プロアクティブ音声対話の実現に向けた聞き手笑いの先読み

内山大希 [1]　三輪拓真 [2]　吉野幸一郎 [1,2]
[1] 東京科学大学大学院　[2] 奈良先端科学技術大学院大学
uchiyama.h.b139@m.isct.ac.jp　miwa.takuma.mx0@is.naist.jp
koichiro@c.titech.ac.jp

概要

音声対話における能動的発話選択には，聞き手の反応を事前に予測する技術が不可欠である．こうした聞き手の反応を事前にリアルタイムで先読みしてプロアクティブに応答を生成する音声対話システムを構築することを指向して，本研究では，full-duplex 音声対話基盤モデル Moshi を拡張し，対話文脈から直後 10 秒以内の笑い発話開始をリアルタイムに推定する線形分類ヘッドを導入した．PodcastFillers の音声データを用いて評価を行い，強いクラス不均衡下でもチャンスレートを上回る性能で笑い開始を先読みできることを示した．

1　はじめに

近年，大規模言語モデル（LLM）の発展により，テキストを介した対話システムは文脈理解や応答生成の面で大きく向上してきた [1, 2]．一方で，テキスト対話では，声のトーンやタイミング，笑いなどのパラ言語情報が欠落しやすく，対話における感情伝達や関係性形成に重要な社会的シグナルを十分に扱えないという課題が残る [3]．この課題意識を背景に，音声を入出力とする対話（speech dialogue）や，音声から音声への生成（speech-to-speech）により，人間同士の会話に近い相互作用を実現しようとする研究が進んでいる．特に近年提案された Moshi は，音声対話を「音声のストリーム同士の生成」として定式化し，話しながら聞くことを可能にする full-duplex な枠組みでリアルタイム対話を実現している [4]．

人間同士の会話では，話し手は相手の反応を逐次予測しながら話題や表現を調整する．このような能動性（proactivity）は，システムがユーザの反応や目標達成を見越して対話を主導・調整する能力として整理されており，近年サーベイも含めて注目を集めている [5, 6]．しかし既存研究の多くは，話題誘導や目標指向等のテキストベースの能動性に焦点を当てており，相槌や笑いなどの音声対話に固有な反応を発話選択に活用する枠組みは十分に検討されていない．

音声対話における将来予測としては，相槌のタイミング予測や形式予測 [7, 8]，あるいは将来の発話活動を自己教師ありに予測する VAP（Voice Activity Projection）[9] などが提案されている．また，笑いに関しても，ユーザの笑いを検出してシステムが応答笑いを返す「共有笑い」生成 [10] や，スタンドアップ動画で観客の笑いを単語列に対する系列ラベリングとして予測する試み [11] は報告されている．しかし，日常的な対話データを対象に，Full-duplex な音声対話モデル上でリアルタイムに聞き手の笑いを先読みする研究は限定的である．

そこで本研究では，Full-duplex 音声対話モデル Moshi[4] を基盤とし，対話の過去文脈から近未来の笑いイベント発生確率を推定するモデルを構築する．具体的には，Moshi の内部表現に軽量な分類器を付加することで，リアルタイム性を損なわずに笑い予測を行う．学習・評価には，笑いを含む多様な音響イベントが付与された PodcastFillers データセット [12] を用いる．本稿では、Full-duplex 音声対話モデルに対して聞き手笑いの先読みを可能にする最小構成の拡張法を提案する。さらに、PodcastFillers を用いた評価設定を整備してリアルタイム運用を想定した観点から性能を検証し、音声対話における能動的な発話選択へ向けた笑い予測の可能性と課題を議論する。

2　関連研究

2.1　プロアクティブ対話システム

人間同士の会話では，話し手は自身の発話が相手にどのような反応を引き起こすかを予測しながら，話題や表現を選択している．このような対話におけ

る「先読み」は，効果的なコミュニケーションを実現するうえで重要である [3, 13]．対話システムにおいても，ユーザーの反応の予測に基づいて発話を選択するプロアクティブ対話システムが提案されている．このシステムでは，システムが複数の発話候補を生成し，それぞれの発話候補を選択した場合にユーザーがどのような反応を発生させるかを予測する．そして，望ましいユーザーの反応が得られると予測される発話を選択する．例えば，発話選択から相手の感情を予測し，ポジティブな感情を引き出す応答を選択する研究が行われており，先読みが感情の誘発に有効であることが示されている [14]．しかし，多くの先読みを用いる研究では音声入力をテキストに落とし込むため，笑い声や声のトーンといった音声特有のパラ言語情報を活用できていない．

2.2 音響イベント予測

音響イベントとは，対話中に発生する相槌や笑い声，うなずきといった発話以外の音声的な振る舞いを指す [15]．これらは対話における聞き手の反応を示す重要なシグナルであり，円滑なコミュニケーションを支える役割を果たしている．対話システムにおいて音響イベントを適切なタイミングで生成することは，自然な対話体験の実現に不可欠である．

音響イベントの予測は，対話システムが自然なリアクションを生成するための基盤技術として研究されてきた．特に相槌予測については多くの研究が行われており，Ruede らは LSTM を用いた相槌予測モデルを提案している [7]．また，Kawahara らは相槌のタイミングだけでなく，「うん」「へえ」といった形態の予測も行い，言語的・韻律的特徴の組み合わせの有効性を実証した [8]．近年では，Ekstedt & Skantze が Voice Activity Projection（VAP）モデルを提案し，生の音声波形から将来の発話活動を予測する自己教師あり学習手法を開発した [9]．VAP モデルはターンテイキングや相槌，オーバーラップなどの音響イベントを統一的に扱うことが可能であり，Inoue らはこれを拡張してリアルタイムかつ連続的な相槌予測を実現している [16]．

一方，笑い声の予測についても研究が進められている．Inoue らはユーザーの笑いを検出してシステムが笑い返す「共有笑い」の生成システムを提案し，ランダム予測を上回る精度を達成した [10]．また，StandUp4AI データセットでは，スタンダップコメディにおいて各単語の後に笑いが起こるかを予測する系列ラベリングタスクが定式化され，高い F1 スコアが示されている [11]．一方でこのタスクはテキスト入力に対してのみ行われ，音声入力に対する議論は行われていない．そこで本研究では音声入力から直接笑いイベントの予測を行うことで，リアルタイムに応答生成の評価可能な線形分類モデルを提案する．

2.3 Fullduplex 対話モデル

Full-duplex 対話モデルとは，システムとユーザーの音声を同時に処理し，リアルタイムで双方向のやり取りを実現する対話モデルである．従来の対話システムでは，ユーザーの発話終了を検知してからシステムが応答を生成する Half-duplex 方式が主流であった．これに対し Full-duplex 対話モデルは，人間同士の会話のように，相手の発話中でも自身の発話を開始したり，相手の反応を聞きながら発話内容を調整したりすることが可能である．このような特性により，Half-duplex 方式と比較してより自然な対話が実現できるだけでなく，相槌や笑い声といった音響イベントも統一的に扱うことができる．例えば Moshi[4] は，リアルタイムの音声対話を実現するために設計された speech-to-speech 基盤モデルであり，話しながら同時に聞くという Full-duplex な対話が可能である．

3 笑いイベントの先読み

3.1 モチベーション

テキスト対話において，相手の反応やマルチターンの会話の内容の先読みに基づいた発話内容の制御が研究されてきた [13]．しかし，テキストベースでの発話制御は笑い声や声のトーン，発話のタイミングといった音声特有のパラ言語情報の活用は十分に議論されていない．一方で，Moshi や GSLM (Generative Spoken Language Modeling) などの Full-duplex 対話モデルなどは，テキストを介さず音声トークンを直接扱うことで，リッチなパラ言語情報の保持とリアルタイムなターン管理を実現しているが，これらは基本的に「現在の文脈に対する尤もらしさ」に基づいており，将来の予測に基づいた生成制御は未開拓である．そこで，本研究ではユーザの感情や関心を直接的に反映した笑いや相槌などの音響イベントに注目する．システム発話後に生じるこれらのイベントが，発話選択の判断材料として有

用であると考え，複数の応答発話候補の中から望ましいユーザーの音響イベントが得られると予測される発話を選択するプロアクティブ対話システムを目指す．本研究はそのような対話システム構築に向け，リアルタイム応答評価モジュールの提案という位置付けである．

3.2 提案法

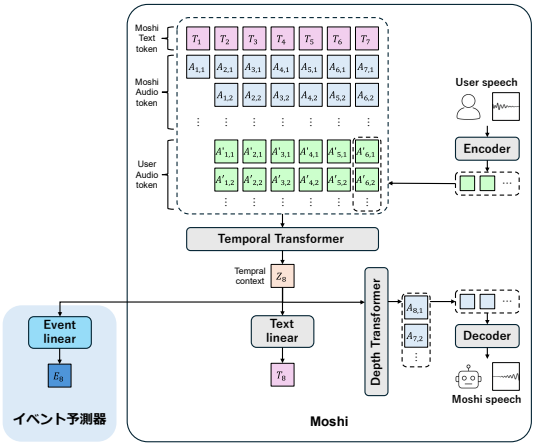

図 1 Moshi のアーキテクチャとイベント予測器

提案法の実装は以下の 3 つのステップからなる．

1. システム発話のみでなく，それに対するユーザの反応も予測するよう Moshi のアーキテクチャを変更
2. Moshi のファインチューニング
3. 音響イベント予測を行うための線形分類の追加および学習

図 1 にこのシステムのアーキテクチャを示す．以下では各ステップについて，詳細に述べる．

Step 1. 本研究では，Moshi のアーキテクチャを拡張し，システム発話の予測と同時にユーザ反応を予測する機構を導入する．Moshi は 80ms の時間フレームごとに，現在のユーザ音声と過去の対話履歴に基づいて次フレームのシステム音声を予測している．本研究ではこのアーキテクチャを変更し，各フレームにおいてシステム音声の予測に加えて，直後に生じる音響イベント（本研究では笑い）の有無を同時に予測するヘッドを追加する．この拡張により，システムの予測を数フレーム先行させて複数の発話候補を生成し，各候補に対してユーザの音響イベント発生確率を推定することが可能となる．

Step 2. Moshi のファインチューニングで行う．pyannote を用いて各エピソードの話者分離を行い，発話時間が最長の話者をシステム音声，その他の話者をユーザー音声として割り当てた．このようにして得られた 2 話者の音声ペアを用いて，Moshi の事前学習と同じ設定で学習を行った．

Step 3. 音響イベントの予測には，Moshi の内部表現から抽出した文脈ベクトル（temporal context）を入力とする線形分類器を用いる．Temporal context は，Moshi のバックボーンである temporal transformer の最終層出力であり，過去の対話履歴とユーザの現在の発話状態を統合した 4096 次元のベクトルである．このベクトルは各時間フレームにおける対話の文脈情報を凝縮しており，次フレームのシステム音声予測の基盤となる表現である．本研究では，この temporal context を入力とし，1 次元のスカラー値を出力する線形分類層を追加する．出力値にシグモイド関数を適用することで，直後の時間フレームにおける笑いイベントの発生確率を推定する．Full-duplex 対話システムに求められるリアルタイム性を維持するために，線形分類器という単純な構造を採用した．この設計により，Moshi の既存の音声生成能力と応答速度を保持しつつ，最小限のパラメータ追加で音響イベント予測機能を実現する．

4 実験

4.1 実験条件

タスク 本研究では，音響イベントの中でも特に笑いイベントに焦点を当てて予測を行った．笑いイベントの予測区間は以下のように定義する．二者間対話において，あるシステム発話の開始から 10 秒以内に笑いイベントが発生する場合，その発話の開始時刻から，同一話者の直前の発話開始時刻までの区間を「笑い予測区間」とする．ただし，この区間が 10 秒を超える場合は，笑い予測区間に含めない．予測タスクの目標は，過去 30 秒間の対話履歴を入力として，次のフレームが笑い予測区間に含まれるかどうかを正しく予測することである．なお，本研究では二者間対話を想定しているため，30 秒間の対話履歴において一方の話者のみが発話しているフレームは予測対象から除外した．

データセット 本研究では，PodcastFillers データセットを用いて笑いイベントの予測の評価を行う．PodcastFillers は 199 のポッドキャストエピソードから構成される英語音声データセットであり，総計 145 時間，350 名以上の話者の音声を含む．このデータセットには 85,803 件の音響イベントが手動

表 1 テストデータにおける混合行列．ただし正例とは，音声入力の終了から 20 秒以内に笑いイベントが発生することである．

	予測ラベル	
真のラベル	負例	正例
負例	1,078,064	71,937
正例	17,114	7,495

でアノテーションされており，そのうち笑い声は 6,623 件含まれている．データセットは train（173 エピソード），validation（6 エピソード），test（20 エピソード）に分割されており，本研究ではこの分割に従って実験を行った．

4.2 評価指標

本研究は将来的に，対話中の相手の音響イベント（ここでは笑い声を伴う発話の開始）を事前に予測し，笑いの多い方向へ対話を生成するためのラベル分類器として用いることを想定する．この用途では，実際には笑い開始が起きないにもかかわらず「起きる」と誤判定して対話生成を誘導してしまうこと（偽陽性）が，不自然な応答やユーザ体験の悪化に直結し得る．そのため本研究では Precision を最重要指標として扱う．

加えて，全体的な正解率として Accuracy，クラス不均衡の影響を緩和する指標として Balanced Accuracy，および正例検出の要約として F1 を併用して報告する．混同行列の要素を TP, FP, TN, FN とすると，

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}, \tag{1}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{2}$$

$$\mathrm{BalancedAcc} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\right). \tag{3}$$

また，F1 は Precision P と Recall R の調和平均であり，

$$\mathrm{F1} = \frac{2PR}{P + R} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

と定義される．

4.3 実験結果と考察

混同行列を表 1 に示す．本結果に基づき，Accuracy は 0.9242，Precision は 0.0944，Balanced Accuracy は 0.6210，F1 は 0.1441 であった．

まず，正例の割合（チャンスレート）は 2.14% と低く，強いクラス不均衡下での評価となっている．この設定において，本モデルの Precision は 0.0944（9.44%）であり，チャンスレート 2.14% を上回った．すなわち，予測を正例とした 79,432 件（7,495 + 71,937 件）のうち，真に正例であったものが 7,495 件含まれており，ランダムに正例を選ぶ場合よりも正例を濃縮して抽出できていることが示唆される．

一方で，Accuracy（0.9242）と比較すると Precision（0.0944）は大きく低く，両者の乖離が大きい．これは，負例が多数を占めるため TN が Accuracy を押し上げやすいのに対し，予測正例の大半が偽陽性（FP=71,937）となっているためである．本研究の想定用途において，偽陽性は不自然な応答やユーザ体験の悪化に直結し得るため，この乖離は運用上のリスクとして重要である．

以上より，チャンスレートを上回る Precision を達成できた一方で，Accuracy と比べて低い Precision の改善が今後の課題である．今後はこの予測結果に基づいて，高確信な場合にのみ笑いが起きると判断して対話生成を誘導する運用を目指す．

5 終わりに

本研究では，full-duplex 音声対話モデル Moshi を基盤とし，対話の過去文脈から近未来における聞き手の笑い発話開始をリアルタイムに先読みする手法を提案した．PodcastFillers データセットを用いた評価の結果，強いクラス不均衡下においても，正例割合であるチャンスレート（2.14%）を上回る Precision を達成し，笑い開始の先読みが一定程度可能であることを示した一方で，Accuracy と比較すると Precision が大きく低いという乖離が確認された．今後は高確信度の予測結果のみを利用する運用戦略や，時間的平滑化，閾値設計の工夫などを通じて，Precision のさらなる向上を図る．

参考文献

[1] Tom B. Brown, et al. Language models are few-shot learners. In **Advances in Neural Information Processing Systems (NeurIPS)**, 2020.

[2] OpenAI. Gpt-4 technical report, 2023.

[3] Sophie K. Scott, Nadine Lavan, Sinead Chen, and Carolyn McGettigan. The social life of laughter. **Trends in Cognitive Sciences**, Vol. 18, No. 12, pp. 618–620, 2014.

[4] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue, 2024.

[5] Yang Deng, Wenqiang Lei, Wai Lam, and Tat-Seng Chua. A survey on proactive dialogue systems: Problems, methods, and prospects. In **Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)**, pp. 6583–6591, 8 2023. Survey Track.

[6] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. Proactive human-machine conversation with explicit conversation goal. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)**, pp. 3794–3804, 2019.

[7] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. Enhancing backchannel prediction using word embeddings. In **Interspeech 2017**, pp. 879–883, 2017.

[8] Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel Ward. Prediction and generation of backchannel form for attentive listening systems. In **Interspeech 2016**, pp. 2890–2894, 2016.

[9] Erik Ekstedt and Gabriel Skantze. Voice activity projection: Self-supervised learning of turn-taking events. In **Interspeech 2022**, 2022.

[10] Koji Inoue, Divesh Lala, and Tatsuya Kawahara. Can a robot laugh with you?: Shared laughter generation for empathetic spoken dialogue. **Frontiers in Robotics and AI**, Vol. 9, , 2022.

[11] Valentin Barriere, Nahuel Gomez, Léo Hemamou, Sofia Callejas, and Brian Ravenet. Standup4ai: A new multilingual dataset for humor detection in stand-up comedy videos. In **Findings of the Association for Computational Linguistics: EMNLP 2025**, pp. 16951–16959, 2025.

[12] Ge Zhu, Juan-Pablo Caceres, and Justin Salamon. Filler word detection and classification: A dataset and benchmark. In **Interspeech 2022**, 2022.

[13] 岸波洋介, 赤間怜奈, 佐藤志貴, 鈴木潤, 徳久良子, 乾健太郎. 対話システムの先読み能力実現に向けた未来の展開まで生成する学習戦略の提案と分析. 人工知能学会全国大会論文集 第 35 回 (2021), pp. 3J2GS6b02–3J2GS6b02. 一般社団法人 人工知能学会, 2021.

[14] Lubis Nurul, Sakti Sakriani, Yoshino Koichiro, and Nakamura Satoshi. Positive emotion elicitation in chat-based dialogue systems. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 27, No. 4, pp. 866–877, 04 2019.

[15] Kouhei Sumi, Tatsuya Kawahara, Jun Ogata, and Masataka Goto. Acoustic event detection for spotting" hot spots" in podcasts. In **INTERSPEECH**, pp. 1143–1146, 2009.

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