は高齢者に対する対話に長けており、特徴的な点が生じたと考えられる。一方で、バックチャネルには介護士と学生のみに有意差が生じ、特に学生が高い結果となった。表 2 から確認できるように、学生は介護士と臨床心理士と比較してバックチャネルの数が多く、特徴的な点をファインチューニングにより捉えることができた可能性がある。しかし、原因の断定にはさらなる検証が必要である。

3.4　非言語特徴の取り込み

ここでは、提案した単一話者特徴モデルが非言語特徴を取り込むことができるかを検証する。

表 5　非言語特徴を追加した場合の比較

イベント	NoXi		P 値	効果量
	ユニモーダル	マルチモーダル		
S	0.702	0.710	0.382	0.376
BC	0.663	0.679	0.004	1.344
OV	0.591	0.639	0.001	2.213

3.4.1　実 験 方 法

先行研究 [14] で提案した視線方向、Action Unit、頭部運動、関節点の 4 つの非言語特徴を採用し、音声特徴のみのモデルと比較してどのように性能が向上するかを確認する。NoXi の映像データを用いて、OpenPose [20]、OpenFace [21] を用いて特徴量を抽出した。非言語特徴においても、音声波形と同様に単一話者の特徴量を入力とした。

3.4.2　実 験 結 果

表 5 に音声特徴のみの場合と非言語特徴を追加した場合の F1 スコアを比較する。その結果、ターンシフトには有意差が生じかなったものの、バックチャネルおよびオーバーラップに有意差が生じ性能向上につながった。

3.4.3　考　　察

ターンシフトは相互作用によるところが大きく、単一話者の非言語特徴を追加しても有意差が生じなかった可能性がある。一方で、バックチャネルやオーバーラップは、単一話者からの動きである程度予測できる部分があると推察される。これは、直感的にも理解ができる点であり、バックチャネルやオーバーラップといった事象は、双方の特徴量に基づくものではなく、単一話者からの特徴に左右される部分が大きいと考えられる。

4.　お わ り に

本研究では、単一話者特徴を用いた Transformer に基づく音声活動予測モデルを提案した。二者の特徴を用いたモデルと比較し、相互作用の大きなターンシフトでは性能がやや低下するものの、バックチャネルやオーバーラップではその低下は小さく抑えられることを明らかにした。その上で、私たちは国や職業ごとにファインチューニングすることで言語や個人差の特性を獲得し性能向上を目指した。国ごとにファインチューニングすることで言語差による特性を獲得し、大幅に精度が向上することを示した。職業ごとにファインチューニングすることでターンシフトなどの部分的に効果があることを示した。さらに、非言語特徴を組み込みモデルにどのような影響を及ぼすか調査した。その結果、バックチャネルとオーバーラップで有意差が得られた。本研究により、音声対話システムのターンテイキング性能の更なる発展が見込まれる。

5.　謝　　辞

本研究は JST-CREST（グラント番号: JPMJCR19M5）の支援によって行われた。

文　　献

[1] G. Skantze, “Turn-taking in conversational systems and

human-robot interaction: a review," Computer Speech & Language, vol.67, p.101178, 2021.

[2] S.C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," Frontiers in psychology, vol.6, p.731, 2015.

[3] B. Inden, Z. Malisz, P. Wagner, and I. Wachsmuth, "Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent," Proceedings of the 15th ACM on International conference on multimodal interaction, pp.181–188, 2013.

[4] B.B. Türker, Z. Buçinca, E. Erzin, Y. Yemez, and T.M. Sezgin, "Analysis of engagement and user experience with a laughter responsive social robot.," Interspeech, pp.844–848, 2017.

[5] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, "Turn-taking prediction based on detection of transition relevance place.," INTERSPEECH, pp.4170–4174, 2019.

[6] T. Itoh, N. Kitaoka, and R. Nishimura, "Subjective experiments on influence of response timing in spoken dialogues," Interspeech, 2009.

[7] N.G. Ward, A.G. Rivera, K. Ward, and D.G. Novick, "Root causes of lost time and user stress in a simple dialog system," Interspeech, 2005.

[8] A. Raux, D. Bohus, B. Langner, A.W. Black, and M. Eskénazi, "Doing research on a deployed spoken dialogue system: one year of let's go! experience," Interspeech, 2006.

[9] M. Takeuchi, N. Kitaoka, and S. Nakagawa, "Generation of natural response timing using decision tree based on prosodic and linguistic information," Eighth European Conference on Speech Communication and Technology, 2003.

[10] S. Fujie, H. Katayama, J. Sakuma, and T. Kobayashi, "Timing generating networks: Neural network based precise turn-taking timing prediction in multiparty conversation," 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021, pp.3771–3775, International Speech Communication Association, 2021.

[11] J. Sakuma, S. Fujie, and T. Kobayashi, "Response Timing Estimation for Spoken Dialog System using Dialog Act Estimation," Proc. Interspeech 2022, pp.4486–4490, 2022.

[12] N.G. Ward, D. Aguirre, G. Cervantes, and O. Fuentes, "Turn-taking predictions across languages and genres using an lstm recurrent neural network," 2018 IEEE Spoken Language Technology Workshop (SLT), pp.831–837, IEEE, 2018.

[13] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks," Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, pp.220–230, 2017.

[14] K. Onishi, H. Tanaka, and S. Nakamura, "Multimodal voice activity prediction: Turn-taking events detection in expert-novice conversation," HAI '23: Proceedings of the 11th International Conference on Human-Agent Interaction, Conference in Human-Agent Interaction '23, 2023.

[15] E. Ekstedt and G. Skantze, "Voice activity projection: Self-supervised learning of turn-taking events," arXiv preprint arXiv:2205.09812, 2022.

[16] E. Ekstedt, "Vap: Voice activity projection," 2022. `https://github.com/ErikEkstedt/vap_turn_taking`.

[17] J.J. Godfrey, E.C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," Acoustics, speech, and signal processing, ieee international conference on, pp.517–520, IEEE Computer Society, 1992.

[18] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The noxi database: multimodal recordings of mediated novice-expert interactions," Proceedings of the 19th ACM International Conference on Multimodal Interaction, pp.350–359, 2017.

[19] K. Yoshino, H. Tanaka, K. Sugiyama, M. Kondo, and S. Nakamura, "Japanese dialogue corpus of information navigation and attentive listening annotated with extended iso-24617-2 dialogue act tags," Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[20] Z. Cao, G. Hidalgo, T. Simon, S.E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," IEEE transactions on pattern analysis and machine intelligence, vol.43, no.1, pp.172–186, 2021.

[21] T. Baltrusaitis, A. Zadeh, Y.C. Lim, and L.P. Morency, "Openface 2.0: Facial behavior analysis toolkit," 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pp.59–66, IEEE, 2018.