相槌・フィラー予測とのマルチタスク学習による ターンテイキング予測

Prediction of Turn-taking Using Multi-task Learning with Prediction of Backchannels and Fillers

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 a turn-taking prediction model that considers related behaviors such as backchannels and fillers. Backchannels are used by listeners to acknowledge that the current speaker can hold the turn. Fillers are used by the prospective speaker to take a turn. Therefore, it is expected that predicting not only the turn-taking behavior but also the two related behaviors leads to the improvement of the accuracy of turn-taking prediction itself. The proposed model is a neural network that performs multi-task learning by sharing an LSTM layer among the multiple prediction tasks. We evaluated the accuracy on the turn-taking prediction with two kinds of dialogue corpora of human-robot interaction. The result demonstrated that the proposed multi-task learning model outperformed the conventional single-task learning model.

1　はじめに

音声対話システムは様々な場面で実用化されているが，ユーザの発話がいつ始まり，そしてそのターンがいつ終わるのかを予測するターンテイキングの問題は，自然で円滑な対話を実現するためには重要な課題である．スマートフォンやスマートスピーカなどでは，発話の開始はボタンやマジックワードなどにより明示的になされる．また，発話内容はコマンドなどの単純な 1 発話になることが多く，ターンの終了はポーズ検出により確実に行うことができる．一方，人間どうしの自然な対話では，1 つのターンに複数の発話が含まれたり，聞き手によって相槌が使用されたりするなど，ターンテイキングは自明でない．従来システムのように，固定長のポーズをターンの終了とみなすと，不自然な間があいてしまったり，発話衝突が起こったりする可能性がある．したがって，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に関して，これまでに多くの研究がなされてきた [1].

本研究では，ターンテイキングの予測において，相槌およびフィラーの予測も 1 つのモデルで同時に行うことを提案する．相槌とフィラーは，ターンテイキングのふるまいと関係していることが知られている [2, 3]. 相槌は，現話者がターンを継続することを認める際に，聞き手によって用いられる．フィラーは，次話者がターンを獲得する際，および現話者がターンを保持する際に，その意図を示すために用いられる．図 1 に，対話相手の発話末において，ターンを獲得する確信度とそれに対応する行動の概念を示す．ターンテイキングが自明であれば，相槌やフィラーは用いられないが，曖昧な場合にはそれぞれを用いることで，対話相手と発話権を調整していると考えられる．したがって，相槌およびフィラーを予測することは，ターンテイキングの予測自体に寄与すると考えられる．本稿では，これらの関係性を考慮するために，3 つの予測タスクを同時に行うニューラルネットワークを，マルチタスク学習の枠組みを用いて実現する．ネットワークの前半を予測タスク間で共有させ，どのタスクにも必要となる

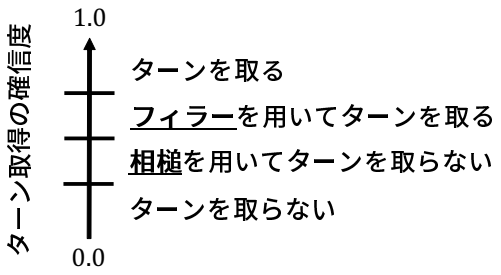

図 1: 相槌・フィラーとの関係を考慮するターンテイキングのふるまい

*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

共通の情報を得ることを狙う．これにより，ターンテイキングの予測単体では学習が困難だった箇所に関して，予測精度の向上が期待される．

2 関連研究

ターンテイキング，相槌，フィラーの特徴および予測に関する関連研究について述べる．

2.1 ターンテイキング予測

ターンテイキング予測の典型的な問題設定は，発話末において，先行発話の特徴量から現話者のターンが終わるか否かを予測することである．先行発話の特徴量として，ピッチやパワーといった韻律特徴が有用であることが示されている [4, 5]．さらに，視線 [6] や呼吸 [7, 8] といった非言語情報も検討されている．予測のモデルには，SVM やニューラルネットワークが用いられている [2, 3]．最近では，LSTM などの再帰型ニューラルネットワークを用いて，フレーム単位の特徴量をそのまま入力する手法も提案されている [9, 10]．

2.2 相槌予測

相槌とは，聞き手によって発話される「うん」や「はい」などの短い発話であり，現話者のターンの保持を促す機能をもつ．過去の研究では，先行発話の韻律特徴 [11, 12] や言語特徴 [13, 14] から相槌の生起を予測している．ここで用いられる韻律特徴は，前述のターンテイキングの予測で用いられていたものと概ね同じである．また，相槌の形態の予測も行われている [15]．

2.3 フィラー予測

フィラーとは,「えー」や「あのー」といった言いよどみ時などに出現する場繋ぎ的な表現である．フィラーは，対話参与者が次発話について考えている状態を示したり，沈黙を和らげたりといった機能を持つ．また，ターンテイキングと関連する場面では，フィラーはターンを保持し続ける意志や，これからターンを獲得する意志を対話相手に示すために用いられる．フィラーの韻律的な特徴を分析する研究はいくつかなされているが [16, 17]，フィラーの予測に関する研究は限られている [18, 19]．

3 コーパス

我々が収録を進めている被験者と自律型アンドロイド ERICA[20, 21] による対話データを本研究で使用した．ERICA は別室のオペレータによって遠隔操作されており，オペレータが発した音声がそのまま ERICA のスピーカから再生されている．したがって，音声による人間どうしの遠隔対話に近い状況といえる．また，ERICA のうなずきや視線などの非言語動作は，オペレータが手元のコントローラを用いて操作している．収録した音声データに基づき，発話の書き起こしを行った．発話単位は，200 ミリ秒のポーズを基準とする間休止単位（IPU）とした．我々はこれまでに，様々な種類の対話コーパスを収録してきたが，本研究では以下の 2 種類を使用した．

3.1 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，その回答に応じて掘り下げるような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 15 対話分のデータを使用する．被験者は対話ごとに異なるが，オペレータは同一の人が複数の対話を兼ねており，オペレータの人数は全体で 4 人であった．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，対話における発話の大半は志願者（被験者）によってなされる．また，面接官は相槌をうつが，逆に志願者が相槌をうつことは少ない．

3.2 傾聴

ERICA が聞き手となり，被験者の話に対して傾聴を行った．語り手である被験者が話を継続しやすくなるように，ERICA は聞き手としての応答を行った．ここでの聞き手応答は，相槌，繰り返し，掘り下げ質問，自己呈示などであり，オペレータに適切なタイミングと種類を判断してもらった．各対話は約 10 分程度であり，ここでは 15 対話分のデータを使用する．面接対話と同様に，被験者は対話毎に異なるが，オペレータは 4 人が交代で務めた．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，聞き手（ERICA）がターンを獲得することは少ないため，学習サンプルの偏りからターンテイキングの予測はかなり難しいといえる．その一方で，聞き手は相槌を頻繁に用いる．

表 1: ターンテイキングの予測タスクにおけるサンプル数の内訳

コーパス	先行話者	ターンテイキング	
		交替	継続
面接	面接官	175	642
	志願者	276	620
傾聴	語り手	306	1,131
	聞き手	162	364

表 2: 相槌の予測タスクにおけるサンプル数の内訳

コーパス	予測対象者	相槌	
		有	無
面接	面接官	223	884
	志願者	52	894
傾聴	語り手	74	626
	聞き手	741	1,175

3.3 対話コーパスによるふるまいの違い

上記のように，対話の種類によって主導権やターンテイキングのふるまいの傾向が異なると考えられる．したがって，対話コーパスの種類毎に予測モデルの学習および評価を行う．さらに，これらのふるまいは対話参与者の役割毎にも異なる．そのため，被験者と ERICA に対しても別々にモデルの学習および評価を行う．

各予測タスクにおけるサンプルの内訳を表 1 から表 3 に示す．ただし，予測タスクによって合計のサンプル数が異なることに注意されたい．

ターンテイキングの予測（表 1）では，IPU 末に続く 1 秒以内にいずれかの参与者が発話した場合が予測点となり，その後続話者がどちらの参与者であるかによって,「交替」または「継続」のラベルが付与される．両対話タスクともに「継続」のサンプル数が比較的多いことがわかる．つまり，1 つのターンが複数の IPU で構成されていることから，これらの対話では，単純ではない自然な発話がなされていたことが示唆される．さらに，傾聴においては，先行話者が語り手のときの「継続」が圧倒的に多いことがわかる．逆に，聞き手はほとんどターンを獲得しないため，先行話者が聞き手の場合のサンプル数は少ない．

相槌の予測（表 2）では，対話相手の IPU 末に続いて予測対象者が相槌をうったか否かがラベルとして付与される．面接においては，志願者はほとんど相槌を使用していないことがわかる．また，傾聴では聞き手による相槌が多く，逆に，語り手は相槌をほとんど使用していない．これは，ターンテイキングの際に，語り手の「継続」が多いことに関係すると考えられる．

表 3: フィラーの予測タスクにおけるサンプル数の内訳

コーパス	予測対象者	フィラー	
		有	無
面接	面接官	473	1,580
	志願者	287	1,766
傾聴	語り手	403	2,213
	聞き手	286	2,330

フィラーの予測（表 3）では，予測対象者と対話相手の IPU 末に続いてフィラーをいったか否かがラベルとして付与される．面接では，面接官のフィラーが志願者に比べて多い．これは，面接官の発話はほとんどが質問であるため，ターンの受け渡しが明示的であり，そのため，志願者はフィラーを用いなくともターンを確実に取得できるためと考えられる．また，傾聴においては，聞き手のフィラーが比較的少ないことがわかる．聞き手がターンを取得すること自体が少ないためと考えられる．

4 マルチタスク学習に基づく統合モデル

提案モデルは，LSTM を用いたモデル [9] をベースラインとしている．このモデルを拡張することで，ターンテイキングと相槌・フィラーを同時に予測する統合モデルを実現する．このモデルは対話参与者の役割ごとに用意する．例えば，面接であれば，面接官のモデルと志願者のモデルをそれぞれ別々に用意する．

4.1 ベースラインモデル

このモデルが相槌やフィラーを考慮しない,「シングルタスク」のモデルに相当する．はじめに，特徴量について述べる．特徴量は，対話参与者ごとに抽出される．その後，予測対象である参与者とその対話相手それぞれの特徴量を 16 次元のベクトルにまとめる．このモデルへの入力は，時間フレーム毎に行われる．ただし，フレームシフトサイズは 50 ミリ秒である．

- ピッチ（3 次元）
  ピッチとその一次および二次変化量．ピッチの抽出には，Praat[1] を用いた．また，対話参与者ごとに z 正規化を施した．

[1] www.praat.org/

- パワー（３次元）
  パワーとその一次および二次変化量．ピッチと同様のツールおよび正規化を用いた．

- スペクトル安定性（１次元）
  以下の式で算出した．

$$S_t = \frac{\sum_{f=1}^{N} \mathrm{abs}(s_{t,f} - s_{t-1,f})}{\sum_{f=1}^{N} s_{t,f}} \quad (1)$$

  ただし，N は周波数ビン数，$s_{t,f}$ は時間 t における f 番目の周波数ビンのパワースペクトルである．パワースペクトルの抽出には，librosa[2]を用いた．また，対話参与者ごとに z 正規化を施した．

- 発話（１次元）
  対象参与者が発話しているか否かの二値．コーパスの書き起こし情報を用いた．

ベースラインモデルは，対象参与者の将来 1 秒以内の発話確率を出力する．出力は 20 次元（=1000 ミリ秒 / 50 ミリ秒）であり，各次元が将来の各フレームにおける発話確率に対応する．ベースラインモデルは，1 層の LSTM（18 ノード），その後に 3 層の全結合層（各層 20 ノード）からなる．

本研究では，上記のモデルを用いて IPU 末におけるターンテイキングを予測する．各 IPU 末において，各参与者のモデルを用いて，将来 1 秒間の各参与者の発話確率を出力する．出力である 20 次元のベクトルの要素和を各対話参与者について計算し，その値が大きい参与者が次話者であると判定する．つまり，先行話者と次話者が異なる場合は「交替」，先行話者と次話者が同じ場合は「継続」がターンテイキング予測の出力となる．

4.2 統合モデル

4.1 節のターンテイキングの予測モデルを，相槌とフィラーも予測するマルチタスクのモデルへと拡張する．図 2 に提案モデルを概要を示す．提案モデルは，最初の LSTM および全結合層において各タスクで共通の特徴を学習する．また，次の各タスク毎に分岐した層において各タスク特有の特徴を学習する．ただし，ターンテイキングの予測は対象参与者の将来 1 秒間の発話確率であったが，相槌およびフィラーの予測では対象参与者とその対話相手のそれぞれが将来 1 秒間以内に相槌とフィラーをそれぞれ発話するか否かである．したがって，計 4 次元（２参与者 × ２タスク）の出力となる．以上により，異なるタスク間の関係性を考慮す

[2] librosa.github.io/librosa/

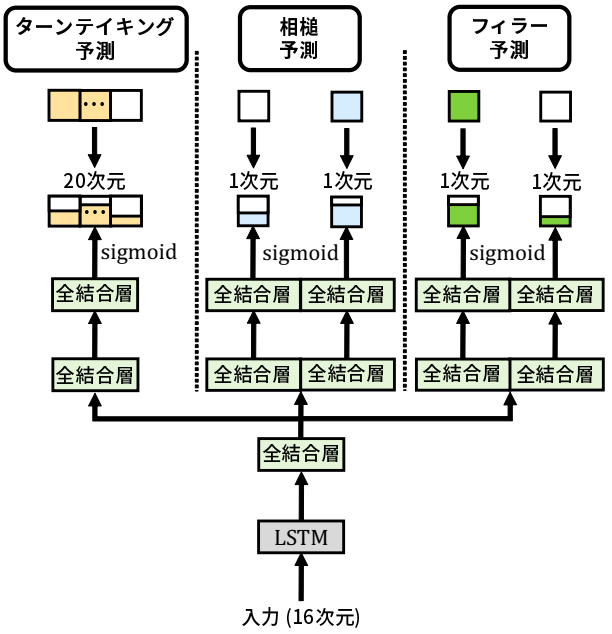

図 2: マルチタスク学習を用いた統合モデル（相槌とフィラーは対話参与者の両者についてそれぞれ予測する）

ることで，ターンテイキングの情報のみでは困難な予測も可能になることを狙う．

学習時の目的関数は以下とする．

$$\mathcal{L} = \alpha\, \mathcal{L}_{turn} + \beta\, \mathcal{L}_{bc} + \gamma\, \mathcal{L}_{filler} \quad (2)$$

ここで，$\mathcal{L}_{turn}$ はターンテイキング予測のための損失関数，$\mathcal{L}_{bc}$ と $\mathcal{L}_{filler}$ はそれぞれ相槌とフィラー予測のための損失関数である．また，α，β，γ は各損失関数の重み係数である．さらに，各予測タスクの損失関数は以下のように算出する．

$$\mathcal{L}_{task} = \begin{cases} r_{task,N} \times \mathrm{MSE}_{task} & (\text{正例の場合}) \\ r_{task,P} \times \mathrm{MSE}_{task} & (\text{負例の場合}) \end{cases} \quad (3)$$

ここで，$r_{task,N}$ はタスク $task$ における負例の割合，$r_{task,P}$ は正例の割合である．これにより，正例と負例とのサンプル数の偏りを補正する．また，MSE_{task} はタスク $task$ における平均二乗誤差である．

5 評価実験

ターンテイキングおよび相槌・フィラーの予測について，提案モデルを評価した．

5.1 条件

各対話コーパスに対して，５分割の交差検定を行った．比較手法は，ターンテイキング予測のみを考慮し

表 4: 面接における被験者（志願者）の発話末におけるターンテイキング予測の結果

モデル	正解率	交替			継続		
		適合率	再現率	F 値	適合率	再現率	F 値
ベースライン	84.9	79.3	69.2	73.9	87.0	91.9	89.4
マルチタスク	87.1	84.2	71.4	77.3	88.1	94.0	91.0

表 5: 傾聴における被験者（語り手）の発話末におけるターンテイキング予測の結果

モデル	正解率	交替			継続		
		適合率	再現率	F 値	適合率	再現率	F 値
ベースライン	80.4	54.6	48.7	51.5	86.5	89.0	87.8
マルチタスク	82.2	60.4	47.4	53.1	86.5	91.6	89.0

たシングルタスクモデル（ベースラインモデル）である．各モデルは，各対話コーパスにおいて対話参与者の役割ごとに学習した．評価指標として，適合率と再現率，および F 値を使用した．また，ターンテイキング予測に関しては，正解率も算出した．

ニューラルネットワークのパラメータの設定は以下の通りである．LSTM 層のノード数は 18，全結合層のノード数は 20 とした．活性化関数は，出力層にはシグモイド関数，隠れ層には ReLU を用いた．学習時のネットワークの重みの更新はミニバッチごとに，学習係数が 1.6×10^{-3} の RMSProp で行なった．各バッチは 20 秒の時系列データを 30 系列含むようにした．学習係数は，10 エポックごとにテストデータの精度が改善しなかった場合に半減させ，4 回半減したところで十分小さい値であるとして固定した．過学習を避けるために各層間にドロップアウトを適用し，その率は 0.2 とした．ニューラルネットワークの実装には，TensorFlow 1.4.1[3]をバックエンドとする Keras 2.1.2 [4] を用いた．

5.2 ターンテイキング予測

各コーパスにおいて先行話者が被験者である場合のターンテイキングの予測について評価した．提案モデルの損失関数（式 2）の重みは，$\alpha = 0.6$，$\beta = 0.2$，$\gamma = 0.2$ とした．結果を表 4 および表 5 に結果を示す．マルチタスクモデルでは，両対話コーパスにおける「交替」の適合率が向上した．これは，ターンを獲得しようとしている次話者が，相槌ではなくフィラーを使用しているという情報を考慮できるようになったためと考えられる．また,「継続」の再現率も両コーパスにおいて向上した．これは，現在の聞き手の相槌および現話者によるフィラーを考慮したためと考えられる．

[3] www.tensorflow.org
[4] github.com/keras-team/keras

5.3 相槌およびフィラー予測

相槌またはフィラーの発話を行いながら，ターテイキングを制御するシステムを想定して，ERICA による相槌とフィラーの予測精度を評価した．ここで，提案モデルのネットワーク構成を変更した．まず，相槌を予測する場合には，ネットワークの出力は，両参与者（２人分）の発話および予測対象参与者の相槌の３つとした．また，損失関数の重みは，両参与者の発話予測はそれぞれ 0.1，予測対象参与者の相槌は 0.8 とした．フィラーを予測する場合は，上記の相槌をフィラーに置き換える．

結果を表 6 から表 8 に示す．面接における相槌予測（表 6）は，マルチタスクモデルにより精度が向上した．この結果は，発話末における ERICA（面接官）のターンテイキングのふるまいが相槌に大きく関係しているためと考えられる．傾聴での相槌予測（表 7）は，マルチタスクモデルによる改善はみられなかった．傾聴においては，現話者のターンを継続させるという目的以外にも相槌が多く使用されるためと考えられる．例えば，現話者の発話に対して共感を示すために相槌が用いられる．そのため，傾聴においては，ターンテイキングのふるまいを考慮しても相槌の予測精度は向上しなかった．面接におけるフィラー予測（表 8）では，マルチタスクモデルによる改善はわずかなものにとどまった．面接では，志願者が十分に回答を述べてから面接官はターンを獲得する．また，志願者は面接官が質問を言い終えるまで確実に待つため，面接官はフィラーを使わなくともターンを保持することができる．そのため，面接においては，ターンテイキングのふるまいを考慮してもフィラーの予測精度は大きく向上しなかった．また，傾聴においては ERICA（聞き手）のターン保持は少なく，したがってフィラーも少ない．そのため，ここではフィラーの予測精度は評価しない．

表 6: 面接における ERICA（面接官）の相槌の予測結果

モデル	適合率	再現率	F 値
ベースライン	25.8	79.4	38.9
マルチタスク	27.8	87.0	42.1

表 7: 傾聴における ERICA（聞き手）の相槌の予測結果

モデル	適合率	再現率	F 値
ベースライン	45.5	89.2	60.3
マルチタスク	44.2	91.4	59.6

表 8: 面接における ERICA（面接官）のフィラーの予測結果

モデル	適合率	再現率	F 値
ベースライン	31.4	84.4	45.8
マルチタスク	31.1	88.8	46.0

6 おわりに

本稿では，相槌とフィラーのふるまいを考慮したターンテイキングの予測について述べた．ここでは，ターンテイキングの予測と同時に相槌とフィラーの予測も行うマルチタスク学習を提案した．提案モデルはニューラルネットワークで構成され，その一部が各予測タスク間で共有されている．提案モデルを，面接と傾聴という，対話参与者のふるまいの傾向が異なる 2 種類の対話コーパスで評価した．その結果，ターンテイキングのみを考慮したシングルタスクモデルよりも，マルチタスクモデルが高い精度を示すことを確認した．今後は，このモデルをアンドロイド ERICA のシステムへと応用し，相槌やフィラーを発話しながらターンテイキングを制御するシステムの実現を目指す．

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

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

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

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

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

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

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

[9] G. Skantze, “Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks,” in *SIGDIAL*, pp. 220–230, 2017.

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

[11] M. Mueller *et al.*, “Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques,” in *HCI International*, pp. 329–340, 2015.

[12] N. G. Ward, “Using prosodic clues to decide when to produce back-channel utterances,” in *ICSLP*, vol. 3, pp. 1728–1731, 1996.

[13] H. Koiso *et al.*, “An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs,” *Language and speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[14] N. Kitaoka *et al.*, “Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems,” *Transactions of the Japanese Society for Artificial Intelligence*, vol. 20, no. 3, pp. 220–228, 2005.

[15] T. Kawahara *et al.*, “Prediction and generation of backchannel form for attentive listening systems,” in *INTERSPEECH*, pp. 2890–2894, 2016.

[16] M. Watanabe, *Features and roles of filled pauses in speech communication: A corpus-based study of spontaneous speech*. Hituzi Syobo, 2009.

[17] S. Nakamura *et al.*, “Analysis of the relationship between prosodic features of fillers and its forms or occurrence positions,” in *INTERSPEECH*, pp. 1726–1730, 2017.

[18] S. Andersson *et al.*, “Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection,” in *Speech Prosody*, 2010.

[19] R. Nakanishi *et al.*, “Generating fillers based on dialog act pairs for smooth turn-taking by humanoid robot,” in *IWSDS*, 2018.

[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.