複数の日本語データセットによる音声活動予測モデルの学習とその評価

Training and Evaluating Voice Activity Projection Models Using Multiple Japanese Datasets

佐藤友紀[1*]　千葉祐弥[2]　東中竜一郎[1]
Yuki Sato[1]　Yuya Chiba[2]　Ryuichiro Higashinaka[1]

[1] 名古屋大学大学院情報学研究科
[1] Graduate School of Informatics, Nagoya University
[2] NTT コミュニケーション科学基礎研究所
[2] NTT Communication Science Laboratories

Abstract: Voice activity projection is crucial in dialogue systems to facilitate natural turn-taking. Currently, models for predicting speech activity in voice conversations and models incorporating non-audio multimodal information have been reported. However, these models primarily handle English audio data, and there is limited research on adaptation to other languages, including Japanese, and the differences between languages. In this study, we utilized the Transformer-based voice activity projection model proposed by Ekstedt and Skanze. and conducted training on multiple Japanese datasets. Additionally, we performed fine-tuning using Japanese datasets based on an English model. Subsequently, we compared the Japanese model with the English model and assessed performance differences based on datasets. The results confirmed the utility of the model trained on Japanese data for predicting Japanese speech activity and the contribution of fine-tuning based on an English model to performance improvement.

1　はじめに

二者間対話において，話し手と聞き手は交互に発話を行うことで内容を共有し，コミュニケーションを行う．この際，対話の参加者が互いに話し手と聞き手を切り替える行為のことをターンテイキングと呼ぶ [1]．音声対話システムで適切なターンテイキングを行うため，機械学習を用いてターンテイキングを予測する研究が行われている．中でも Ekstedt らの提案した音声活動予測（Voice Activity Projection; VAP）モデルは，音声活動を予測し，その出力を用いることで将来のターンテイキングを予測する [2]．また，このモデルを発展させ，視線方向など音声以外のマルチモーダル情報を取り入れたモデルが提案されている [3]．さらに，実際の対話システムでの利用に対応するために，単一話者の発話のみという，より少ない情報源から二者の音声活動を予測するモデルも提案されている [4]．

しかし，これらの研究は英語のデータセットを中心に扱っており，日本語やその他の言語に関する研究は十分とは言えない．文献 [5] では日本語のデータセットを用いているものの，利用している対話データは観光案内対話データ（3.1 節）だけであり，日本語の大規模なデータセットに関する実験は十分とは言えない．

そこで，本研究では，日本語の対話音声に対して，音声活動予測モデルがどのような性能を示すのかを網羅的に調査する．具体的には，Ekstedt らが提案した Transformer ベースの音声活動予測モデルを用い，内容や対話環境，話者の年齢や関係性などが異なる複数の日本語データセットによる学習と評価を行う．本研究の貢献は次のとおりである．

1. 音声活動予測モデルに内容・対話環境の異なる複数の日本語データセットを混合して適用し，どのような性能を示すかを調査した．
2. 音声活動予測モデルに内容・対話環境の異なる複数の日本語データセットを個別に，または組み合わせを変えて適用し，どのような性能を示すかを調査した．
3. 他言語の音声活動予測モデルを元にファインチ

*連絡先：名古屋大学大学院情報学研究科
〒 464-8601 名古屋市千種区不老町
E-mail: sato.yuki.y1@s.mail.nagoya-u.ac.jp

ューニングを行うことで，性能が向上することを示した．

2 音声活動予測モデル

ここでは，音声活動予測（VAP）モデルについて説明する．この手法では，特定のターンテイキングイベントを直接予測するのではなく，将来の音声活動を予測する．音声活動（Voice Activity; VA）とは，発話の内容とは関係なく，ある区間の発話の有無を表す．音声活動予測を中間表現として用いることで，複数の種類のターンテイキングイベントを 1 つのモデルで予測できるという利点がある．

図 1 に，Cross-Attention-Transformer に基づく二者音声入力型音声活動予測モデル音声活動予測モデルの概要を示す．本モデルは，二者間対話を想定したモデルである．過去 20 秒間の両話者の発話を入力とし，将来の 2 秒間の 2 人の音声活動を予測する．予測する音声活動は 8 つのビンを用いて表現される．これらのビンは，各話者の発話における将来の 0-200ms，200-600ms，600-1200ms，1200-2000ms の区間に対応し，「音声あり」または「音声なし」に離散化されている．この表現では，2 人の音声活動は 256 種類のクラスで表される．したがって，音声活動予測モデルは将来の音声活動を 256 クラスで分類するように学習される．

しかしながら，音声活動予測の分類結果をそのまま直感的に解釈したり，ターンテイキングイベントの予測に使用するのは困難である．そこで，解釈を容易にするために，予測結果を要約した 2 つの尺度が提案されている．一つは話者 s の短期的な音声活動予測である $p_{now}(s)$ である．これは各話者のビンの 0-200ms および 200-600ms 区間の確率値を合算し，その合計に softmax を適用することで計算され，「次の 600ms 以内に話者 s が話す確率がどのくらいか」を表す尺度である．もう一つは，長期的な音声活動予測を表す $p_{future}(s)$ である．これは，600-1200ms および 1200-2000ms 区間に対して $p_{now}(s)$ と同様の計算をすることで求められる．

なお，本研究で用いたソースコードは，先行研究[1]によって公開されている．

3 データセット

3.1 使用データセット

本研究では，文献 [5] で用いられているデータセットを含む，様々な日本語の音声対話データセットを学習に利用する．モデルの学習およびテストに用いる 6 つの異なるデータセットについて述べる．

[1] https://github.com/ErikEkstedt/VAP

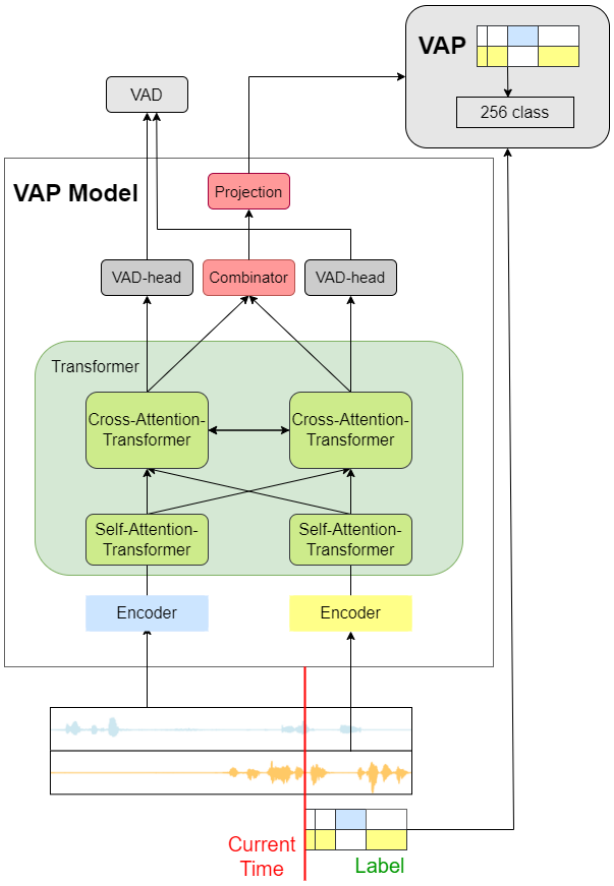

図 1: Cross-Attention-Transformer に基づく二者音声入力型音声活動予測モデル

Switchboard Corpus (Sw) [6]
543 人の参加者，2,438 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の自由対話が収録されている．合計で約 259 時間の対話が収録されており，言語は英語のみである．

日本語日常会話コーパス（CEJC） [7]
日常場面の中で自然に生じる日本語の会話が収録されている．原則として各話者に対応するボイスレコーダーを設置して録音した音源となっている．自然な日常の環境で録音されているため，環境音によるノイズや他の話者の発話が聞こえる部分が非常に多いという特徴がある．本研究では，二者間の対話で，各話者に対応する音源があるもののみを選択した．使用した対話数は 167 対話，合計時間は約 53.3 時間である．

日本語話し言葉コーパス（CSJ） [8]
日本語の自発音声を大規模に集めて形態論情報や係り受けなどの様々な研究用情報を付加した話し言葉研究用のデータベースである．本研究では，インタビュー形式，ギャラタスク形式の二者間対

話のみを選択した．使用した対話数は 58 対話，合計時間は約 12.2 時間である．

CALLHOME Japanese Speech (Call) [9]
北米在住の日本語母語話者が海外（主に日本）の知人に電話を掛けることで始まる日本語の自由対話が収録されている．ほとんどの音声は家族や親しい友人同士の対話である．本データセットのうち，ほとんどは二者間対話であるが，一部に他の話者に交代したり，子供が割り込む音声が存在する．本研究ではそのようなデータを除いた純粋な二者間対話のみ使用した．さらに，本データセットは各対話の一部分（10 分程度）にしか書き起こしおよびタイムスタンプが付与されていない．そこで，前処理として，タイムスタンプの存在する部分のみを切り出した音源を作成した．使用した対話数は 105 対話，合計時間は約 16.2 時間である．

接客対話データ (Sales)
我々が独自に構築したデータセットで，人間同士の模擬接客対話データが収録されている．具体的には，接客業経験のある店員役を用意し，所定の商品カテゴリに属する複数の商品について顧客役に 15 分程度の接客を行う．対話は Zoom によるビデオ通話によって行われ，各チャネルの音声が収録されている．1 回の対話は 15 分以上である．使用した対話数は 236 対話, 合計時間は約 64.8 時間である．

観光案内対話データ (Travel) [10]
人間同士の模擬接客対話データが収録されている．具体的には，接客業経験のある店員役を用意し，旅行代理店における観光相談を模した対話を行う．顧客役は，児童から高齢者まで，幅広い年齢層の話者で構成されている点が特徴である．対話は Zoom によるビデオ通話によって行われ，各話者の音声が別々のチャネルで収録されている．1 回の対話は 20 分程度である．この中には音源長の調整が必要なものが見られたため，一部のデータを除外して使用した．使用した対話数は 233 対話, 合計時間は約 82.1 時間であった．本データセットは，文献 [5] でも使用されている．

3.2 データの整形と分割

Sw，Call はアップサンプリングを，Sales，Travel はダウンサンプリングを行い，すべての音声を 16kHz に統一した．それぞれのコーパスは，Train，Validation，Test データに分割した．データセットによって時間にばらつきがあるものの，日本語データセットの総数としては Sw と同程度のサンプル数が得られた．データの内訳を表 1 にまとめる．

表 1: 学習データの内訳（単位は時間）

言語	データセット	Train	Val.	Test
英語	Sw	207.8	38.4	13.0
日本語	CEJC	41.4	8.7	3.1
	CSJ	9.5	1.9	0.8
	Call	12.8	2.6	0.9
	Sales	51.6	9.6	3.6
	Travel	65.4	12.0	4.6
	合計	180.7	34.8	13.0

4 実験

4.1 実験条件

音声活動予測モデルの Transformer は隠れ層サイズ 256，自己注意機構 1 層，相互注意機構 3 層，4 ヘッドで構成した．ドロップアウト 0.1，早期停止基準を 10 エポックとし，最適化手法 AdamW，学習率 3.63e-4，バッチサイズ 4 で学習を行った．学習においては，出力する音声活動予測と正解ラベルのクロスエントロピー損失を損失関数として用い，Validation データに対する損失が最も小さくなる時点のモデルを最良のモデルとした．seed を 0 から 4 まで変化させて学習したモデルそれぞれについて評価を行い，評価値の平均を最終的な評価値として用いた．

4.2 比較モデル

日本語の対話音声に対する音声活動予測モデルの性能を網羅的に調査するため，以下のように学習内容の異なるモデルを用意した．

Switchboard 学習モデル
先行研究の再現およびベースラインとして，Sw による学習を行う．

日本語データセット学習モデル
音声活動予測モデルが日本語の対話音声に対してどのような性能を示すのかを調べるため，3 節で述べた 5 種類の日本語データセットすべてを用いて学習を行う．

日本語データセット単体学習モデル
音声活動予測モデルが内容や対話環境の異なる対話音声に対してどのような性能を示すのかを調べるため，3 節で述べた 5 種類の日本語データセットをそれぞれ個別に用いて学習を行う．

日本語データセット w/o 学習モデル

音声活動予測モデルが内容や対話環境の異なる対話音声を複数学習した際にどのように性能を示すのかを調べるため，日本語のデータセット 5 種類のうち，1 種類ずつデータセットを除外して残りすべてを用いて学習を行う．

他言語モデルをベースにしたモデル

ターンテイキングにおける言語非依存な特徴を学習に役立てることができるかを調査するため，Sw で学習したモデルを元に，日本語データセットによるファインチューニングを行う．

5 性能評価

5.1 評価指標

テストデータに対する損失（loss）によって，モデルの性能を比較する．しかしながら，音声活動予測の精度だけでは，実際のターンテイキングイベントをどの程度予測できるのか評価できない．そこで，Ekstedt らは，音声活動予測を用いて，ゼロショットで検出されたターンテイキングのイベントを予測する方法を提案している．本研究ではこのうち，評価指標として，Shift/Hold，Shift-prediction の 2 つを使用する．以下に，これらの評価指標について説明する．なお，話者が交代することを Shift，交代せずに続けることを Hold と定義する．

5.1.1 Shift/Hold (S/H)

この評価指標は，両方の話者が発話していない無声区間において，次の話者をモデルがどれだけ正確に予測できるかを評価する．つまり，現在の話者がターンを保持するか，次の話者にターンが移るかを予測するものである．本研究では，25ms 以上の無声区間を評価対象とし，前後の話者から Shift/Hold イベントのどちらであるかを判定したうえで，この範囲の各フレームの出力について評価を行う．Shift イベントの場合は，この範囲で無声区間後の話者 s_0 に対して $p_{now}(s_0)$ を抽出し，閾値を超えていれば正解とする．Hold イベントの場合は，この範囲で無声区間後の話者 s_0 に対して $1-p_{now}(s_0)$，つまりもう一方の話者が話す確率 $p_{now}(s_1)$ を抽出し，閾値を下回っていれば正解とする．ただし，判定に用いる閾値は，学習に用いた Validation セットで最も評価の高かった値を採用する．

5.1.2 Shift-prediction (S-pre)

この評価指標は，話者がまだ発話を続けている状態で，近い将来の話者交代をどれくらい正確に予測できるかを評価する．つまり,「まもなく話し終わり，ターンが移る」ことを予測するものである．本研究では，無声区間前の 500ms の範囲を Shift と予測すべき区間とする．また，単一の話者のみが発話をしており，もう一方の話者の次の発話から十分に（2 秒以上）離れた領域の 500ms を Hold と予測すべき区間とする．この範囲で p_{future} を抽出し，Shift/Hold と同様の方法で判定を行う．

表 2: Switchboard 学習モデルの評価値

	テストデータ				
評価値	CEJC	CSJ	Call	Sales	Travel
loss	3.80	3.47	4.28	2.48	2.90
S/H	0.582	0.742	0.714	0.680	0.509
S-pre	0.464	0.593	0.666	0.625	0.573

表 3: 日本語データセット学習モデルの評価値

	テストデータ				
評価値	CEJC	CSJ	Call	Sales	Travel
loss	3.40	3.07	3.11	2.08	2.21
S/H	0.616	0.787	0.796	0.829	0.740
S-pre	0.588	0.697	0.760	0.771	0.716

5.2 結果

まず，先行研究に基づき，Sw で学習したモデルを Sw でテストした場合の評価値は loss が 2.48，S/H が 0.81，S-pre が 0.71 であった．続いて，表 2 に Sw で学習したモデルを各日本語データセットでテストした際の評価値を示す．全体として，Sw でテストした場合に比べ，性能が低下していることがわかる．

表 3 に，5 種類の日本語データセットすべてを学習したモデルについて，各日本語データセットで評価した際の評価値を示す．なお，同一のデータセットであっても，評価に用いたデータは学習に用いたデータに含まれていない．表 2 と比較して，全体に改善が見られる．特にターンテイキング予測の評価指標では，CEJC を除いて，英語モデルの英語での評価と同程度の値を示している．このことから，音声活動予測モデルは日本語においても同等の精度でターンテイキングを予測できることがわかる．しかしながら，CEJC のテストデータに対しては，どの評価指標でも性能が低い結果となっている．CEJC は他のデータと比較して環境雑音や他の話者の音声の入り込みが多く，特に予測が困難であったと考えられる．また，表 2，3 を総合して，同一のテストデータに対しては loss が減少すると S/H，S-pre の値が上がる傾向にあることがわかる．しかし，異なるテストデータに対してはほとんど相関がない．例

表 4: 日本語データセット単体・w/o 学習モデルの比較（loss）．w/o は 5 種類のデータセットのうちそのデータセット以外で学習したことを示す．

	テストデータ				
学習データ	CEJC	CSJ	Call	Sales	Travel
CEJC	3.39	3.47	3.78	3.02	2.94
CSJ	5.24	3.11	3.83	3.14	3.31
Call	4.94	3.70	3.11	3.21	3.09
Sales	4.61	3.80	3.82	2.11	2.73
Travel	4.12	3.58	3.78	2.78	2.23
w/o CEJC	4.09	3.10	3.11	2.08	2.22
w/o CSJ	3.41	3.29	3.11	2.09	2.21
w/o Call	3.40	3.07	3.59	2.08	2.22
w/o Sales	3.40	3.06	3.10	2.74	2.21
w/o Travel	3.39	3.07	3.09	2.12	2.58

表 5: 英語モデルをベースに日本語データでファインチューニングしたモデルの評価値

	テストデータ				
評価値	CEJC	CSJ	Call	Sales	Travel
loss	3.39	3.05	3.09	2.09	2.20
S/H	0.617	0.806	0.884	0.812	0.743
S-pre	0.597	0.706	0.820	0.755	0.720

えば，表 3 で本モデルの Call と Sales に対する性能について比較すると，loss の比較では Call が 3.11，Sales が 2.08 となっており，Sales への対応に優れていると読み取れる．一方，S/H の比較では Call が 0.796，Sales が 0.829 となっており，Call への対応に優れていると読み取れる．したがって，単一のテストデータの loss を比較し，これを改善することはターンテイキング予測の精度向上につながるが，異なるテストデータの loss を比較してもターンテイキング予測の精度には結び付かないと言える．

表 4 に，各日本語データセットを個別に学習したモデルと，5 種類の日本語データセットのうち 1 種類を除外して学習したモデルについて，各日本語データセットで評価した際の音声活動予測と正解ラベルのクロスエントロピー損失（loss）を示す．表 4 の上部は，各日本語データセットを個別に学習したモデルの評価値である．対角要素は学習したデータセットと同一のデータセットでテストした場合を表す．テストデータと同一のデータセットを学習することで性能の向上が見られる．また，対面，Zoom など，対話環境が似ているデータを学習したモデルでは，他のモデルに比べて性能の向上が見られる．対話環境によって応答タイミングが変化する場合があるが，本モデルはこのような対話環境による特徴を捉えている可能性がある．しかし，対話環境によって録音品質が異なる可能性もあるため，原因の特定にはより詳細な分析が必要である．

表 4 の下部は，5 種類の日本語データセットのうち 1 種類を除外して学習したモデルの評価値である．対角要素は学習に含まれないデータセットで評価した場合を表す．対角要素の損失が他の要素に比べて大幅に大きくなっていることから，性能の改善には内容が類似したデータでの学習が必要であることがわかる．一方で，学習に用いていないデータの評価では，単体学習モデルよりも高い性能を示した．例えば，w/o CEJC で学習したモデルの CEJC での評価結果は 4.09 となっているが，これは CSJ，Call，Sales，Travel でそれぞれ学習したモデルの CEJC での評価結果 4.12，4.61，4.94，5.24 よりも良い評価結果である．このことから，学習する音声の量と多様性を増やすことで，未知の音声にも対応できる可能性があるといえる．

表 5 に，英語モデルをベースに 5 種類の日本語のデータセットすべてでファインチューニングを行ったモデルの評価値を示す．表 3 と表 5 を比較すると，英語モデルをベースにファインチューニングを行うことで，多くの場合で性能が向上している．この結果は，ターンテイキングには言語に依存しない特徴があり，複数の言語のコーパスを学習することでその共通要素を学習できることを示唆している．特に，Call を対象としたテストでは大幅な性能の向上が見られる．Sw と Call はともに電話による対話音声であり，特に共通性が高かったと考えられる．これに対して，Sales や Travel に対しては性能がほとんど変化しないか，下がっている．これらのデータは Zoom での対話音源であり，かつ，基本的に初対面の相手との対話であるため，Sw との類似性が特に低かった可能性が考えられる．しかし，Sales に対する評価値は他のデータセットと比べてもともと高く，英語モデルをベースにすることで性能が向上する余地がなかった可能性も考えられる．また，英語モデルをベースにすることで日本語データへの精度が向上することから，日本語データのみを用いるモデルの精度には改善の余地があることがわかる．より大規模かつ入力音声と類似したデータセットを追加して学習することで，性能の向上が期待される．

以上の結果を比較すると，本研究で実験したモデルでは，英語モデルをベースにすべての日本語データを学習させたモデルが最も良い性能を示すことがわかる．

6 おわりに

本研究では，Transformer ベースの音声活動予測モデルを用い，内容や対話環境，話者などが異なる複数の日本語データセットによる学習と評価を行った．その結果，条件の類似した音声を学習させていれば，日

本語でも英語に匹敵する性能を発揮することが明らかになった．そのうえで，英語音声による学習モデルをベースにファインチューニングを行うことで，直接学習を行う場合よりも精度が向上することを示した．この結果から，ターンテイキングには言語に依存しない特徴があることが示唆された．

しかしながら，日本語での音声活動予測モデルの精度には改善の余地があることも示された．入力音声と類似した音声データを追加することで改善すると予想されるが，どのような点で類似している音声データが必要なのかは明らかになっていない．そのため，今後より詳しい検証を行い，音声活動予測モデルの精度を上げる方法を検証していく必要がある．

謝辞

本研究は，ムーンショット目標 1「2050 年までに、人が身体、脳、空間、時間の制約から解放された社会を実現」（JPMJMS2011）の支援を受けた．また，新学術研究「人間機械共生社会を目指した対話知能システム学」（19H05692）の支援を受けた．

参考文献

[1] Skantze, G.: Turn-taking in Conversational Systems and Human-Robot Interaction: A Review, *Computer Speech & Language*, Vol. 67, pp. 1–26 (2021).

[2] Ekstedt, E. and Skantze, G.: Voice Activity Projection: Self-Supervised Learning of Turn-Taking Events, in *Proceedings of the Interspeech*, pp. 5190–5194 (2022).

[3] Onishi, K., Tanaka, H. and Nakamura, S.: Multimodal Voice Activity Prediction: Turn-Taking Events Detection in Expert-Novice Conversation, in *Proceedings of the 11th International Conference on Human-Agent Interaction*, pp. 13–21 (2023).

[4] 大西一誉, 田中宏季, 中村哲：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, HCG シンポジウム, pp. 1–5 (2023).

[5] Inoue, K., Jiang, B., Ekstedt, E., Kawahara, T. and Skantze, G.: Real-time and Continuous Turn-Taking Prediction Using Voice Activity Projection (2024), arXiv preprint: 2401.04868.

[6] Godfrey, J., Holliman, E. and McDaniel, J.: SWITCHBOARD: Telephone Speech Corpus for Research and Development, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 517–520 (1992).

[7] 小磯花絵, 天谷晴香, 居關友里子, 臼田泰如, 柏野和佳子, 川端良子, 田中弥生, 伝康晴, 西川賢哉, 渡邊友香：『日本語日常会話コーパス』設計と構築, 国立国語研究所論集, Vol. 24, pp. 153–168 (2023).

[8] Maekawa, K.: Corpus of Spontaneous Japanese: Its Design and Evaluation, in *Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 1–6 (2003).

[9] Wheatley, M. K., Barbara and Kobayashi, M.: CALLHOME Japanese Speech LDC96S37. Web Download. Philadelphia: Linguistic Data Consortium (1996).

[10] Inaba, M., Chiba, Y., Higashinaka, R., Komatani, K., Miyao, Y. and Nagai, T.: Collection and Analysis of Travel Agency Task Dialogues with Age-Diverse Speakers, in *Proceedings of the 13th Language Resources and Evaluation Conference*, pp. 5759–5767 (2022).