移行適格場の情報を考慮したターンテイキング予測

Turn-taking Prediction Considering Transition Relevance Place

原 康平 *　井上 昂治　高梨 克也　河原 達也
Kohei Hara　Koji Inoue　Katsuya Takanashi　Tatsuya Kawahara

京都大学 大学院情報学研究科
Graduate School of Informatics, Kyoto University

Abstract: We address turn-taking prediction where dialogue systems decide whether to take the conversational floor. In this study, we take into account the transition relevance place (TRP) for turn-taking prediction. TRPs are defined as places where the floor can be exchanged between dialogue participants. We presume that participants predict the places of TRP and then decide to take the floor in some of the TRP places. Previous studies did not consider TRP, instead directly predicted turn-taking labels, which were actually difficult to predict. The proposed model is based on a two-step prediction by distinguishing the prediction of TRP and the prediction of turn-taking behaviors under the places of TRP. We also manually annotated labels that approximate TRP and used the labels to train the proposed model. The experimental results demonstrate that the proposed model outperforms the conventional model that directly predicts turn-taking labels in several dialogue tasks.

1　はじめに

音声対話システムが自然で円滑な対話を実現するためには，ユーザの発話中にポーズが検出されたときに，発話権を取得するべきかを判断するターンテイキング予測が重要である．スマートフォンの音声アシスタントやスマートスピーカでは，ユーザの発話はコマンドなどの単純なものからなることが多く，固定長のポーズを検出することで確実に発話権を取得する．一方，人間どうしの自然な対話では，各ターンはポーズによって区切られた複数の発話で構成される．そのため，ターンテイキングでは，話し手のフィラーや聞き手の相槌などによる高度な調整が行われる．音声対話システムがより多くの実用的な場面で利用されるためには，このような自然な対話に対応することが望ましい．しかし，現在の多くのシステムのように，ターン終了の検出を固定長のポーズによって行う場合，不自然なポーズがあいてしまったり，両者の発話が重なる発話衝突が発生したりする．したがって，これらを回避しつつすばやいターンテイキングを実現するという目的で，ターンテイキング予測に関する多くの研究がなされてきた [1, 2, 3]．これらの研究では，先行発話の情報からターンの終了あるいは継続を予測する問題として定式化しており，本研究でもこれを踏襲する．

本研究では，ターンテイキング予測において，移行適格場（TRP: Transition Relevance Place） の情報を利用することを提案する．TRP とは，ターンを構成する基本単位であるターン構成単位が完結可能な時点を指す [4]．従来のターンテイキング予測では，この情報は考慮されず，コーパスにおいて実際に起こった事象をもとに直接予測が行なわれていた．しかし，本来聞き手は発話末において，まずターンが替わりうるかを判断し，その後そうであればターンを獲得するかを判断する．本研究では，この関係性を考慮した上で，TRP の概念に基づいた 2 段階のターンテイキング予測を行うことを提案する．ただし，TRP を客観的に定義することは難しいため，ここでは TRP の近似ラベルを導入し，アノテーションの安定性を確保する．

*連絡先：京都大学 大学院情報学研究科 知能情報学専攻
京都市左京区吉田本町
E-mail: hara@sap.ist.i.kyoto-u.ac.jp

2　従来研究と移行適格場の導入

まず従来のターンテイキング予測について述べ，その後，移行適格場とその導入の必要性について述べる．最後に，それらを踏まえたうえで，本研究の提案手法である移行適格場の情報に基づいたターンテイキング予測について述べる．

2.1　従来のターンテイキング予測

ターンテイキング予測では，いつ話者が交替するかを予測する．典型的な問題設定は，発話末において，話者が交替するか継続するかの二値予測である．特徴量とし

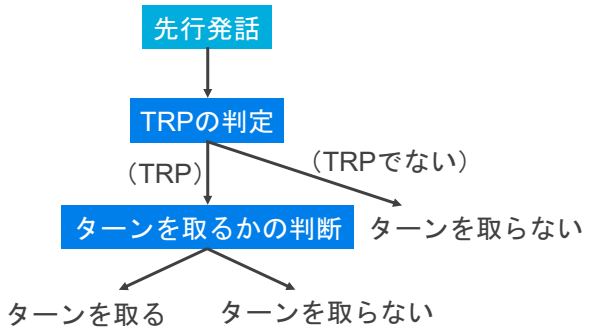

図 1: TRP とターンテイキングとの関係

ては，先行発話の末尾周辺の基本周波数（F0）やパワーなどの韻律的特徴が主に用いられている [5, 6, 7]．また，文節間の依存関係といった統語的特徴も有用とされている [8]．さらに，視線 [9, 10] や呼吸 [11, 12] などの非言語情報も検討されている．予測モデルには，サポートベクトルマシン（SVM）やニューラルネットワークが用いられている [13, 14]．最近では，Long short-term memory（LSTM）などの再帰型ニューラルネットワークを用いて，フレーム単位で特徴量を入力する手法が研究されている [15, 16, 17]．また，ニューラルネットワークを階層的に拡張することで，対話の履歴といった大局的な情報も考慮できるようになっている [2, 3]．

本来，聞き手は，まずいつターンが替わりうるのか（いつターンが終了しうる時点が到来するのか）を見定め，その後，その時点において，ターンを獲得するか否かの判断をしていると考えられる．従来のターンテイキング予測では，このような複雑な関係をコーパス上で実際にみられたターンテイキングの事象のみからモデル化しようとしていたと考えられる．

2.2 移行適格場（TRP）の導入

上記の関係性を考慮するために，ターンが終了しうる時点として移行適格場（TRP）の概念を導入する．本研究における TRP の定義は，Sacks の研究 [4] を参考とする．ここでは，TRP はターンを構成する基本単位であるターン構成単位が完結可能な時点を指す．また，TRP の時点では，ターン割り当て規則が適用されることによって，円滑なターンテイキングがなされる，とされている．実際のターンテイキングのふるまいは，このターン割り当て規則が適用された結果であると解釈することができる．このことから，TRP とターンテイキングの間には図 1 に示す関係があると考えられ，この関係性を考慮するために，その時点が TRP であるか否かの情報を用いることは，ターンテイキング予測に有用であるといえる．

ただし，本来その時点が TRP でありターン割り当て規則が適用されるのかについては，対話参与者の内部で判断されるため，第三者が後から判断することは原理的にはできない．そこで，TRP のアノテーションでは，TRP 自体を直接扱うのではなく，何らかの基準に基づいた近似や評定が行われている．小磯 [7] の研究では，長い発話単位（LUU: Long utterance unit） [18] の末尾を，TRP として近似している．長い発話単位は，統語的・談話的・相互行為的な境界を含んでいるが，LUU でも TRP でないところが多数みられる．一方，Kane ら [19] は，間休止単位（IPU: Inter-pausal unit）の末尾において，話者交替が生起するか否かを第三者に予測してもらった．ここでの目的は，ターンテイキング予測の対象箇所を，第三者でも予測可能である箇所のみに絞ることであったが，この人手による予測は，TRP の判断に近いものである．本研究ではこれに近い TRP のアノテーションを行う．

2.3 TRP を考慮したターンテイキング予測

本研究では，図 1 の関係を考慮した上で，以下に示す 2 段階の予測によって，ターンテイキング予測を行う．1 段階目では，先行発話末が TRP であるか否かを判定する．2 段階目では，TRP の箇所において話者交替または継続（ターンテイキング）を判断する．これらの結果を統合することで，最終的なターンテイキングの予測を行う．具体的なモデルや予測方法については 5 章で述べる．

本研究では，TRP を近似したラベルを第三者にアノテーションしてもらい，それを TRP の代わりに使用する．アノテーション方法の詳細は 4 章で述べる．

3 対話コーパス

被験者と遠隔操作された自律型アンドロイド ERICA [20, 21] による対話コーパスを使用した．ERICA は別室の操作者によって遠隔操作されており，操作者がマイクに向かって話した音声を ERICA のスピーカから再生している．また，ERICA のうなずきや視線などの非言語動作は，操作者の手元にあるコントローラで操作されている．

本研究では，これまでに収録を行った以下の 3 種類の対話タスクを使用した．

- 面接
  ERICA が面接官役，被験者が志願者役として，就職試験の模擬面接を行った．この面接では，志望動機やスキルに関する質問が志願者に投げかけられ，その回答に応じた掘り下げ質問が適宜なされた．各面接は約 10 分程度であり，ここでは 13 対話分のデータを使用する．このような面接対話で

は，面接官（ERICA）が対話の主導権を持つが，発話の大半は志願者（被験者）によるものである．

- 傾聴
  被験者が特定のテーマについて語る状況で，ERICA は聞き手役としてその語りを促進させるための聞き手応答を発する傾聴を行った．語りのテーマは，「印象に残っている旅行」や「最近食べておいしかったもの」などである．各対話は約 10 分程度であり，ここでは 13 対話分のデータを使用する．傾聴対話では，語り手（被験者）が対話の主導権を持ち，発話の大半を占める．したがって，語り手がターンを保持するのに比べて，聞き手（ERICA）がターンを獲得することは少なく，発話数に偏りが生じ，ターンテイキングの予測が難しい．
- お見合い
  ERICA が女性参加者役，被験者が男性参加者役となり，お見合いの練習を行った．この対話の目的は，初対面の相手と適切に対話を進め，親睦を深められるようになることである．対話の内容は，趣味や嗜好について，お互いに質問したり，それに対して反応したりする．各対話は約 10 分程度であり，ここでは 33 対話分のデータを使用する．お見合い対話は，被験者と ERICA の間で対話の主導権が頻繁に入れ替わる混合主導であり，発話数の偏りは少ない．

なお，被験者は対話ごとに異なるが，オペレータは全体で 5 人で，各人物が複数の対話に参加している．

4 アノテーション

3 章で述べたコーパスに対して，TRP のアノテーションを行い，この結果をターンテイキング予測モデルの学習に利用する．実際には，TRP の近似ラベルをアノテーションする．以下では，コーパスに対して既になされていたアノテーションおよび本研究で行った TRP のアノテーションについて述べ，最後に TRP のアノテーション結果の分析について述べる．

4.1 既存のアノテーション

3 章で述べたコーパスには，以下に挙げる情報がアノテーションされている．

- 間休止単位（**IPU: Inter-Pausal Unit**）
  書き起こしを 200 ミリ秒のポーズで区切った発話単位．通常，音声対話システムでは，ポーズごとに音声認識が行われる．
- 長い発話単位（**LUU: Long Utterance Unit**）
  統語的・談話的・相互行為的な境界によって区切られる発話単位．通常，話者交替はこのような境界において起こりうると考えられる．
- 対話行為
  発話に対し，一般目的機能の分類に基づき質問や応答といった対話行為のラベルを付与している．対話行為の認定対象の発話単位として，LUU を採用している．また，複数の発話間の対話行為の関係を表すために，「質問→応答」といった隣接ペアのアノテーションも行なわれている．これらは，質問の後は必ず話者交替が起こるといった情報や同一の対話行為間では話者交替が起こりにくいといった情報を考慮するためにターンテイキング予測に利用することができる．

4.2 TRP のアノテーション

2.2 節で述べたように，その時点が TRP であり，ターン割り当て規則が適用されたのかについて，第三者が客観的に判断することは難しい．したがって，本研究では，TRP 自体ではなく，TRP を近似したラベル（以下，aTRP: approximate TRP）を導入し，その定義は以下とした．LUU の末尾において，聞き手の視点に立ったときに，ターンを取得しようとすれば可能であると判断できる箇所は aTRP である．逆に，ターンを取得してはならないと判断できる箇所は非 aTRP である．ただし，その後に続く発話内容や実際のターンテイキングの事象は考慮しない．これらの判断材料の一部として，隣接ペアおよび対話行為ラベルの情報を用いた．例えば，質問に対する回答において，回答が途中である場合には，その LUU 末は aTRP ではないとする．また，aTRP であるか否かが曖昧な時点については，曖昧である旨を表すラベルを別途用意した．この曖昧な時点のラベルが付与された数は比較的少ないが，分析を行ったところ，隣接ペアの位置やポーズの有無などから 7 つに分類することができた．ここでは，そのうち aTRP に近いと思われる 4 種類を aTRP に，残りを非 aTRP とした．

4.3 アノテーション結果の分析

アノテータ間でのラベル一致度を調査した．はじめに，2 名のアノテータに，各対話タスクで 2 対話ずつをアノテーションしてもらい，Cohen の κ 係数を算出した．その結果，面接では $\kappa = 0.792$，傾聴では $\kappa = 0.784$，お見合いでは $\kappa = 0.817$ となり，それぞれ十分な一致率が観測された．したがって，aTRP のアノテーションは，アノテータ間での安定性が高いことがわかったた

表 1: 移行適格場の近似ラベル（aTRP）とターンテイキングのラベルの内訳（数値は IPU の数．括弧内は aTRP の時点におけるターンテイキングのラベルの内訳.）

タスク	aTRP		ターンテイキング	
	aTRP	非 aTRP	交替	継続
面接	363	1,515	293（ 289）	1,585（ 74）
傾聴	891	2,420	580（ 474）	2,731（ 417）
お見合い	3,089	3,798	2,259（1,988）	4,628（1,101）

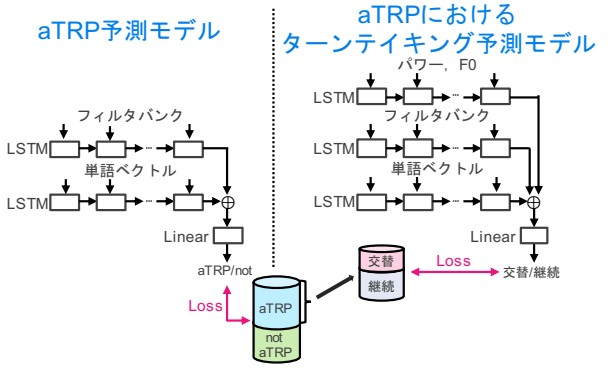

図 2: aTRP 予測モデルおよび TRP におけるターンテイキング予測モデルの学習

め，残りの全対話は，この 2 名のうち 1 名のみによってアノテーションを行った．以降では，このアノテータがアノテーションしたデータを用いる．

aTRP のラベルおよびターンテイキングのラベルの内訳を表 1 に示す．ただし，単位は，後述のターンテイキング予測と同じ IPU である．いずれの対話タスクにおいても以下の傾向が共通してみられた．ターンテイキングにおける交替の数よりも，aTRP の数が多くなった．したがって，継続の数よりも，非 aTRP の数が少なくなった．これは，実際には交替でなかったが，ターンが交替してもおかしくなかった時点が，aTRP と認定されているためである．ターンテイキングの内訳をみてみると，全体では交替よりも継続の数が多かったが，aTRP においては交替のほうが多くなっている．したがって，その時点が aTRP であるという情報は，話者交替であることを予測するのに有用であるといえる．なお，LUU の数は，面接が 754，傾聴が 1,557，お見合いが 4,380 であった．このことからも，すべての LUU が aTRP にはなっていないことがわかる．また，aTRP ではない箇所で交替が起こっている箇所も少数ではあるが存在した．aTRP の定義より，通常このような現象は起こりにくいため，その原因を調査した．これらを分析したところ，聞き手の割り込みや「そうですか」などの聞き手応答によるターン獲得が見受けられた．これらの箇所は，その参与者にとっては TRP であると判断されたといえるが，現状の TRP の近似ラベルではカバーされない．

5 提案モデル

本研究では，TRP を予測するモデルと TRP においてターンテイキングを予測する 2 つのモデルを用いてターンテイキングを予測する．ただし，両予測ともに IPU 末で行う．

5.1 aTRP 予測

各 IPU の末尾において aTRP であるか否かを予測する．図 2 の左側に学習の概要を示す．モデルは先行研究 [17] で用いられた LSTM に基づくものである．入力は，40 次元の対数メルフィルタバンクと 100 次元の単語ベクトルである．ただし，単語ベクトルは Word2Vec モデルを用いて抽出する．教師データは，4.2 節で述べた人手でアノテーションされた aTRP のラベルである．

5.2 aTRP におけるターンテイキング予測

IPU 末が aTRP である箇所についてターンテイキングを予測する．図 2 の右側に学習の概要を示す．モデルは TRP 予測の場合と同様の LSTM に基づくものである．特徴量は，対数メルフィルタバンクと単語ベクトルに加えて，韻律的特徴としてパワーと基本周波数（それぞれ一次と二次の変量を含む）の 6 次元を用いる [1]．教師データは，人手でアノテーションされたターンテイキングのラベルである．

5.3 aTRP 予測に基づくターンテイキング予測

5.1 節と 5.2 節のモデルを用いて，ターンテイキングを予測する．IPU 末における予測手順を以下に示す．

1. 5.1 節のモデルを用いて，aTRP である確率 $P(aTRP)$ を算出する．
2. 5.2 節のモデルを用いて，その時点が aTRP であるという仮定のもとでの話者交替である確率 $P(Take|aTRP)$ を算出する．

[1] aTRP 予測においても韻律的特徴の使用を試みたが，予測精度が向上しなかったため，今回は aTRP 予測では用いていない

表 2: ターンテイキングの予測結果

タスク	モデル	正解率	適合率	再現率	F 値	F 値マクロ
面接	ベースライン	90.4	69.4	68.3	68.8	81.6
	提案	92.9	72.8	86.7	79.1	87.4
傾聴	ベースライン	81.8	28.8	2.6	4.7	47.3
	提案	81.9	48.1	44.1	46.0	67.6
お見合い	ベースライン	76.0	65.9	55.5	60.3	71.5
	提案	76.4	77.5	39.4	52.2	68.2

3. $P(Take, aTRP) = P(aTRP) \times P(Take|aTRP) > 0.5$ ならば話者交替，そうでなければ話者継続を予測結果とする．

3. の条件式は，図 1 や 4.3 節で示したように，aTRP でない箇所では話者継続であるという前提に基づいている．

6 評価実験

提案モデルの有効性を評価するために，aTRP の情報を用いないベースラインモデルとの比較を行った．

6.1 実験条件

3 章で述べたアンドロイド ERICA と被験者との対話データを使用した．aTRP のラベルは対話タスクに依存しないと考えられるため，提案モデルの前段である aTRP の予測モデルは全タスクのデータを用いて学習を行った．一方，後段の aTRP におけるターンテイキングの予測は，対話タスクに依存すると考えられるため，モデルの学習は対話タスク毎に行った．これらの学習および評価は，5 分割交差検定により行った．ベースラインモデルは，aTRP の情報を考慮せずに，ターンテイキングのラベルのみを用いて学習したものである．具体的には，提案モデルにおける後段の LSTM と同じで，学習には，aTRP でない箇所も含むターンテイキングのラベルを用いる．また，本研究では，ユーザ発話の IPU 末におけるターンテイキングのみに着目した．したがって，アンドロイド ERICA の発話の IPU 末は対象外である．評価指標として，正解率，適合率・再現率・F 値，正例と負例の F 値の平均であるマクロ平均を用いた．

モデルの実装方法について述べる．活性化関数は，出力層では Softmax，その他の層では ReLU を用いた．ミニバッチサイズは 32 とし，パラメータの更新には学習係数 10^{-4} の RMSProp を用いた．損失関数には交差エントロピーを用いた．各層間のドロップアウト率は 0.2 に設定した．ベースラインモデルと aTRP 予測モデルには 3 層の LSTM，aTRP におけるターンテイキング予測モデルには 1 層の LSTM を使用した．各層のノード数は 128 とした．実装には，Pytorch 0.4.1 を用いた．パワーおよび F0，対数メルフィルタバンクは，IPU 末から過去 2 秒間をフレームシフトサイズ 10 ミリ秒で抽出し，IPU 毎に平均 0，分散 1 に正規化した．単語ベクトルは，MeCab [2] で分割された単語系列を用いて，gensim [3] で Continuous Bag-of-Words（CBOW）モデルを学習した．そのため，単語ベクトルの抽出は単語の出現に合わせて行われる．ただし，全学習データを用いた．

6.2 実験結果

ターンテイキングの予測結果を表 2 に示す．面接と傾聴では，F 値およびそのマクロ平均から，提案モデルによる精度向上が確認された．特に，再現率が大きく改善されている．今回の対話データは，1 つのターンに複数の IPU が含まれる自然な対話であるため，ターンテイキングのラベルに関しては，継続の割合が大きい．また，この継続の要因には，非 TRP であることやタスクに依存するものが混在していると考えられる．提案モデルではこれらを分割して学習を行っているため，それぞれの学習が改善され，結果として全体の精度が向上したと考えられる．ただし，お見合いでは精度の改善がみられず，逆に，再現率が下がっている．この要因として，提案モデルの一段階目である aTRP の予測において，再現率が低く，結果として全体的な再現率も低くなっていると考えられる．

提案モデルは二段階のモデルで構成されるため，それぞれの精度についても調べた．はじめに，aTRP 自体の予測精度を調べた．結果を表 3 に示す．正解率のチャンスレベルは，面接が 80.7，傾聴が 73.1，お見合いが 55.1 であるため，高い精度で予測できているといえる．ただし，前述の通り，お見合いの再現率は低い．続いて，テストデータを aTRP の時点のみに絞り，二段階目のモデルのみを用いてターンテイキングの予測精度を調べた．結果を表 4 に示す．正解率のチャンスレベルは，面接が 79.6，傾聴が 53.2，お見合いが 64.4 であり，こちらは若干の改善にとどまっている．この予測に関しては，対話タスクに依存する部分でもあるため，

[2] http://taku910.github.io/mecab/
[3] https://radimrehurek.com/gensim/

表 3: aTRP の予測結果（提案モデルの一段階目の評価）

タスク	正解率	適合率	再現率	F 値	F マクロ
面接	93.3	77.9	89.3	83.2	89.4
傾聴	82.4	72.9	55.6	63.1	75.8
お見合い	81.4	84.6	71.5	77.5	80.8

表 4: aTRP におけるターンテイキングの予測結果（提案モデルの二段階目の評価）

タスク	正解率	適合率	再現率	F 値	F マクロ
面接	81.3	83.0	96.2	89.1	61.2
傾聴	54.2	55.4	71.1	62.3	52.0
お見合い	68.3	70.1	88.3	78.2	60.1

これを反映させた特徴量の追加が必要だと考えられる．

7 おわりに

本稿では，移行適格場（TRP）の情報を用いたターンテイキングの予測手法を提案した．聞き手によるターンテイキングの判断は，その時点が TRP である，そしてその TRP の時点でターンを獲得する，という二段階で構成されると考えた．提案モデルは，はじめに TRP を予測し，そのうえで TRP におけるターンテイキング予測も行う．また，TRP を第三者がアノテーションすることは困難であるため，これを近似するラベルを導入し，アノテーションを行った．実験の結果，面接と傾聴という対話タスクにおいて，提案モデルは，ターンテイキングを直接学習するベースラインに比べて，予測精度が向上することを確認した．今後の課題として，特徴量の追加およびより安定的と考えられる長い発話単位（LUU）のラベルの利用を検討している．

謝辞

本研究は，JST ERATO 石黒共生ヒューマンロボットインタラクションプロジェクト JPMJER1401 の支援を受けて実施した．

参考文献

[1] 駒谷和範, “円滑な対話進行のための音声からの情報抽出,” **電子情報通信学会誌**, vol. 101, no. 9, pp. 908–913, 2018.

[2] M. Roddy *et al.*, “Multimodal continuous turn-taking prediction using multiscale rnns,” in *ICMI*, pp. 78–86, 2018.

[3] R. Masumura *et al.*, “Neural dialogue context online end-of-turn detection,” in *SIGDIAL*, pp. 224–228, 2018.

[4] S. Harvey *et al.*, “A simplest systematics for the organization of turn-taking for conversation,” *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[5] M. Zellers, “Perception of pitch tails at potential turn boundaries in Swedish,” in *INTERSPEECH*, pp. 1944–1948, 2014.

[6] O. Niebuhr *et al.*, “Speech reduction, intensity, and F0 shape are cues to turn-taking,” in *SIGDIAL*, pp. 261–269, 2013.

[7] 小磯花絵, “話者交替における統語的・韻律的特徴の役割：日本語三者会話の定量的分析に基づく考察,” **音声研究**, vol. 14, no. 3, pp. 13–26, 2010.

[8] Y. Ishimoto *et al.*, “End-of-utterance prediction by prosodic features and phrase-dependency structure in spontaneous japanese speech,” in *INTERSPEECH*, pp. 1681–1685, 2017.

[9] K. Jokinen *et al.*, “Turn-alignment using eye-gaze and speech in conversational interaction,” in *INTERSPEECH*, pp. 2018–2021, 2010.

[10] R. Ishii *et al.*, “Predicting next speaker and timing from gaze transition patterns in multi-party meetings,” in *ICMI*, pp. 79–86, 2013.

[11] R. Ishii *et al.*, “Analyzing mouth-opening transition pattern for predicting next speaker in multi-party meetings,” in *ICMI*, pp. 209–216, 2016.

[12] M. Włodarczak *et al.*, “Respiratory turn-taking cues,” in *INTERSPEECH*, pp. 1275–1279, 2016.

[13] G. Skantze *et al.*, “Turn-taking, feedback and joint attention in situated human–robot interaction,” *Speech Communication*, vol. 65, pp. 50–66, 2014.

[14] N. G. Ward *et al.*, “Dialog prediction for a general model of turn-taking,” in *INTERSPEECH*, pp. 2662–2665, 2010.

[15] M. Roddy *et al.*, “Investigating speech features for continuous turn-taking prediction using lstms,” in *INTERSPEECH*, pp. 586–590, 2018.

[16] R. Masumura *et al.*, “Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks,” in *INTERSPEECH*, pp. 1661–1665, 2017.

[17] D. Lala *et al.*, “Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios,” in *ICMI*, pp. 78–86, 2018.

[18] Y. Den *et al.*, “Two-level annotation of utterance-units in japanese dialogs: An empirically emerged scheme.,” in *LREC*, 2010.

[19] J. Kane *et al.*, “Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions,” in *INTERSPEECH*, pp. 1681–1685, 2017.

[20] K. Inoue *et al.*, “Talking with ERICA, an autonomous android,” in *SIGDIAL*, pp. 212–215, 2016.

[21] T. Kawahara, “Spoken dialogue system for a human-like conversational robot ERICA,” in *IWSDS*, 2018.