修士論文

2者対話の返答音声におけるタイミングと自然さの関係分析

吉川 禎洋

主指導教員　岡田 将吾

北陸先端科学技術大学院大学
先端科学技術研究科
(情報科学)

令和６年９月

Abstract

The timing estimation models in spoken dialogue systems (SDSs) have typically been trained by human responses in order to achieve the appropriate response timing. However, human response timings are not always appropriate: in a previous experiment in which annotators listened to responses with the timings replaced by fixed values, some responses with the mode value sounded more realistic than actual human responses. Since this previous experiment was a small-scale preliminary one that only showed that some speakers tended to be significantly preferred, in the current study, we conducted an experiment on about 1,700 human responses, and scored whether they could be replaced with the mode value. The results showed that the annotators tended to feel that mode (or perhaps from 0 ms to 400 ms) responses are more appropriate than actual overlappings. We determined the responses that could and could not be replaced with the mode value by a chi-square test and then formulated a detection task to predict them from the scores. The evaluation results showed that our proposed simple model outperformed random selection with the AUC of 0.650. On the basis of these results, we discuss about the challenging for implementing SDSs, using the score to predict which responses or response timings are appropriate for the SDS users. Our findings may suggest a more efficient way to determine the appropriate response timing for SDSs compared to training models by corpus data.

目 次

図 目 次

表 目 次

第1章　はじめに

音声対話システムにおいて自然な返答を実現するには，人間が適切と感じる返答タイミングを予測し，適切なタイミングでシステムに応答させることが重要である．しかし，人間同士の日常会話における適切な返答タイミングは明らかでなく，その適切さは人間の感覚に依存する．たとえば，発話者に返答する話者は必ずしもそのタイミングに注意を払うわけではないため，一部の返答は発話者あるいは観察者にとって不適切に (早くまたは遅く) 聞こえる可能性がある．加えて，先行研究 [1] では，返答タイミングを特定の秒数に置き換えた一部の返答がコーパスに収録された返答よりも実際の会話らしく聞こえたことが報告されている．このことは，コーパスに記録された返答タイミングを正しいデータとして返答生成モデルを学習させたとしても，必ずしも適切なタイミングで返答できるとは限らないことを暗に意味している．したがって，適切な返答タイミングを実現するには，コーパスデータによってモデルをトレーニングするだけでなく，各返答にたいして人間が適切と認識する返答タイミングを分析することも重要である．

しかしながら，上記の先行研究 [1] は話者単位に有意な回答としての選択されやすさがあることを示すのみの小規模な予備実験であり，コーパスに収録された返答タイミングと置き換えた返答の一方がどのように適切なのかを示していなかった．加えて，最頻値に置き換えた返答とコーパスの返答では評価結果が同率であったため，その先行研究においてはコーパスで学習することが人間の適切な返答タイミングを実現に繋がるのかが明らかではなかった．そこで本研究では，人間が適切だと認識する返答タイミングを音声対話システムへ応用することを目指すうえで，まず音声対話においてどういった返答がそのタイミングを特定の秒数に置き換えられると適切に聞こえるようになるのかを詳しく分析すべきだと考えた．その分析のために, まず先行研究と同様にタイミングを置き換えた返答を人間が実際に聴いて評価する聴取評価を実施した. そこから先行研究を拡張し，その評価結果は話者単位ではなく話者に共通する返答タイミングの特徴を返答単位で分析した．コーパスに含まれるすべての返答を評価することはコストが大きいため，置き換えられる前の返答タイミングの違いによる評価の違いを詳細に分析できるように，返答タイミングの分布から速い返答，遅い返答，その間の返答の 3 種類を均等に抽出した．利用するコーパスには，日常会話における会話パターンを分析するため，単一話者の多様な会話が収録された日本語の 2 者対話データを選択した．そしてこれらの多様な返答パターンとその評価結果から, 音声対話に適した返答タイミングがどのようなものであるかを分析した.

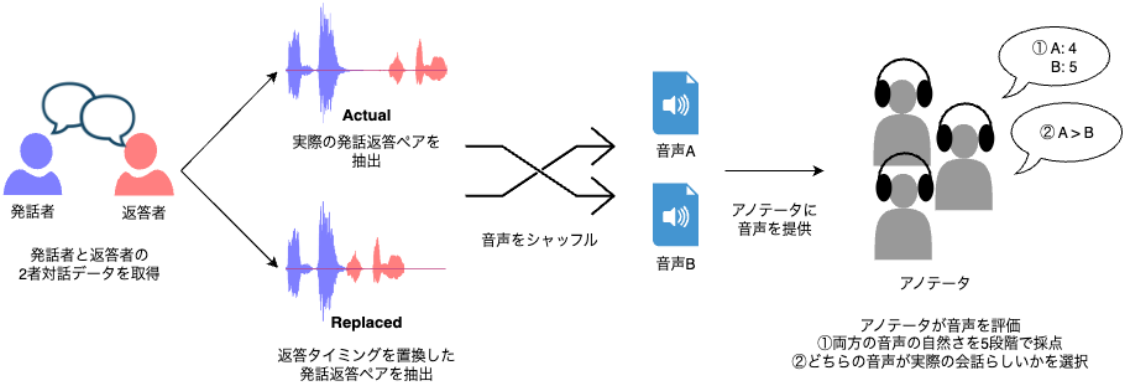

図 1.1: 返答タイミングの自然さの聴取評価の概要

本研究における聴取評価の概要を図 1.1 に示す．実験に用いる 2 者対話データを取得するため，コーパスから 8 名の返答音声を約 1,700 件抽出し，各返答へ 17 名のアノテータが割り当てられるように参加者を集めて実験を行った．実験のなかで，アノテータはコーパスに収録された返答 (Actual と呼ぶ) と特定の秒数に置き換えた返答 (Replaced と呼ぶ) の両方を聴き比べてその自然さを評価した．置換する秒数は先行研究 [1] と同様にコーパスの最頻値を用いた．自然さの評価では，音声そのものの自然さと返答タイミングの置き換えによる変化を明らかにするため，アノテータは①両方の音声の自然さを 5 段階で採点と②どちらの音声が実際の会話らしいかを選択した．それらの評価結果をもとにどういった返答がタイミングの影響を受けるのかを実際の返答タイミングごとに分類して分析した．

音声対話システムへの応用に向けて，返答タイミングにおける人間の感覚を自動評価するモデルを想定し，本研究ではアノテータから収集した評価結果を予測するモデルの構築も試みた．本研究では，音声対話システムの実装までは実施せず，ベースラインモデルとして簡易な構成のニューラルネットワークモデルを学習させてその予測結果を分析することで，どういった返答や評価指標において予測精度が高いのか，言い換えると，どういった音声対話システムへの応用が可能なのかを考察した．

これらの実験を通して，2 者対話における返答タイミングと人間が感じる自然さの関係性において，本研究での成果は以下の 6 点である．

1. 2 者対話の返答音声を聴き比べたとき，返答タイミングが遅くなるほど自然さの評価が低くなることを示した

2. 2 者対話の返答音声を聴き比べたとき，実際の返答タイミングが 0 秒から−1 秒で少し発話が重複する場合，返答タイミングを 0 秒に置き換えたほうが適切であると感じる傾向にあることを示した

3. 2 者対話の返答音声を聴き比べたとき，実際の返答タイミングが 0 秒から 400 ミリ秒で少し間を置いて返答する場合，返答タイミングを 0 秒に置き換える

よりも実際の返答タイミングで返答したほうが適切であると感じる傾向にあることを示した

4. 2者対話の返答音声を聴き比べたとき，置き換える返答タイミングと実際の返答タイミングの違いが数十ミリ秒でも評価に差が生じることを示した

5. 返答音声の自然さをニューラルネットワークモデルで予測した場合，低評価である返答の予測精度が低く，高評価である返答の予測精度が高くなることを示した

6. 返答タイミングを0秒に置き換えるべきかどうかをニューラルネットワークモデルで予測した場合，返答音声を入力にすると6割以上の精度で予測ができるが，十分なデータ量がなければ機能しない可能性があることを示した

本論文の構成を説明する．まず，第2章では返答タイミングの自然さの評価に関連する研究を紹介する．次に，第3章では返答タイミングの分析に利用するデータセットを作成するための手順を説明する．具体的には，2者対話コーパスから分析対象の返答音声を抽出する方法と聴取評価の実施方法を説明する．そして第4章では聴取評価の結果について分析し，第5章ではその聴取評価によって得られた評価スコアを予測するモデルを構築してその性能を分析する．その後，第6章で本研究では分析しきれなかった点や音声対話システムへの応用について述べ，第7章で本研究をまとめる．

第2章　関連研究

2.1　音声の自然さの評価

音声の自然さの評価のうち, 最も代表的な人間の評価値のひとつに Mean Opinion Score(MOS) がある．MOS 評価では，聴取した音声に対してその品質を主観的かつ絶対的に 1(bad) から 5(excelent) の 5 段階で評価する．VoiceMOS Challange[2] と呼ばれる MOS を予測するモデルを構築することで合成音声の自動評価を目指すコンテストにおいては，その MOS 予測モデルを利用してノイズが多く含まれる Youtube 等のデータから音声合成に重要な音声のみを学習する応用例 [3] が存在する．また，Non-intrusive Objective Speech Quality Assessment(NISQA)[4] と呼ばれる，オンライン通話音声等にたいして MOS やノイズネス，ラウドネスを人手で評価したデータセットを学習したモデルも存在する．NISQA はモデルの重みを公開しており，それを使って音声品質の低下要因を容易に見積もることができる．本研究における聴取評価と RS はそれらと同様に，不確実性のあるコーパスに収録された返答タイミングから高品質な返答を実現するための自動評価を目指している．

2.2　返答タイミング

返答タイミングはいくつかの先行研究により，タスク会話かどうかや簡単に答えられる質問かどうか等，会話の状況によって分布が異なることが知られている [5, 6]．また Kendrick と Torreira によるコーパス分析 [7] では，返答タイミングが 700 ミリ秒の場合，好ましくないアクションの割合が好ましいアクションの割合よりも大幅に大きいことが示唆されている．加えて，Roberts と Francis による返答タイミングを調整して実際に会話を行う実験 [8] では，評価スコアとして実験参加者が返答を聞いたときに感じる返答者の意欲 (willingness) の度合いを測定したが，その評価は返答タイミングが 600 ミリ秒を超えると低下し始め，それ以降は 700 ミリ秒から 800 ミリ秒にかけて大幅に低下することが示された．これらの分析結果は音声対話システムの評価においても非常に有用であり，本研究もこれらと同様に，返答タイミングによる人間の感じ方の違いを明らかにする試みであると言える．ただし，本論文では返答タイミングごとの違いを言及するにとどめ，質問にたいする返答であるかどうかや返答そのものに対する感情的な印象と, 本研究における返答の自然さとの関係分析は今後の課題とする．

返答タイミングの対話システムへの応用研究としては，バス情報案内の自動通話システムにおけるユーザーでの実証実験 [9] を始め，LSTM モデルを利用した返答タイミング予測モデルの構築 [1] や，音声認識による遅延の削減 [10] が進められている．一方で，その実際の返答タイミングの一部がコーパスにおける最頻値に置き換えた音声よりも実際の会話らしくないことが報告 [1] されている．その先行研究では，いくつかのフィルタ条件を定義してコーパス全体の発話から速い返答 (early と呼ばれた) と遅い返答 (late と呼ばれた) を抽出し，opposite(early であれば late 全体の平均値，late であれば early 全体の平均値) または mode(コーパス全体の最頻値) に返答タイミングを加工する．そして，アノテータは実際の返答タイミング (true と呼ばれた) と加工された返答音声の両方を聴き，「どちらが実際の会話で発生しうる返答タイミングでしたか？」という質問に回答する．そして，その回答結果から実際の返答タイミング (true) と加工音声 (opposite もしくは mode) のどちらに統計的な有意な回答としての選択されやすさがあったかについて返答した話者ごとに p<0.05 レベルの統計検定を実施した．

先行研究の結果，true vs opposite では 10/16 が true を統計的に有意に選択していた．その一方で，true vs mode では 3/16 が true を統計的に有意に選択し，3/16 が mode を統計的に有意に選択しており評価がわかれた．そこで本研究では，先行研究において評価がわかれた true vs mode について調査を行う．本研究における Actual vs Replaced はこの true vs mode の関係を表す．

第3章　データセット

本章では，返答音声を人間が自然に感じるかどうかを分析するためのデータセットをどのように作成したのかとそのデータの特徴について説明する．図1.1に示したように，本研究では2者対話データから返答にたいする発話と返答を抽出し，それらの音声をアノテータが聴取して評価することで，返答音声を人間が自然に感じるかどうかのデータセットを構築した．その2者対話データとアノテータによる聴取評価データの作成方法について説明する．

3.1　2者対話データ

3.1.1　コーパス

2者対話データには林ら [11] の研究で利用された日本語での対面会話における2者間のやりとりが含まれたコーパスを利用した．このコーパスの主な特徴は単一話者の多様な音声サンプルが収録されている点である．コーパス収録に参加する話者は友人3人および初対面3人と対話する．本研究では返答タイミングを分析するため，返答あるいは返答者を基点としてデータを整理する．本研究では，このコーパスから4人組の友人グループ2つを選択し，計8名を分析対象の話者とした．つまり，返答者はすべて友人グループに含まれており，発話者は返答者にとって友人あるいは初対面の話者となる．本研究においては友人と初対面の返答タイミングの違いには言及しない．

3.1.2　返答の定義

本研究に利用するコーパスは対面会話における2者間のやりとりであるため，どちらが発話者でどちらが返答者かは曖昧である．よって，返答タイミングを分析するうえで返答とは何かを定義する必要がある．本研究では返答を①発話者が発話して返答者が返答した音声であり②発話よりも短い音声と定義する．したがって，返答音声には発話者の質問にたいする回答だけでなく，相槌，笑い声，驚くなどの感情的な反応，あるいは発話者の相槌に続く返答者の発話も含まれる．

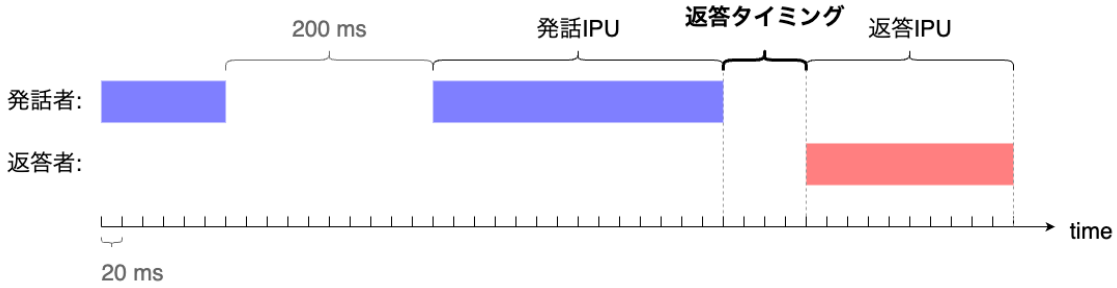

図 3.1: 返答タイミングの定義

3.1.3 返答タイミングの定義

返答タイミングを分析するうえで，コーパスに収録された音声にたいして発話区間をアノテーションする必要がある．本研究で利用するコーパスには人間の手によるアノテーションが存在しないため，音声区間検出のオープンソースソフトウェア https://github.com/wiseman/py-webrtcvad を利用してアノテーションを行った．このコーパスは静かな環境で収録されており騒音ノイズはほとんどなかったためノイズ補正は実施していないが，僅かな呼吸音を発話と認識してしまうケースがあったため音圧の小さいフレームについてはツールによって音声区間だと認識されても音声区間ではないと判定した．

返答タイミングの定義を図 3.1 に示す．発話区間は 20 ミリ秒単位の音声フレームごとに音声区間を判定し，音声区間となるフレームが 200 ミリ秒続いた場合に発話開始時点とし，音声区間ではないフレームが 200 ミリ秒続いた場合に発話終了時点とした．発話区間として判定された音声は間休止単位 (IPU) として抽出する．そして，返答タイミングは発話者の IPU から返答者の IPU に話者交替するまでの時間として定義する．

これらの定義から抽出された本研究の分析対象である 8 名の返答者の返答タイミングの分布を図 3.2 に示す．返答タイミングの平均値は 82 ミリ秒，中央値は 60 ミリ秒，最頻値は 0 ミリ秒である．ここでの最頻値を本研究で利用する置き換える返答タイミングの秒数として定義する．

3.2 聴取評価データ

本研究における自然さの評価のための手順を図 1.1 に示した．本セクションではこの手順についてそれぞれ具体的な方法を示す．

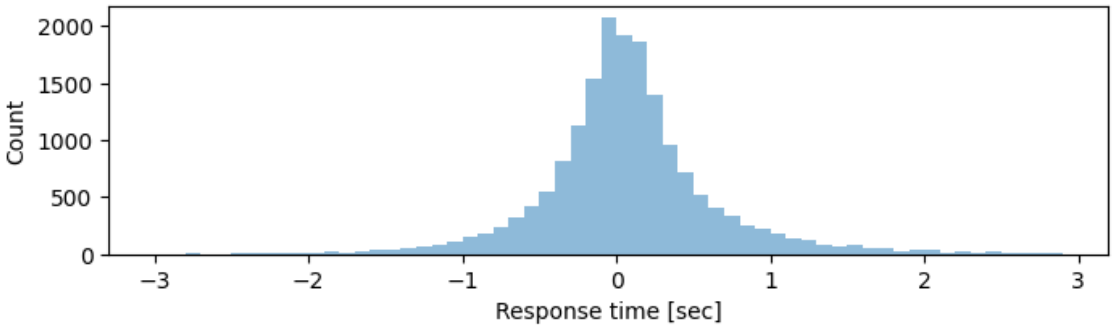

図 3.2: 分析対象話者の返答タイミングの分布

3.2.1 評価方法

本研究では，分析対象の話者から評価対象となる返答を抽出して聴取評価を行った．聴取評価のために，抽出された評価対象の返答 (Actual) にたいして返答タイミングを特定の秒数に置き換えた返答 (Replaced) を作成した．そして，各返答へ 17 名のアノテータが割り当てられるように実験参加者を集め，アノテータは Actual と Replaced の両方を聴き比べて，それらの自然さを評価した．本研究におけるアノテータは日本語を母国語としており, 自然言語処理や音声情報処理におけるアノテーションの経験がある者で構成されている．

先行研究 [1] に倣い，置換する秒数には分析対象の話者の返答全体の最頻値である 0 ミリ秒を利用した．返答単位で独立した評価を行うため，聴き比べる返答と，評価対象となる返答の時系列をシャッフルした．自然さを評価をするうえでは，アノテータは①両方の音声の自然さを 5 段階で採点し，②どちらの音声が実際の会話らしいかを選択した．

3.2.2 評価対象

本研究では，分析対象である 8 名の話者について，そのすべての対話者 (各 6 名) との対話における返答音声から評価対象としてそれぞれ 32 返答を抽出した．実験中の問題により一部の返答が除外されたが合計 1720 返答が収集された．これは分析対象の話者の返答全体の約 10%にあたる．本研究の目的に合わせて評価対象の返答を抽出するにあたり，以下の抽出方法を適用した．

多様な返答タイミングの収集 ：各話者のさまざまな返答タイミングを収集するため，返答を Early，Late，Medium の 3 つのタイプに分割した．各話者の応答時間の統計を計算し，30 パーセンタイルより小さい返答 (Early)，70 パーセンタイルより大きい返答 (Late)，および Early と Late の間 (Medium) として抽出した．各タイプごとに 4 返答を収集した．0 ミリ秒よりも速い返答 (オーバーラップ) のうち，発話を妨げる返答 [12] はほとんどなかった．

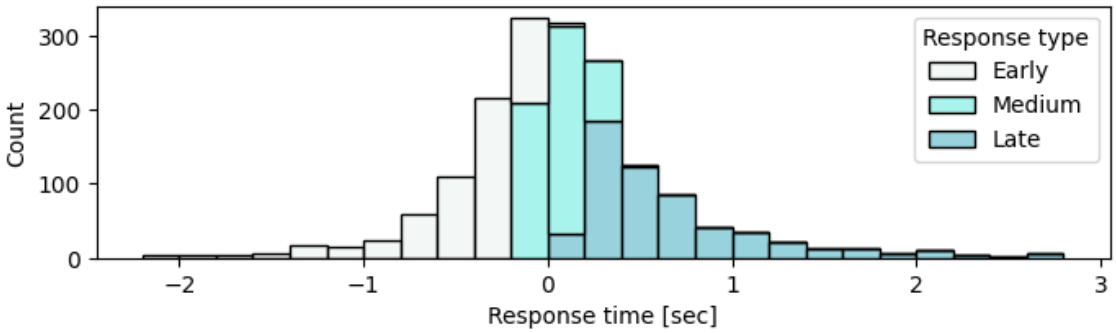

図 3.3: 評価対象である返答のタイミングの分布

極端な値の除外：−20 ミリ秒以降かつ 20 ミリ秒未満の返答タイミングは，0 ミリ秒に置き換えてもほとんど差がないと判断して除外した．加えて，極端な値を避けるため，0.1 パーセンタイル未満または 99.9 パーセンタイルを超える返答も除外した．

不自然な IPU の除外：IPU が不自然に中断されている等，音声品質以外の要因によって返答タイミングの変更が容易に認識される可能性があるものは除外した．

これらの抽出方法を実施して抽出された，評価対象となる返答の分布を図 3.3 に示す．この図と図 3.2 を比較すると，特に Medium の領域についてデータ量が少なくなっているが，他の領域の分布は大きく変化させないまま，さまざまな秒数の返答タイミングを抽出できていることがわかる．

3.2.3 評価スコア

評価結果の分析や予測モデルの構築のため，アノテータによる 2 種類の評価を 3 種類の評価スコアとして定義する．

Actual Response Naturalness Score (ARNS)：Actual の自然さを 5 段階で採点した数値の平均値．評価値は [0,5] の範囲となる．

Replaced Response Naturalness Score (RRNS)：Replaced の自然さを 5 段階で採点した数値の平均値．評価値は [0,5] の範囲となる．

Realness Score (RS)：Actual と Replaced を聴き比べたときに，実際の会話らしい音声として Actual が選ばれた場合に+1，Replaced が選ばれた場合に-1 した数値の平均値．評価値は [-1,1] の範囲となる．

ARNS と RRNS は返答音声そのものの自然さも含めた評価スコアであり，RS は返答音声の品質によらず返答音声を特定の固定値に置き換えられるかについての評価スコアである．また，本実験における 5 段階評価は MOS とは異なるため Naturalness Score と命名した．なぜなら MOS は絶対評価であり他の音声とは独立して評価するが，本実験では同一返答音声のタイミングを置き換えたものを同時に聴取したためか，評価結果に相対的な優劣が確認されたためである．本研究では，この 3 種類の評価スコアと返答タイミングの関係性を分析する．

第4章　データ分析

聴取評価を行なった評価対象のデータセットを分析し，音声の自然さと返答タイミングの関係性を考察する．本章では，返答タイミングと聴取評価における評価スコアの関係性について分析する．

4.1　返答音声の自然さの評価

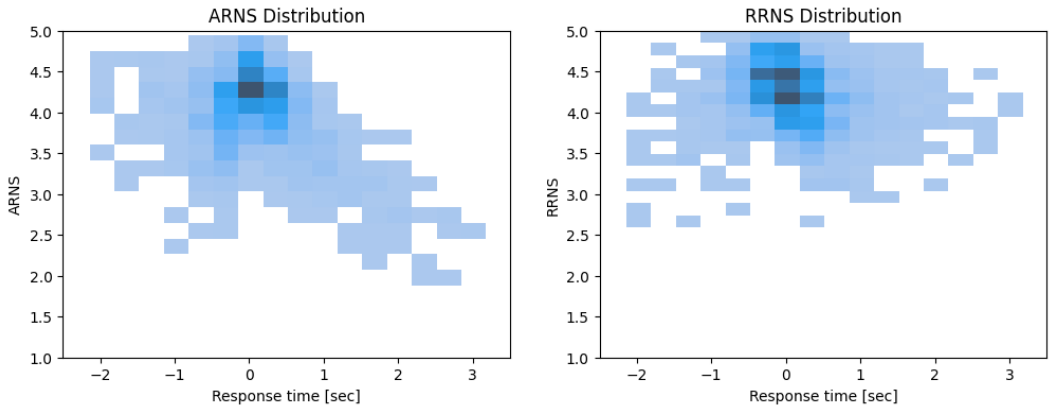

図 4.1: 返答タイミングごとの ARNS と RRNS の分布

まず音声の自然さの評価として ARNS と RRNS の実験結果を図 4.1，表 4.1 に示す．この図における相関係数は返答タイミングと ARNS(あるいは RRNS) にたいしてピアソンの相関係数で算出したものを表す．全体の傾向として RRNS のほうが ARNS よりも平均が高く，ARNS は返答時間が遅くなるほどスコアが低かった．また標準偏差は RRNS のほうが小さく，返答タイミングを固定したほうが分散が小さくなることが示された．また各返答における ARNS もしくは RRNS のどちら

	平均	標準偏差	最小	最大	相関係数 (返答タイミング)
ARNS	4.02	0.46	1.88	4.94	−0.34
RRNS	4.22	0.36	2.59	5.00	−0.15

表 4.1: ARNS と RRNS の統計情報

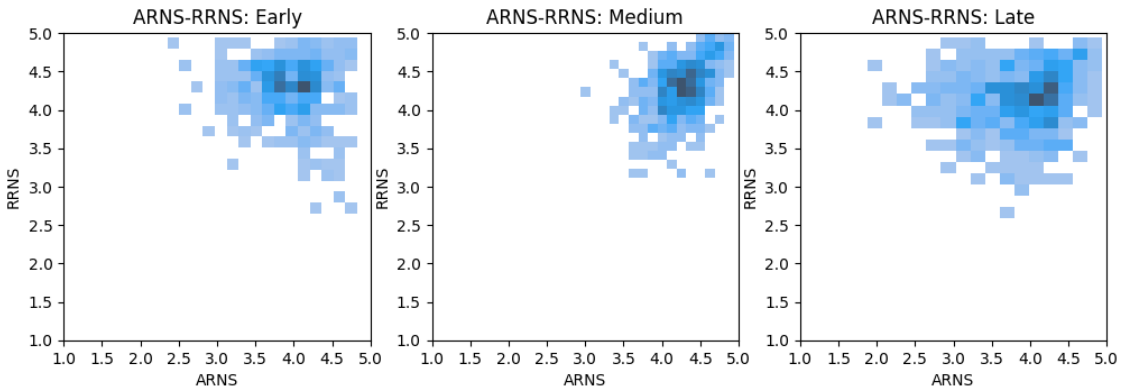

図 4.2: 返答タイプごとの返答タイミングの分布

か一方は少なくとも 3.5 以上であり，このスコアは聴取評価対象である音声において返答タイミングを考慮しないで評価した場合の評価水準を表していると言える．

4.1.1 評価スコアでの比較

返答タイプごとの ARNS と RRNS の関係性を図 4.2 に示す．ここでの相関係数は ARNS と RRNS にたいしてピアソンの相関係数で算出したものを表す．この図から Medium の分散が小さいことが示された．また表 4.2 に示すように Medium の ARNS と RRNS の相関係数は 0.43 であり強い相関があるとは言えず，Medium でも固定値に置き換えることで返答の評価に差が発生することが示された．他にも，この表から ARNS の標準偏差は Late が一番大きいことが示された．また RRNS の平均は Early が一番高く，この結果から，速い返答タイミングである音声は 0 ミリ秒に置き換えたときに他の返答タイプと比較して最も自然に聴こえる，あるいは相対的に置き換えによる自然さが向上する度合いが高い可能性が示された．

返答タイプ	ARNS		RRNS		相関係数
	平均	標準偏差	平均	標準偏差	(ARNS,RRNS)
Early	3.94	0.41	**4.28**	0.35	−0.06
Medium	4.27	0.30	4.25	0.34	**0.43**
Late	3.86	**0.53**	4.13	0.38	0.12

表 4.2: 返答タイプごとの ARNS と RRNS の統計情報

4.2 返答タイミングの評価

次に，Actual と Replaced のどちらの返答タイミングが適しているかを調べるために RS を分析する．RS の返答タイミングごとの結果を図 4.3 に示す．この図に

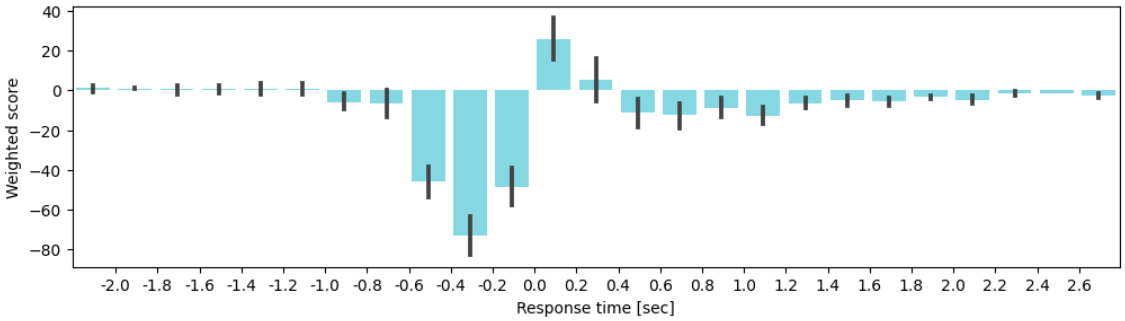

図 4.3: 返答タイミングごとに RS と返答数を乗算した加重平均の分布

おける各バーの垂直線は加重平均からの標準偏差を表す．この図は返答タイミングごとに RS と返答数を乗算した加重平均であり，もし RS がすべて 1 であれば返答タイミングの分布 (図 3.3) と同じ形状になる．よって，この図から，全体の傾向としてアノテータは 0 ミリ秒から 400 ミリ秒までは Actual を多く選択し，-1000 ミリ秒から 0 ミリ秒までは Replaced を多く選択したことが示された．

返答タイプごとの RS の統計情報を表 4.3 に示す．この表から，最大の RS は 0.88 であり，すべてのアノテーターが Actual を選択した返答が存在しなかったことが示された．また Early と Late の平均が負であり，アノテーターがこの 2 つのタイプで Replaced を多く選択したことが示された．そして，Early と Late の間での標準偏差の差は 0.002 程度であり，本研究においては Early と Late の間で Actual と Replaced のどちらを選択するかのバラつきにほとんど差がなかったことが示された．

	平均	標準偏差	最小	最大
Early	−0.25	0.38	−1.00	0.88
Medium	0.02	0.32	−0.88	0.88
Late	−0.14	0.38	−1.00	0.77
合計	−0.13	0.37	−1.00	0.88

表 4.3: 返答タイプごとの RS の統計情報

4.2.1 有意差検定

中程度の RS にはランダムな選択による誤差が含まれるため，RS にたいして統計検定を実行し，Actual と Replaced の間でどういった返答に統計的に有意に回答としての選択されやすさ (Statistically Significant Preferences(SSP) と呼ぶ) が現れるかを確認した．検定にはカイ二乗検定を使用し，各返答音声においてアノテータの Actual あるいは Replaced への選択頻度に偏りがないかを p<0.05 レベルで確

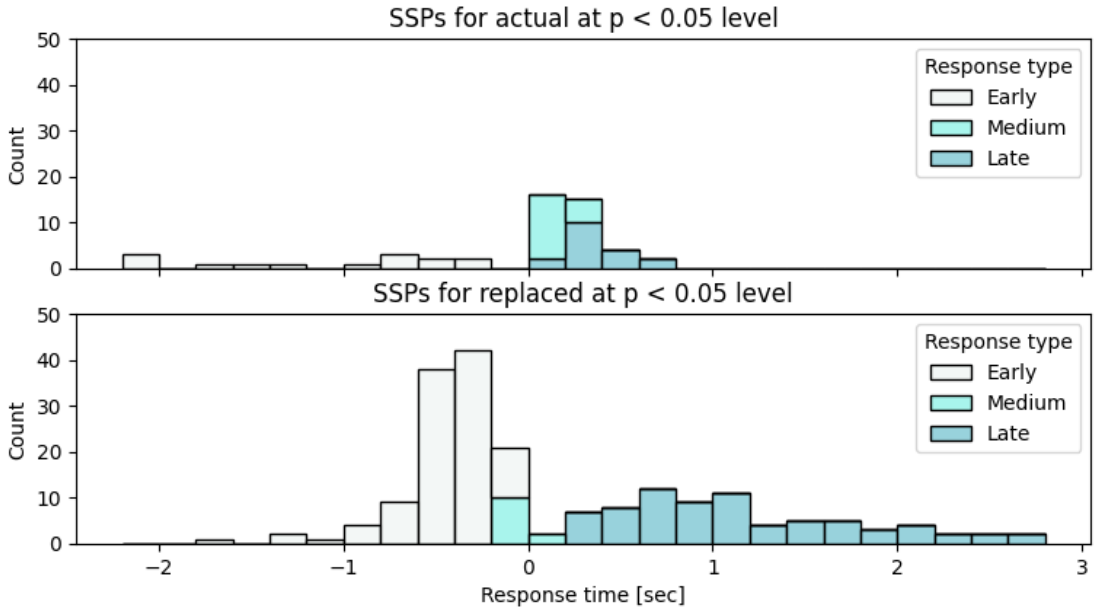

図 4.4: 返答タイミングごとの SSP に現れた返答の件数の分布

かめた．その結果，返答音声のうち 52 件の返答に Actual への SSP が，193 件の返答に Replaced への SSP が現れた．Actual に SSP が現れた音声のうち，最も置き換えた返答タイミング (0 ミリ秒) に近い返答時間は 60 ミリ秒であり最も遠いものは−2140 ミリ秒であった．Replaced に SSP が現れた音声のうち，最も置き換えた返答タイミング (0 ミリ秒) に近い返答は−20 ミリ秒であり，最も遠いものは 2800 ミリ秒であった．

SSP の現れた返答の件数を表 4.4 に示す．この表における Total (%) は評価対象となる返答全体における割合を表す．返答タイプの Early が最も Actual に有意な回答としての選択されやすさのある返答が少なく，最も Replaced に有意な回答としての選択されやすさのある返答が多かった．また，返答時間ごとの SSP の分布を図 4.4 に示す．Medium において Actual に有意な回答としての選択されやすさがあった返答はすべて正の値だった．聴取評価の返答全体として，0 ミリ秒よりも速い，つまりオーバーラップのある返答時間の返答はそれ以外よりも Actual で SSP が現れた返答が少なく，Replaced に SSP が現れた返答が多かった．これらの結果から，音声対話においてオーバーラップのある返答は人間には Replaced(ある

SSP (at p<0.05)	返答タイプ			Total (%)
	Early	Medium	Late	
Actual	14	19	19	52 (3.0)
Replaced	107	12	74	193 (11.2)

表 4.4: 聴取評価の結果における SSP の現れた返答の件数

いは 0 ミリ秒から 400 ミリ秒の返答) よりも不適切だと感じられやすい可能性が示された.

第5章　スコア予測モデル

本章では，返答タイミングにおける人間の感覚を自動評価し，音声対話システムへ応用することを目指して，本研究における返答タイミングの評価スコア (ARNS, RRNS，RS) を予測するモデルを構築し，それらの予測精度を検証した．本研究における評価スコア (ARNS，RRNS，RS) を予測するモデルが構築された先行研究は存在しないため，あくまでベースラインモデルとしてその性能を検証し，詳細なハイパーパラメータチューニングは実施しない．また，どういった特徴量や評価スコアだとどのような結果になるのかを明確にするため，同一モデルを用いて特徴量と評価スコアを切り替えて評価を実施する．

5.1　モデルの構築

VoiceMOS Challenge の先行研究 [13] から音声を入力特徴量としたスコア予測に Self-Supervised モデルが有用であることがわかっている．Self-Supervised 学習フレームワークは大量の音声特徴を学習するものであり，本研究ではこれを 5 音声特徴抽出器として使用した．そしてベースラインモデルとして，その Self-Supervised モデルに Projection 層を追加したシンプルなモデルを用意した．Projection 層には活性化関数として ReLU[14] を採用した 1536 次元の全結合層とスコア出力のために 1 次元の全結合層を利用した．データセットは日本語であるため，日本語音声で事前学習された公開モデルである rinna Co., Ltd. の HuBERT[15] https://huggingface.co/rinna/japanese-hubert-base を利用した．

入力特徴量は先行研究 [13] に倣って音声のみを利用する．本実験ではモデルの性能やデータ量は同一にして，入力特徴量や評価対象を別にすることで比較を行う．またデータセットの特性上，ある会話では返答者であった話者が別の会話では発話者となることと，本研究では音声対話システムの返答タイミングに着目しており，システムは常に返答者であると想定することから，発話者と返答者の両方を学習データとする実験は本研究の対象外とする．音声には返答タイミングは抽出せずに返答者あるいは発話者の音声のみを入力する．データセットの特性から，単一話者の音声は返答音声のほうが多いため，返答音声による予測精度を重点的に分析する．

5.1.1 モデルの学習

モデルの学習は k-fold クロスバリデーション方式で，各 fold における Out-Of-Fold (OOF) の予測スコアを利用する．本実験ではベースラインとしてすべてのモデルを同一の学習方法で精度を検証する．クロスバリデーションの k には 6 を採用し，各 fold に含まれる話者データと対話者データの件数を均等にする．オプティマイザには Adam を使用する．強力な正則化をすると学習結果が過度に平均的な出力となってしまうため，learning rate は 1e-06，weight decay は 1e-05 とした．epoch 数は 70 であり，スケジューラは使用しなかった．損失関数には L1 loss を使用した．またクロスバリデーションは 5 seed で行い，評価には各 seed で算出された評価スコアの平均値を用いた．

5.2 返答音声の自然さを予測するモデルの評価

返答音声の自然さを加味した自動音声評価モデルの構築を目指して，ARNS, RRNS を予測するモデルを構築する．聴取評価から RRNS のほうが平均値が高く標準偏差も低かったため，RRNS の予測精度を重点的に分析する．具体的には，まず返答音声を入力特徴量として ARNS と RRNS の予測モデルによる予測スコアと評価スコアの誤差を分析し，返答タイミングが固定されている RRNS と固定されていない ARNS でどれほどの予測精度の差が出るのかを分析する．その後，入力特徴量を発話音声と返答音声の 2 種類で RRNS の予測モデルを学習し，発話音声からどれほど返答の自然さが予測できるのかを分析した．

5.2.1 評価スコアでの比較

まずは返答音声を入力とした場合の評価を実施する．比較対象として ARNS の予測モデルも構築し，RRNS との精度の違いを確認する．前述のモデルの構築と学習を実施し，それぞれの予測結果を図 5.1 と表 5.1 に示す．本研究における ARNS と RRNS の評価指標は VoiceMOS Challenge[13] に倣い，平均二乗誤差 (MSE)，ピアソンの相関係数 (LCC)，スピアマンの順位相関係数 (SRCC)，ケンドールの順位相関係数 (KTAU) とした．この実験では返答タイミングは抽出せずに返答音声のみを入力としたため，返答タイミングが発話によって異なる ARNS とすべて同じ返答タイミングである RRNS の予測精度に差が生まれた．よって，返答音声の自然さの評価にはその返答タイミングの違いも考慮してモデルを構築する必要性を示唆している．また，図 5.1 から RRNS の予測モデルにおいて高いスコアの予測精度が示された．

今回の結果を先行研究 [13] における VoiceMOS の予測モデルの精度と比較すると，特に低いスコアの予測精度が低いことがわかる．この精度の違いが生まれる要因として，単純にデータ量が 3.6 倍ほどあることや VoiceMOS の学習データの標

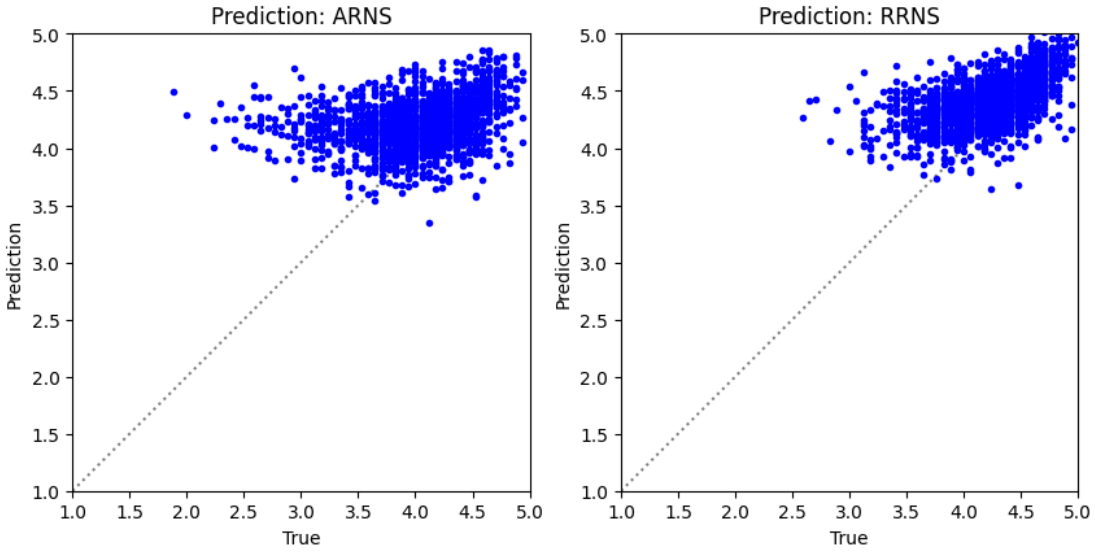

図 5.1: ARNS と RRNS の予測モデルによる予測スコア (Prediction) とアノテータによる評価スコア (True) の分布

準偏差は RRNS の約 2.5 倍ほどありスケールに幅があったこと，あるいは返答の自然さには文脈情報も加味する必要があるかもしれない等が考えられ，今後の分析が求められる．

	MSE	LCC	SRCC	KTAU
ARNS	0.245	0.242	0.273	0.189
RRNS	0.134	0.513	0.521	0.376

表 5.1: RRNS の予測モデルの精度

一方で，本研究における RRNS では MOS のように絶対的なスコアではなく相対的であるため，その相対誤差が予測精度を下げている可能性を分析する．評価スコアの相対誤差と予測精度の相関を確認するため，学習した各返答について，予測されたスコアから実際の評価スコアを引いた絶対誤差と，RRNS から ARNS を引いた絶対誤差との相関係数を計算する．この 2 つの値の相関を計算することで相対的に評価に差のあった返答と予測に大きく失敗していた返答の相関が計算できる．計算にはピアソンの相関係数を用いる．その結果，ARNS の予測スコアにおける相関は 0.697，RRNS の予測スコアにおける相関は 0.148 となった．つまり ARNS の予測には返答タイミングの評価における相対誤差が相関していた一方で，RRNS の予測スコアに大きな相関はなく，全体としては評価スコアの相対誤差に大きな影響を受けていない予測結果であることが示された．

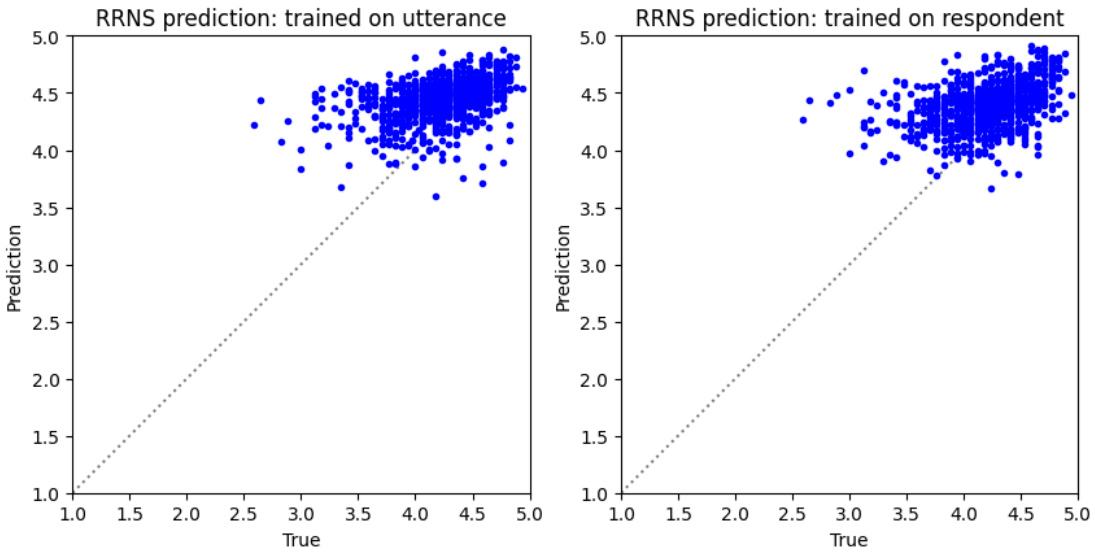

図 5.2: RRNS の予測モデルの特徴量として発話者の音声 (utterance) と返答者の音声 (respondent) を入力した際の予測スコア (prediction) とアノテータによる評価スコア (true) の分布

5.2.2 入力特徴量での比較

次に，同様のモデルの構築方法で入力を発話者の音声，あるいは返答者の音声とすることで入力特徴量を発話者とすることでどれほど精度が向上するのかを分析する．ここで，発話者と返答者の音声に含まれる人数は異なる (返答者は友人グループから選ばれるが，発話者は友人と初対面の両方が含まれる) ため，学習データは友人同士の会話データに絞り，全体の半数のデータで学習を実施した．データ数に合わせて epoch 数は 140 とした．

予測結果を図 5.2 と表 5.2 に示す．返答者の音声を学習したほうが精度は高かったが，ARNS と RRNS ほどの大きな精度の差はなく，RRNS の予測モデルにおいて発話音声と返答音声の特徴量に大きな差はないことが示された．この予測モデルを音声対話システムの返答へ応用することを想定すると，返答にたいする任意のユーザーの発話で RRNS の予測モデルを機能させる必要があり，発話音声の話者数は本研究のようにまとまった人数で取得することはできず，話者の特性やデータ量に偏りがあったりノイズが含まれる可能性が高い．したがって，予測精度が

特徴量	学習対象	学習量	epoch 数	MSE	LCC	SRCC	KTAU
返答音声	友人	半数	140	0.150	0.385	0.410	0.290
発話音声	友人	半数	140	0.166	0.383	0.403	0.285

表 5.2: RRNS の予測モデルの入力特徴量を返答者と発話者の音声とした際の精度比較

同程度なのであれば，RRNS の予測モデルを構築するうえでは発話音声よりも返答音声を学習するほうが適していると言える．

5.3 返答タイミングの適切さを予測するモデルの評価

返答タイミングの適切さを自動的に予測するモデルを目指して，RS を予測するモデルを構築し，その予測精度を評価する．RS は聴取評価でアノテータが Actual と Replaced を聴き比べたときに， 実際の会話らしい音声として Actual が選ばれた場合に+1 ，Replaced が選ばれた場合に-1 した数値の平均値であるため，0 に近い中程度の RS には選択においてランダム性が含まれる．よって，本実験における RS の予測モデルはスコアそのものの予測ではなく，前章における有意差検定において Actual と Replaced の間に統計的に有意な回答としての選択されやすさ (Statistically Significant Preferences(SSP) と呼ぶ) が現れた発話を識別できるかどうかを予測精度を比較するうえでの評価対象とした．具体的には，Actual に SSP が現れた返答は置換した秒数へ返答タイミングを置き換えられないもの，Replaced に SSP が現れた返答は置き換えられるものと定義し，その両方の返答を RS の予測スコアを利用して正しく分類できた割合を評価する．

評価のために，Actual の SSP の再現率 (R_a) と，Replaced の SSP の再現率 (R_r)，およびそれらのバランス精度 (Balanced Accuracy(BA) と呼ぶ) の 3 つを評価指標として定義した．これらは次のような数式で表される．

$$R_a = \frac{T_a}{T_a + F_r} \qquad R_r = \frac{T_r}{T_r + F_a} \qquad BA = \frac{1}{2}\left(R_a + R_r\right) \tag{5.1}$$

where:

- $T_{a/r}$ は Actual/Replaced に SSP が現れた返答において，返答タイミングを固定値に置き換えが可能/不可能であると予測された返答の数
- $F_{a/r}$ は Replaced/Actual に SSP が現れた返答において，返答タイミングを固定値に置き換えが可能/不可能であると予測された返答の数

これらの評価指標において返答タイミングを固定値に置き換えが可能/不可能であることを予測するうえで，予測スコアの閾値には，実際の RS の平均を設定する．具体的には，予測スコアが RS の平均を下回った場合は Replaced が選択されたとして置き換えが可能，上回った場合は Actual が選択されたとして置き換えが不可能だと予測されたと判断する．

さらに，ROC 曲線と AUC を使用して各予測スコアの閾値におけるパフォーマンスを評価するため，Actual が正しく選択された割合 (True Actual Rate(TAR) と呼ぶ) と Actual が正しく選択されなかった割合 (False Actual Rate(FAR) と呼ぶ) を定義した．

$$TAR = \frac{T_a}{T_a + F_r} \qquad FAR = \frac{F_a}{F_a + T_r} \tag{5.2}$$

加えて，RS の比較において利用する返答タイミングごとの Actual と Replaced に有意な回答としての選択されやすさ (SSP) のある返答のデータ量は限られるため，RS の予測モデルの比較は ARNS と RRNS とは別の方法で実施する．

5.3.1 返答音声を入力特徴量とした際の比較

RS の予測モデルを評価するうえで，まず返答音声においてデータ量が増加するとどのように精度が変化するのかを分析する．データ量を比較するうえで，学習データを発話者が返答者の友人である場合と友人あるいは初対面である場合 (すなわち全評価対象データ) の 2 種類に区別して RS の予測モデルの評価指標の比較を実施する．そのうえでまず SSP が現れたデータの件数を表 5.3.1 に示す．この表における Total (%) は評価対象となる返答全体における割合を表す．発話者が返答者の友人である返答は全評価対象データの半数であるが，この表から，SSP について Actual に現れる SSP の件数は全体の約 60%，Replaced に現れる SSP の件数は全体の約 50%であることがわかる．つまり，評価対象となる返答の数においても，友人のデータで学習した場合は両方で学習した場合の半数程度となる．また学習データ量を考慮し，友人のデータで学習した場合の epoch 数は 140 とする．

発話者の区分	SSP (at p<0.05)		Total (%)
	Actual	Replaced	
友人	32	97	129 (7.5)
友人+初対面	51	193	244 (14.2)

表 5.3: 返答タイミングごとの SSP の現れた返答について，発話者が返答者にとって友人なのか初対面なのかを区別して集計した際の件数

返答音声を入力特徴量として RS の予測モデルを学習した際の評価結果を図 5.3 と表 5.5 に示す．これらの結果から，返答音声を入力特徴量にした場合は学習データを増やしたほうが予測精度が向上することが示された．また AUC が 0.636 であるため，ランダムに返答を置換された秒数に置き換えられるかどうかを選択するよりも精度が高いことが示された．また，学習データ量を増やすと R_r よりも R_a の精度が高くなっており，返答者の特徴から返答タイミングを置換された秒数に置き換えられるべきではない音声を予測する性能が置き換えられる音声を予測する精度よりも高いことが示された．

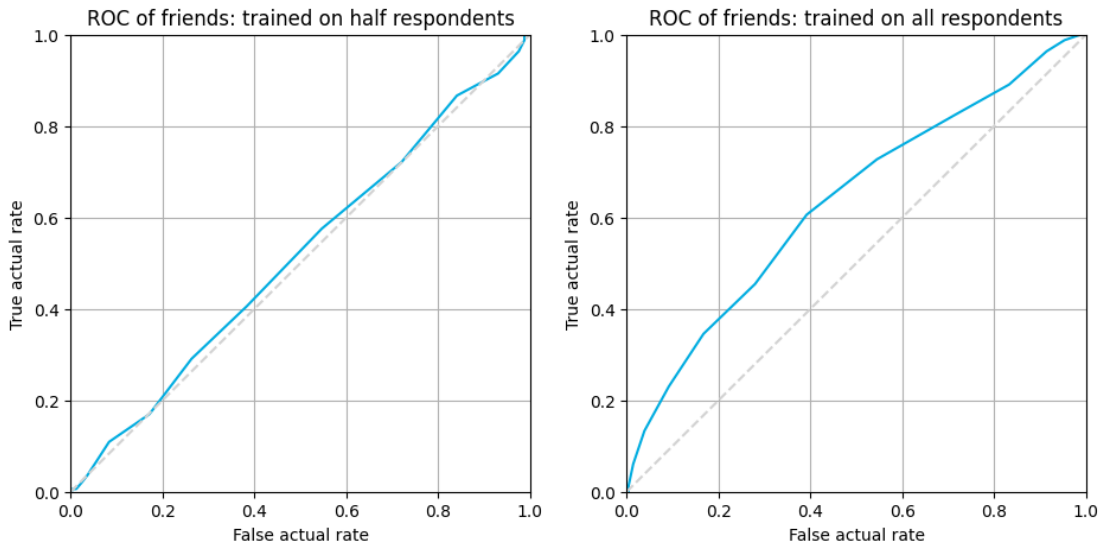

図 5.3: 返答音声を入力特徴量として，発話者が返答者の友人であった際の RS の予測モデルの予測精度について，友人のみのデータ (half) とすべての発話者のデータ (all) で学習した場合の評価指標の比較

学習量	epoch 数	AUC
半数	140	0.519
全体	70	**0.636**

表 5.4: 返答音声を入力特徴量として，RS の予測モデルの学習対象を区別して予測精度を算出した際の評価指標の比較

特徴量	学習対象	学習量	epoch 数	R_a	R_r	BA	AUC
返答音声	友人	半数	140	0.473	0.576	0.523	0.519
	友人	全体	70	0.642	0.584	0.613	**0.636**

表 5.5: 返答音声を入力特徴量として，RS の予測モデルの学習対象を区別して予測精度を算出した際の評価指標の比較

5.3.2 発話音声を入力特徴量とした際の比較

次に，同様に発話者の音声でも学習を実施する．前述のとおり，返答音声を入力特徴量とする際に，友人のデータのみで学習した場合の RS の予測精度が著しく低かったため，比較対象として発話者の友人データで学習したものだけでなく，友人と初対面の両方で学習したものも比較対象とする．ただし，前述の通り，データセットの特性として，発話者と返答者の音声に含まれる人数は異なる (返答者は友人グループから選ばれるが，発話者は友人と初対面の両方が含まれる) ため，発話者と返答者の音声を学習データとして比較することは厳密な比較とはならない

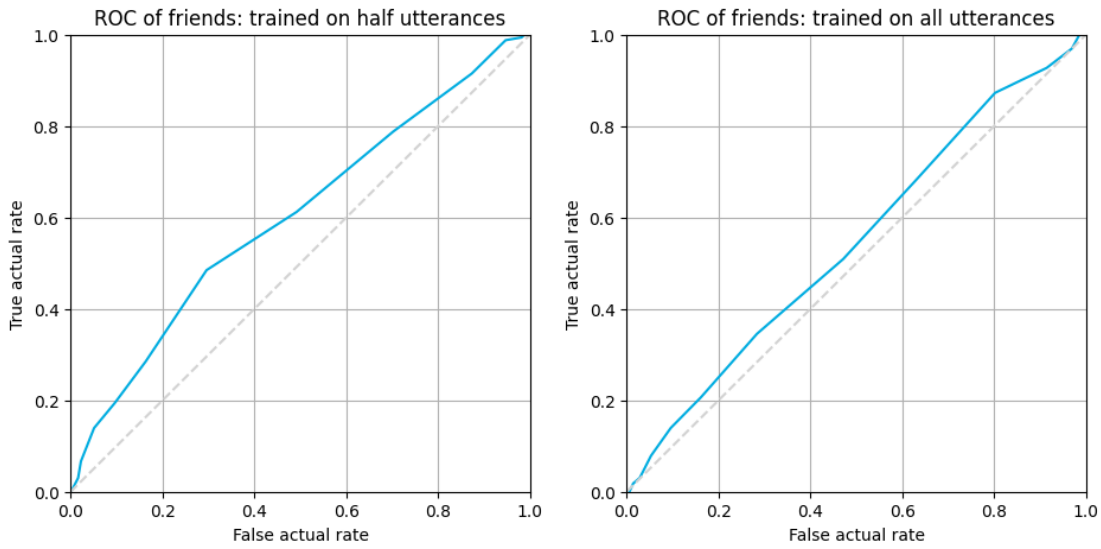

図 5.4: 発話音声を入力特徴量として，発話者が返答者の友人であった際の RS の予測モデルの予測精度について，友人のみのデータ (half) とすべての発話者のデータ (all) で学習した場合の評価指標の比較

が，学習データの合計を一致させることを目的として比較を行う．

発話音声を入力特徴量として RS の予測モデルを学習した際の評価結果を図 5.6 と表 5.4 に示す．返答音声を入力特徴量として学習したときとは異なり，学習データ量を増やしたことで R_a が下がった．この結果から，学習データ量を増やすあるいは評価対象の話者を増やすと返答タイミングを置換された秒数に置き換えられるべきではない音声の予測精度が低下する可能性が示された．また，R_r の精度は学習データ量を増やしても低下しておらず，発話音声を入力特徴量としたとき，返答タイミングを置換された秒数に置き換えられる音声の予測精度が置き換えられない音声を予測する精度よりも高いことが示された．

特徴量	学習対象	学習量	epoch 数	R_a	R_r	BA	AUC
発話音声	友人	半数	140	0.515	0.641	0.578	**0.601**
	友人+初対面	全体	70	0.388	0.647	0.517	0.540

表 5.6: 発話音声を入力特徴量として，RS の予測モデルの学習対象を区別して予測精度を算出した際の評価指標の比較

5.3.3 追加実験: 初対面の発話者における予測精度

前述のとおり，入力特徴量を返答音声あるいは発話音声とした場合で予測精度を検証したが，返答音声と発話音声での比較結果に大きな違いが出たため，友人

との会話だけでなく初対面との会話についても分析を実施した．SSP の件数を表 5.3.3 に示す．この表における Total (%) は評価対象となる返答全体における割合を表す．発話者が返答者にとって初対面である返答は全評価対象データの半数であるが，この表から，SSP について Actual に現れる SSP の件数は全体の約 60%，Replaced に現れる SSP の件数は全体の約 50%であることがわかる．つまり，評価対象となる返答の数は，初対面のデータ学習した場合は両方で学習した場合の約 40%程度となる．また学習データ量を考慮して，初対面のデータで学習した場合の epoch 数は 140 とする．

発話者の区分	SSP (at p<0.05)		Total (%)
	Actual	Replaced	
初対面	19	96	115 (6.7)
友人+初対面	51	193	244 (14.2)

表 5.7: 返答タイミングごとの SSP が現れた返答について，発話者が返答者にとって友人なのか初対面なのかを区別して集計した際の件数

初対面との会話を入力特徴量として RS の予測モデルを学習した際の評価結果を図 5.5 と表 5.8 に示す．これらの結果から，返答音声を入力特徴量とした場合には R_a の予測精度の向上が共通して示された一方で，発話音声を入力特徴量とした場合にも R_a の予測精度の向上が示されており，この結果は友人との会話を入力特徴量としたときの結果とは異なる．したがって，本実験における発話音声を入力特徴量とした実験において，友人と初対面の発話音声が混在すること，あるいは学習データ量を増やしたことで予測精度が低下した可能性が示唆される．この予測モデルを音声対話システムの返答へ応用することを想定すると，返答にたいする任意のユーザーの発話で RS の予測モデルを機能させる必要があり発話音声の話者が混在することは容易に想定される．したがって，発話音声を用いた RS の予測モデルを構築するうえでは複数の発話者を学習しても予測精度が落ちないモデルの構築が今後の課題として挙げられる．

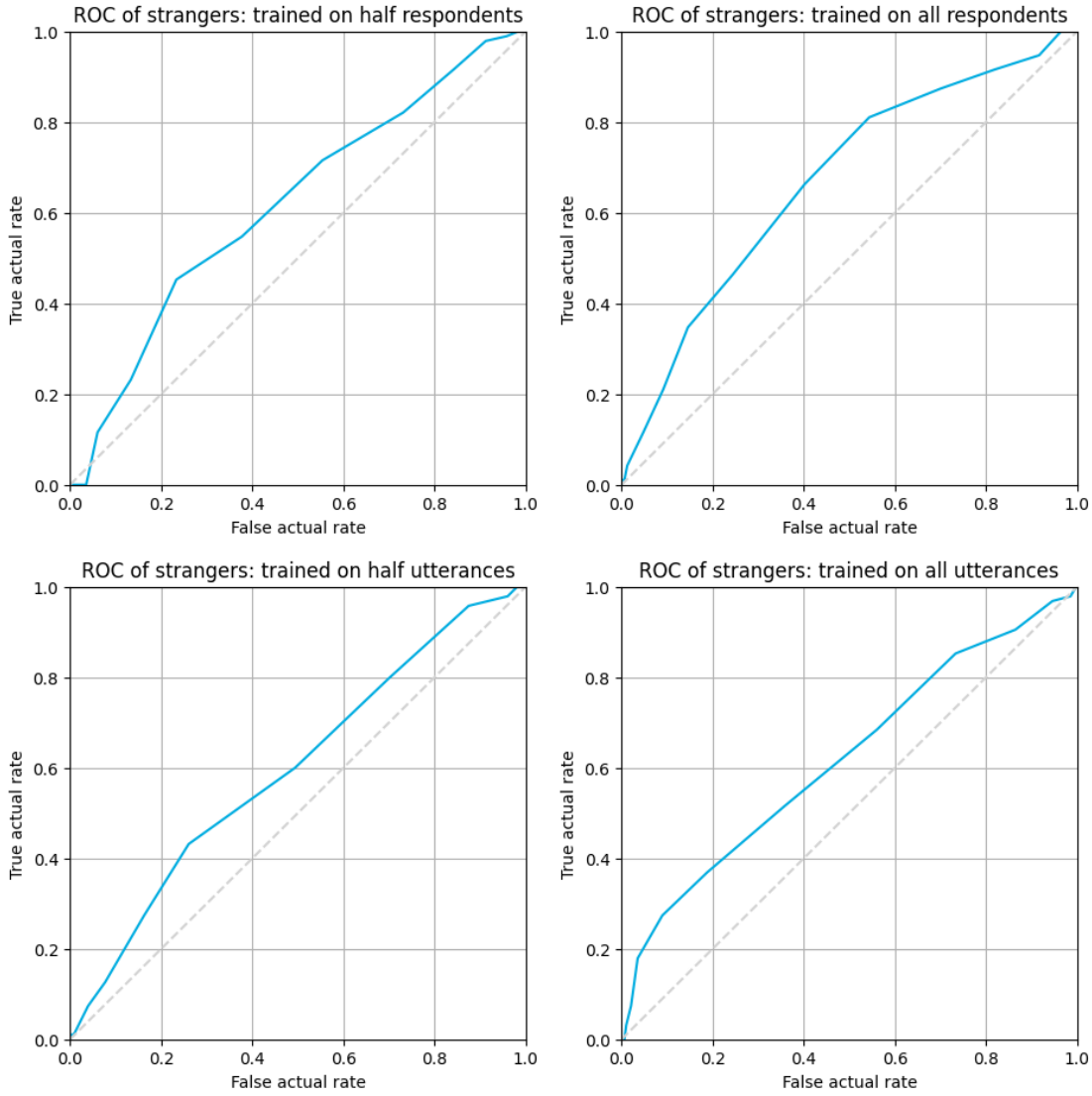

図 5.5: 発話者が返答者にとって初対面であった際の RS の予測モデルの予測精度について，入力特徴量を返答音声 (respondents) と発話音声 (utterances) にした場合と，初対面のみのデータ (half) とすべての発話者のデータ (all) で学習した場合の評価指標の比較

特徴量	学習対象	学習量	epoch 数	R_a	R_r	BA	AUC
返答音声	友人	半数	140	0.611	0.581	0.596	0.624
	友人	全体	70	0.716	0.550	0.633	0.671
発話音声	初対面	半数	140	0.463	0.677	0.570	0.598
	友人+初対面	全体	70	0.547	0.602	0.575	0.620

表 5.8: 発話者と返答者が初対面の会話を対象として，RS の予測モデルの特徴量と学習対象を区別して予測精度を算出した際の評価指標の比較

第6章　今後の課題

6.1　データ分析

本研究では，先行研究 [1] を拡張して返答タイミングを最頻値に置き換えたときの人間の返答タイミングの評価がどれほど変化するかを分析したが，最頻値はあくまで一例であり，他の値における検証が必要である．今回収集した RS の返答時間ごとの平均においては，0 ミリ秒から 400 ミリ秒においては正の数，つまり実際の返答のほうが適切な返答タイミングと感じたという回答が多かったため，日常会話における適切な返答タイミングを分析するうえで，この範囲のうちのいずれかの秒数で返答タイミングを置き換えたときに本研究における結果とどれほど異なるのかを分析する必要がある．加えて，本研究ではあくまで返答タイミングの違いに着目して分析を行ったが，返答時間は返答内容によっても異なること [6] が明らかであるため，本研究におけるデータセットから返答内容を分類し，どのような返答内容においてどのような返答タイミングが適切であるかを分析する必要がある．

分析に利用したデータの収集方法について，本研究における聴取評価はアノテータにたいして会話のトピックや前後の文脈情報を共有せずに発話と返答のみで評価を実施していた．よって，文脈情報を含めたときにどのように返答タイミングの評価が変わるのかを調べる必要がある．また一方で，音声通話と Face-to ー Face の対話だとオーバーラップ頻度が異なることが知られているため [12]，動画や VR を用いたマルチモーダルな加工による実証実験では新たな実験結果やモデル性能が明らかになるかもしれない．加えて，言語による返答時間に差があること [6] は明らかであるため，日本語以外における研究も必要となる．

6.2　スコア予測モデル

本実験の結果から，RRNS の予測精度が高いスコアにたいして高く，低いスコアに対して低いことが明らかになった．一方でなぜ低いスコアに対しての予測精度が低くなっているのかは明らかではなく，原因分析が求められる．加えて，今回のスコアである ARNS と RRNS はあくまで相対的なスコアであり，直接的に MOS との比較を実施できていない．本研究においては，返答を置き換えたときにどれほど差があったのかを分析するうえでスコアを定義したが，MOS など既存の評価

指標と今回のような返答タイミングを加味した評価指標を直接比較できるフレームワークを用意して評価を実施する必要がある．

RS の予測モデルにおいては，今回の実験に用いたデータセットでは発話者の音声を使った予測モデルの性能を十分に検証することができなかった．具体的には，学習に含まれる発話者の数が増えたときに予測精度がどのようになるのかについての検証が求められる．加えて，音声対話システムへの応用を考えたとき，発話者の音声にはノイズが含まれるリスクが大きい．よって，発話者の数が増えたときに精度が向上したとしてもノイズが含まれる音声であってもロバスト性を持って予測精度を維持できるのかどうかを検証する必要がある．

6.3 音声対話システムへの応用

今回構築した予測モデルは最頻値 (0 ミリ秒) に置き換えられるかを元に構築されており，音声対話システムの返答あるいは VAD による発話区間終了の検出には一定のレイテンシが発生するため，現在あるいは過去の発話区間を入力として未来の発話区間を予測する Voice Activity Projection[16] が目指しているような未来の発話終了を事前に予測できるモデルがない限り音声対話システムへそのまま応用することはできない．加えて，たとえ返答タイミングを予測するモデルが正しく予測を実現していたとしても，その返答タイミングが適切かどうかを決定するのは音声対話システムと対話するユーザーであり，そのユーザーがどう感じるのかを実際に検証して効果が立証されなければ，音声対話システムへ応用が実現できたとは言えない．よって，RRNS や RS をどのように音声対話システムへ応用するかだけでなく，実際に音声対話システムへ応用したときにどのようにユーザーの会話にどのような影響を与えるのかを検証する必要がある．

また返答タイミングは発話区間終了のアノテーションによって定義されるため，音声対話システムに導入する VAD ツールの精度も重要である．発話区間終了の検出はアノテータ (VAD ツールを含む) によって異なるため，学習データと音声対話システムの発話区間終了の検出には同じ VAD ツールを使用することを推奨する．加えて，VAD ツールの数十ミリ秒あるいは 100 ミリ秒単位での予測精度が返答タイミング予測精度にも直結するため，音声対話システムで実現したい返答タイミングの精度に合わせて VAD ツールを選択することが重要であり，その精度の差によってどれほどユーザーとの会話に変化を与えるのかは検証する余地がある．

第7章　おわりに

本研究では，人間同士の会話から約 1,700 返答を抽出し，各返答にたいして 17 人のアノテータを集めた聴取評価により，どういった返答タイミングの返答を最頻値に置き換えれば人間にとって適切な応答タイミングとなるかを分析し，簡単なベースラインモデルでそれらを予測できる可能性を示した．データセットを作成する過程におけるフィルタリング処理により観測できないパターンがいくつかあったが，今回の研究で利用したコーパスが静かな環境で収録されていたため除外されたパターンは少なく，話者の返答パターンはほぼカバーされていた．一方で，返答タイミングを人間がどう感じるのかについての先行研究はターンテイキング研究と比較すると非常に少なく，研究すべき課題は多い．また，今回はあくまで予測モデルの検証に留まったが，本研究にて提案した RRNS や RS の予測モデルは音声対話システムへの応用を目指している．これらのモデルは従来の返答タイミング推定モデルや他の音声合成モデルとともに利用することも考えられる．したがって，スコア予測モデルに限らず，他のモジュールと組み合わせた音声対話システムへの実装を進めることが重要である．加えて，人間にとって快適な音声対話システムを構築するにはそのユーザーの評価が不可欠である．そのうえで返答タイミングの快適さをどのように評価すべきかも検討すべきであり，人間にとって快適な音声対話システムの実現への道のりは遠い．しかし，返答タイミングを人間に快適なように制御できる音声対話システムの実現は，人間と AI の共生社会において人間と AI のコミュニケーション領域を拡張し，そのコミュニケーションを豊かにするだろう．快適な返答タイミングの音声対話システムの実現のために，今後の研究が求められる．

参考文献

[1] M. Roddy and N. Harte, "Neural Generation of Dialogue Response Timings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jul. 2020, pp. 2442–2452.

[2] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The voicemos challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2023.

[3] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari, "Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection," *ArXiv*, vol. abs/2210.14850, 2022.

[4] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Interspeech 2021*, Aug. 2021, pp. 2127–2131.

[5] 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.

[6] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K.-E. Yoon, and S. C. Levinson, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, Jun. 2009.

[7] K. H. Kendrick and F. Torreira, "The Timing and Construction of Preference: A Quantitative Study," *Discourse Processes*, vol. 52, no. 4, pp. 255–289, May 2015.

[8] F. Roberts and A. L. Francis, "Identifying a temporal threshold of tolerance for silent gaps after requests," *The Journal of the Acoustical Society of*

America, vol. 133, no. 6, pp. EL471–EL477, 05 2013. [Online]. Available: https://doi.org/10.1121/1.4802900

[9] A. Raux and M. Eskenazi, "Optimizing the turn-taking behavior of task-oriented spoken dialog systems," *ACM Transactions on Speech and Language Processing*, vol. 9, no. 1, pp. 1:1–1:23, 2012.

[10] J. Sakuma, S. Fujie, and T. Kobayashi, "Response Timing Estimation for Spoken Dialog Systems Based on Syntactic Completeness Prediction," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, Jan. 2023, pp. 369–374.

[11] T. Hayashi, C. O. Mawalim, R. Ishii, A. Morikawa, A. Fukayama, T. Nakamura, and S. Okada, "A Ranking Model for Evaluation of Conversation Partners Based on Rapport Levels," *IEEE Access*, vol. 11, pp. 73 024–73 035, 2023.

[12] G. Skantze, "Turn-taking in Conversational Systems and Human-Robot Interaction: A Review," *Computer Speech & Language*, vol. 67, p. 101178, May 2021.

[13] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2021.

[14] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML'10. Madison, WI, USA: Omnipress, 2010, pp. 807–814.

[15] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. rahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[16] E. Ekstedt and G. Skantze, "Voice Activity Projection: Self-supervised Learning of Turn-taking Events," May 2022.