Voice Activity Projection とガンマ分布ハザード関数を用いて対話の間を考慮したターンシフトモデル

大西 一誉 [1,2] 大中 緋慧 [1,2] 吉野 幸一郎 [3,2,1]
[1] 奈良先端科学技術大学院大学
[2] 理化学研究所 ガーディアンロボットプロジェクト [3] 東京科学大学
{kazuyo.onishi.oi5,onaka.hien.oj5,koichiro}@naist.ac.jp

概要

音声対話システムにおいて，自然なタイミングでのターンテイキングは極めて重要な課題である．近年提案された Voice Activity Projection (VAP) モデルは高い予測精度を持つ一方で，具体的な発話開始タイミングを決定する行動決定の枠組みを欠いる．本研究では，VAP による将来の発話意図予測に，人間のターンシフト速度の分布をモデル化したガンマ分布のハザード関数を統合する手法を提案する．実験の結果，従来のベースラインと比較してターンシフト予測における F1 スコアが大幅に向上し，さらに，1 秒以内のタイミングでターンを決定できた割合は 38.55% から 60.62% へと改善した．

1 はじめに

対話におけるターンテイキングは，単に話者が交代する現象という側面を越えて，対話参与者が社会的・認知的・感情的な意図を共有するための動的な合意形成プロセスである [1, 2]．特に，ターンが移行する際のタイミングは，即応，同調，思考，あるいは拒否といった重要な伝達的意味を内包している [3, 4]．自然な対話システムを実現するためには，適切なタイミングでの発話開始が不可欠であるが，従来の多くの研究は**誰が次に話すか**という決定に焦点を当てており，具体的な発話開始タイミングの制御については議論が不十分であった．

近年，自己教師あり学習を用いた将来の音声活動予測モデルである Voice Activity Projection (VAP) [5] が提案され，次発話者予測に関して高い予測精度を達成している．VAP は将来の音声活動を確率分布として出力する優れたモデルであるが，本質的には将来の発話確率 $p = (speaker = i|t + \Delta)$ を示しているに過ぎない．そのため，現実の対話において今この瞬間にターンを開始すべきかという具体的な行動決定を下すための数理的な枠組みを欠いている．この課題により，標準的な VAP 出力を用いたターンシフト予測では，人間らしい間の取り方を考慮できず，実際に人間が行うターンテイキングと比較してタイミング誤差が大きくなる傾向がある．

そこで本研究では，VAP の発話予測能力を活かしつつ，人間らしい適切なタイミングでのターンシフトを実現する予測モデルを提案する．具体的には，VAP から得られる発話意図の強さに，人間のターンシフト速度を数理化したガンマ分布のハザード関数を統合し，時間的な自然さを考慮したターンタイミング調整指標を定義する．このスコアに基づく行動決定アルゴリズムを導入することで，沈黙や重なりを同一の枠組みで扱いながら，高精度なタイミングでのターンシフトを可能にする．さらに，このタイミングモデルでは，ターンタイミング調整指標を特定のドメインにパラメータ調整を行うことで，個別の対話状況に適応したターンテイキングが可能であることを示す．

2 VAP とその課題

2.1 VAP の概要

Voice Activity Projection (VAP) は，音声対話における将来の音声活動を自己教師あり学習を用いて予測するモデルである [5, 6] (図 1)．VAP のアーキテクチャは，対話の文脈を捉えるための CPC (Contrastive Predictive Coding) 特徴量 [7] と，自己・相互注意機構を備えた Transformer で構成される．このモデルは，将来 2 秒間にわたる音声活動を離散的な出力窓として定義し，各話者が将来のどの時点で発話しているかの確率分布を同時に出力する．VAP は，ターンテイキングそのものを直接予測するのではなく，

音声活動の投影を学習するため，下流タスクへの汎用性が高いという特徴を持つ．

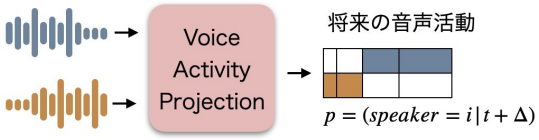

図 1: VAP による将来の音声活動の予測

2.2 既存のターンシフト決定手法

VAP の出力を用いて話者 B から話者 A へターンが移行するかを決定する従来の手法では，話者ごとの将来の発話確率の比率に基づく閾値判定が行われる [5]．具体的には，将来の予測窓において話者 A が話す確率が高い窓の合計と，話者 B が話す確率が高い窓の合計の比率を計算し，その値が特定の閾値を超えたタイミングをターンシフトとみなすアルゴリズムが一般的である．Ekstedt らは，話者 A が将来にかけて話す確率が高い bin を P_A，話者 B が継続して話す確率が高い bin を P_B として図 2 のように定義した．

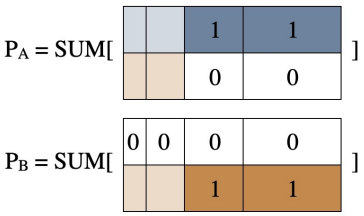

図 2: ターンシフト判定における P_A および P_B

ここで，色の薄い bin は 0 でも 1 でもどちらでも良いことを示している．これらの比率として，$P = P_A/P_B$ が閾値を超えたタイミングで話者 A へターンシフトをする (Positive) と予測したとみなした．これは，正解のタイミングよりも前の段階で予測ができていれば正解とみなす．言い換えれば，既存のターンシフト決定手法で利用されているのは，正解のタイミングよりも前の段階で予測ができていれば正解と見なすアルゴリズムである．

2.3 ターン決定における課題

VAP は高い予測精度を誇る一方で，ターンテイキングのタイミングを制御する観点からは以下の課題が存在する．

行動決定メカニズムの欠如: VAP の出力は将来の時点 $t + \Delta$ における発話確率分布 p を示しているに過ぎない．そのため，**今この瞬間にターンを開始すべきか**という具体的な行動決定を下す仕組みをモデル自体が持っていない．先行研究の評価指標（F1 スコア）では，正解の位置より前で検出されていれば正解とみなされている．つまり，この正解は実時間におけるターン取得のタイミング予測という観点で十分に評価・最適化されたものではないため，そのまま閾値判定結果を用いた場合，正解よりも大幅に早くターンを予測してしまう早合点が生じてしまう．この枠組みではターンテイキングのタイミングに含意される対話意図としての間を十分に表現することができない．

時間分解能とタイミング: ターンシフトの速度は 0.2 秒の差で伝達的意味が大きく変わるが，従来の P は 0.6 秒先以降の予測窓の出力を重視してターンシフト予測をしているため，直近の変化を十分に捉えられない．VAP は潜在的にターンシフトのタイミングを学習しており，それを引き出すにはより直近の bin を考慮する必要がある．

3 提案手法

前節で述べた 2 つの課題を解決するため，以下の二段階の提案を行う．第一に，短期的な発話意図を反映した新たなターンシフト予測指標 $P(t)$ の導入，第二に，**人間らしい間**を考慮した行動決定指標 $T(t)$ の構築である．

3.1 発話意図を考慮した指標 $P(t)$ の提案

まず，従来のターンシフト予測指標 P に対して，直近の発話意図を感度良く捉えるための改良を加える．従来，将来の予測窓における話者 A と話者 B の確率比 $P = P_A/P_B$ が用いられてきたが，本研究ではこれを以下のように修正し，提案手法の基礎とする．

$$P(t) = \frac{P_A(t)}{P_B(t)} \times E_i(t) \tag{1}$$

ここで，$E_i(t)$ は VAP の出力から導出される特定話者 i の発話意図であり，図 3 ように定義される．

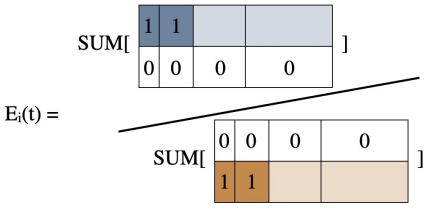

図 3: 発話意図 $E_i(t)$ の定義

$E_i(t)$ は，VAP の出力のうち直近の予測窓（例：

0 ~ 600ms）における話者 i の音声活動確率の累積値である．

従来の指標 P にこの $E_i(t)$ を乗算する意図は，遠い将来（0.6 秒以降）の予測確率が閾値を超えていても，直近の窓で発話の兆候が見られない場合にはターンシフトの判定を抑制するためである．これにより，モデルは将来的に交代が起こる可能性だけでなく，今まさに交代が始まろうとしているかという短期的なダイナミクスを反映でき，予測の感度と精度の両立が可能となる．

3.2 タイミング調整指標による行動決定

前述の $P(t)$ は予測の精度を向上させるが，それ単体では依然としていつ発話を開始すべきかという具体的なタイミング（Action Decision）を制御しきれず，早合点が発生する．そこで本研究では，VAP の予測能力に時間的な自然さを統合するターンタイミング調整指標 $T(t)$ を提案する．

$$T(t) = E_i(t) \times h(t - t_0) \tag{2}$$

ここで，$h(\tau)$ は次節で詳述するガンマ分布に基づいたハザード関数である．ハザード関数は，沈黙が継続するにつれてターンが移る確率がどのように推移するかという人間の時間知覚をモデル化しようとするものである [8, 9]．

このスコア $T(t)$ が特定の閾値（θ_{TAKE}）を超えた瞬間を実際のターン開始点と定義する．これにより，VAP が導き出す**発話意図**と，ハザード関数が示す**時間的な妥当性**を同一の数理枠組みで扱うことが可能となる．この二つの指標の相乗効果により，沈黙や重なりを伴う多様なターンシフトを，人間らしい自然なタイミングで実現することを目指す．

3.3 ガンマ分布による間のモデル化

人間のターンシフト速度の分布は右に裾を引く非対称な形状をしており，以下のような 3 パラメータガンマ分布で適合することが可能である [8, 9]．

$$f(x; k, \theta, \mu) = \frac{(x-\mu)^{k-1} e^{-(x-\mu)/\theta}}{\theta^k \, \Gamma(k)}, \quad x > \mu. \tag{3}$$

この分布の累積分布関数 $F(\tau)$ および確率密度関数 $f(\tau)$ を用い，ハザード関数 $h(\tau) = f(\tau)/(1 - F(\tau))$ を定義する．ハザード関数はまだターンが移行していない状態で，今この瞬間に移行が発生する条件付き確率を意味する．このガンマ分布のパラメータ（形状 k，尺度 θ，位置 μ）は，ターゲットとなるドメインの平均ターンシフト速度とその分散により，以下のように調整可能である．

$$\theta = \frac{\mathrm{Var}(X)}{\mathbb{E}[X] - \mu}, \quad k = \frac{\left(\mathbb{E}[X] - \mu\right)^2}{\mathrm{Var}(X)}, \quad \mu = \mu_0. \tag{4}$$

本研究では，このパラメータを調整することで，話者の役割に適応したタイミング制御を行う．

4 実験設定

提案手法の有効性を検証するため，以下の条件で実験を行った．

- **モデル**: 標準的な VAP モデル [6] [1] を使用し Switchboard コーパス [10] および NoXi Database [11] で学習した．音響特徴量としてメルスペクトログラムを，文脈特徴量として CPC を使用する．
- **評価データセット**: Switchboard コーパスに加えて日本語対話データセットである Japanese NoXi Database [12] を使用．専門家と初学者の対話 22 セッション（約 6.8 時間）を対象とした．
- **評価指標**: 従来の評価指標である F1 スコアに加え，実際のターン開始位置と予測位置の差を時間軸で計測する Mean Error (s) および標準偏差 (SD) を用いた．
- **比較対象**: ハザード関数による補正を行わず，VAP の $P = P_A/P_B$ が閾値を超えた瞬間をターンとみなす Baseline および，Baseline の予測時刻に 0 秒から 2 秒のランダムな遅延を付与した Random 手法と比較を行う．

5 実験結果

表 1 にターンシフトの予測性能を示す．提案手法（Proposed）を導入した結果，Switchboard では F1 スコアが Baseline の 0.6568 から 0.9020 へと大きく向上し，Precision および Recall もそれぞれ 0.9071, 0.8970 と高い値を示した．また，NoXi Database においても，Baseline の F1 スコア 0.6654 に対し，提案手法では 0.8549 を達成しており，一貫した性能向上が確認された．これは，単なる確率比だけでなく $E_i(t)$ を掛け合わせたことで，直近の発話意図の変化をより正確に捉えられるようになったためである．

次に，ターンシフトのタイミング精度に関する結果を表 2 に示す．従来の Baseline では，平均誤差（Mean）が -0.93 秒となっており，実際のターンシフ

1） 実装は GitHub にて公開されている：https://github.com/maai-kyoto/maai

表 1: ターンシフト予測性能の比較評価

Dataset	Method	F1	Precision	Recall
Switchboard	Baseline	0.6568	0.4944	**0.9779**
	Proposed	**0.9020**	**0.9071**	0.8970
NoXi Database	Baseline	0.6654	0.4993	**0.9971**
	Proposed	**0.8549**	**0.8484**	0.8614

表 2: ターンシフトのタイミング精度評価（NoXi Database）

	Timing (s)		Accuracy (%)	
Method	**Mean**	**SD**	**0.5s**	**1.0s**
Baseline	-0.93	1.05	18.07	38.55
Random	-0.20	1.24	27.98	51.30
w/ full data Gamma	**0.01**	1.07	26.42	52.10
w/ expert data Gamma	0.03	1.06	**32.12**	**60.62**

トよりも早いタイミングで発話開始を予測する傾向が見られた．これは VAP がターンシフトの兆候を早期に検出できていることを示す一方で，タイミング精度の観点では改善の余地があることを意味している．これに対し Random 手法では，平均誤差が -0.20 秒まで緩和され，誤差 0.5 秒以内および 1.0 秒以内の的中率もそれぞれ 27.98%，51.30% へと向上した．全データから推定したガンマ分布ハザード関数を適用した w/ full data Gamma では，平均誤差が 0.01 秒となり，ほぼバイアスのないタイミング予測が可能となった．さらに，話者の役割（専門家）に基づいてパラメータを調整した w/ expert data Gamma では，平均誤差が 0.03 秒，誤差 0.5 秒以内の的中率が 32.12%，1.0 秒以内では 60.62% に達し，全手法中で最も高いタイミング精度を示した．これらの結果から，ガンマ分布に基づくハザード関数を用いた時間補正は，単なる遅延付与とは異なり，ターンシフトの発生タイミングそのものを統計的にモデル化できている点において有効であることが示唆される．

6 考察

実験結果から，VAP の予測能力とハザード関数の統合がタイミング制御において極めて有効であることが示された．Baseline において生じていた大きな負の誤差は，VAP が持つ**将来のいずれかの時点で発話が起こる**という仮定を，現在の行動決定に直結させてしまっていたことが原因である．これに対し，提案手法のハザード関数 $h(\tau)$ は，時間経過とともに今，発話すべき確率を調整するフィルタとして機能し，人間らしい待機の時間を数理モデルとして与えることができる．

しかし，本手法の限界として，達成できているのはあくまでターゲットとなるドメインや話者の役割におけるターンシフト速度の全体的な統計分布の模倣にとどまっている点が挙げられる．表 2 に示す通り，提案手法によって精度は劇的に向上したものの，誤差 1.0 秒以内の的中率は 60.62%である．これは，依然として半数近い事例において 1 秒以上のタイミング誤差が生じていることを意味しており，個別のターンシフト事象における微細な速度変化に完全にアジャストできているわけではない．

この精度上の課題は，本手法が音声活動という音響的な側面にのみ依拠しており，対話の内容を考慮できていないことに起因すると考えられる．実際の対話では，直前の発話が質問であるか，あるいは同意を求めるものであるかといった，言語的な完了感によって許容される間の長さは動的に変化する．より高精度なタイミング制御，すなわち的中率のさらなる向上を実現するためには，VAP の予測に大規模言語モデル等から得られる言語情報を統合し，文脈に基づいてハザード関数の挙動を動的に変化させる枠組みが必要であると考えられる．

7 おわりに

本研究では，VAP の発話予測にガンマ分布のハザード関数を統合することで，ターンシフトのタイミングを精密に制御する新たな予測モデルを提案した．実験により，従来モデルの課題であった早合点を効果的に抑制し，個人特性に応じた行動決定が可能であることを示した．

今後の課題として，言語的コンテキストを統合し，発話の意図や意味論的な完了感に応じた動的なタイミングの微調整を行う手法の確立が挙げられる．これにより，統計的な妥当性を超え，対話の文脈に基づいたより自然なインタラクションの実現を目指す．また，多人数対話への拡張についても検討を進める．

謝辞

本研究は，科研費 23K24910 の助成を受けて実施された．また，理化学研究所大学院生リサーチアソシエイトプログラムの一環として実施された．

参考文献

[1] Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: a review. **Computer Speech & Language**, Vol. 67, p. 101178, 2021.

[2] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn taking for conversation. In **Studies in the organization of conversational interaction**, pp. 7–55. Elsevier, 1978.

[3] Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. **Frontiers in psychology**, Vol. 6, p. 731, 2015.

[4] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. Universals and cultural variation in turn-taking in conversation. **Proceedings of the National Academy of Sciences**, Vol. 106, No. 26, pp. 10587–10592, 2009.

[5] Erik Ekstedt and Gabriel Skantze. Voice activity projection: Self-supervised learning of turn-taking events. **arXiv preprint arXiv:2205.09812**, 2022.

[6] Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. Real-time and continuous turn-taking prediction using voice activity projection. **arXiv preprint arXiv:2401.04868**, 2024.

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. **arXiv preprint arXiv:1807.03748**, 2018.

[8] Kyoko Matsuyama, Kazunori Komatani, Ryu Takeda, Toru Takahashi, Tetsuya Ogata, and Hiroshi G Okuno. Analyzing user utterances in barge-in-able spoken dialogue system for improving identification accuracy. In **INTERSPEECH**, pp. 3050–3053, 2010.

[9] Kazuyo Onishi, Hien Ohnaka, and Koichiro Yoshino. Modeling turn-taking speed and speaker characteristics. In Frédéric Béchet, Fabrice Lefèvre, Nicholas Asher, Seokhwan Kim, and Teva Merlin, editors, **Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue**, pp. 21–31, Avignon, France, August 2025. Association for Computational Linguistics.

[10] John J Godfrey. C. holliman & jane mcdaniel. 1992. switchboard: Telephone speech corpus for research and development. In **Proceedings of the International Conference on Audio, Speech and Signal Processing**, pp. 517–520.

[11] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. The noxi database: multimodal recordings of mediated novice-expert interactions. In **Proceedings of the 19th ACM International Conference on Multimodal Interaction**, pp. 350–359, 2017.

[12] Kazuyo Onishi, Hiroki Tanaka, and Satoshi Nakamura. Multimodal voice activity projection for turn-taking and effects on speaker adaptation. **IEICE Transactions on Information and Systems**, Vol. advpub, p. 2024HCP0002, 2024.