親密性を考慮した対話システムのための マルチモーダル雑談コーパスの構築と分析

Construction and analysis of multimodal chat-talk corpus for dialog systems considering the closeness between speakers

山﨑善啓 * 千葉祐弥 能勢隆 伊藤彰則
Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito

東北大学 大学院工学研究科
Graduate School of Engineering, Tohoku University

Abstract: In recent years, multimodal dialog systems that can make chat-talks with facial expression, gesture, and gaze have been expected as a next-generation dialog-based system. On the other hand, some studies suggested that changing a form of a system utterance depending on the closeness to a user improves user impression. However, a sufficient amount of a multimodal dialog corpus based on the closeness between speakers that contains clear audio and visual information has not constructed so far. In this paper, we constructed a multimodal Japanese chat-talk corpus and analyze the dialog behaviors toward the modeling of the dialog strategy considering the closeness to the user. The constructed dialog corpus contains 19,303 utterances (10 hours) from 19 pairs of participants. We compared the dialog behavior between the different closeness of the speakers. The analysis shows that closeness affects dialog acts and facial expressions, and we obtain clues to model target dialog strategy.

1 はじめに

近年，音声対話システムが幅広く用いられるようになり，特に雑談を行うことでユーザを楽しませる雑談対話システムの研究が盛んに行われている．これらの研究では深層学習に基づく応答生成 [1] を始めとして大規模な対話データを必要とする場合が多い．また言語情報・音響情報に加えて表情・視線・ジェスチャなどを含めたマルチモーダル情報を用いることでユーザとのより円滑なコミュニケーションが可能になると考えられ，マルチモーダル対話システムに関しても様々な研究がなされている [2, 3]．マルチモーダル対話システムにおいて統計的な対話モデルを構築するためには，大量のマルチモーダル対話コーパスが必要になる．言語的な対話データに関してはこれまで多数のコーパスが構築されてきたが，マルチモーダル情報を含んだ対話データの集積は十分ではない．

一方で，雑談対話システムの役割として「ユーザに対話を楽しんでもらうこと」が重要であり，特に従来の研究 [4, 5] によってユーザとの親密性に応じて応答を変化させる対話システムがユーザ評価の向上に有用であることが示されている．ここで，従来の研究では口調などの対話の表層的な要素に着目しているが，対話者の対話行為，韻律，表情といった発話内容以外の要素も話者間の親密性に依存して変化すると考えられる．また，従来の研究では親密性の変化による対話制御を経験的に決めていたが，対話システムへの応用を考えると対話のモデルは集積されたデータから統計的に獲得できるのが望ましい．

これに対して，本研究では話者間の親密性を考慮した高品質かつ大規模なマルチモーダル雑談コーパスの構築を目指す．本稿では，先行して収録した 19 名の対話に関して対話行為・発話の韻律・表情の出現傾向について親密性による差を分析する．

*連絡先：東北大学 大学院工学研究科
〒 980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-05
E-mail: yoshihiro.yamazaki.t2@dc.tohoku.ac.jp

2 マルチモーダル雑談コーパスの構築

2.1 雑談対話の収録

対話システムへの応用を想定して，音声と動画像を含んだ人間同士の一対一の雑談対話を収録した．図 1 に

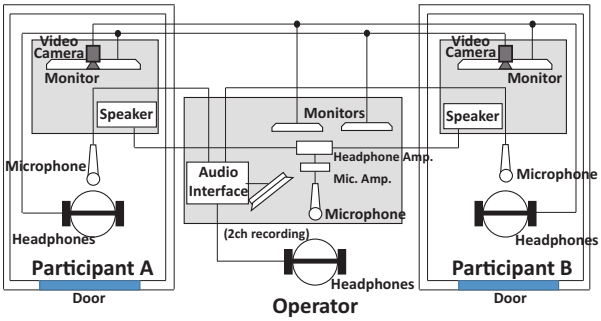


図 1: 収録環境

表 1: 収録時に設定した話題

番号	話題
1	好き，嫌いな食べ物・飲み物
2	好き，嫌いな音楽
3	好きな読み物
4	好きな映画，アニメ
5	今まで行った場所の中で良かった，悪かった所
6	ファッションの好み
7	休日の過ごし方
8	最近，仕事 (学校) で楽しんでいること，満足していること
9	仕事の同僚 (学校の友達) をどう思っているか
10	仕事 (授業) で得意，苦手な所

収録環境を示す．従来の音声対話コーパスは話者ごとに音声が分離されていないものが多く，信号処理を用いた分析には扱い難い．そのような問題を回避するため，対話者を別々の防音室に配置し個別のダイナミックマイク (AT4055) を使用し各話者の発話を異なるチャネルで録音した．また，話者の表情やジェスチャーが映るように調整したビデオカメラ (GoPro HERO7) で話者の上半身を撮影し，撮影した動画像と音声をもう一方の防音室のディスプレイとヘッドフォンに互いに出力した．動画像の伝送にはほとんど遅延がなく，対話者は大きな違和感を感じることなく対話を行うことができた．

対話者は「対話相手とより親しくなること」を目的として対話を行った．対話の話題として，Jourard によって提案された 60 個の自己開示項目 [6] のうち “Work (or studies)” と “Tastes and interests” を参考に，対話の目的を達成するのに適切であると考えられる 10 個の話題を用意した．表 1 に選択した話題をまとめる．それぞれの対話者は収録前に自分が話したいと思う話題を 5 つ選び，最終的な話題は両者の選択を基準にして実験者が決定した．

対話収録の参加者は日本人の大学生・大学院生であり，本稿では収集したデータのうち 19 名 (男性 15 名，女性 4 名) の対話を分析に用いる．ここで，1 つの話題に関する対話を 1 対話とした．結果として，全 19 ペアから計 95 対話 (約 10 時間) のデータが収集された．図 2 に収録された対話の例を示す．

2.2 話者間の親密性を表すラベル

収録前に対話者は (1) 対話相手と面識があるかどうかと (2) 対話相手との付き合いの長さを回答し，(3) 相手との仲の良さに関する 5 段階 (1: 全く当てはまらない，5: 非常に当てはまる) の主観評価を行った．項目 (2) と (3) については，面識のあるペアのみが回答した．ここで，項目 (2) の相手との付き合いの長さの平均値と標準誤差は 0.88 ± 0.32[年] であった．また，項目 (3) の親密性に関する主観評価のスコアの平均値と標準誤差は 4.00 ± 0.161 であった．そのため，相手と面識のあるペアは概ね親しい関係であることが示唆される．

図 2: 収録された対話の例

表 2: 付与した対話行為タグ

タグ名	内容	実際の発話例
自己開示	嗜好や感情を開示している発話	豆全般もあんま得意じゃない。
情報提供	客観的な情報を伝えている発話	まあ山に登るサークルですね。
提案	相手に対する何らかの提案を表す発話	あれは観た方がいいですよ。
要求	相手に対する何らかの要求を表す発話	えーやめてよー。
確認	相手が伝聞・理解したことを確認する発話	ああ、そうなの？
質問	相手の返答を期待した発話	最近漫画、他に買ってる？
同意・共感	相手の意見への同意・共感した発話	うん。 / 確かにそんな感じしますもんねー。
不同意・不共感	相手の意見への不同意・共感していない発話	え似合わないことないとおもうけどな。
相槌	対話相手の発話を促す発話	はいはい。 / うんうん。
感嘆	感心・驚きを表す発話	えっ。 / はあー。
フィラー	意味を持たないが間をつなぐための発話	えっとー。 / なんか。
その他	以上の対話行為のどれにも属さない発話	ホヤって。(発話が中断)

2.3 発話内容の書き起こしと対話行為タグの付与

収録した対話は 5 名のクラウドワーカが書き起こし，第一著者が表記ゆれや句点の誤りを修正した．“意味が完結しているまとまり” を基準として発話に分割したところ，19,303 発話が得られた．

各発話に対しては 1 つの対話行為タグを付与した．本研究では SWBD-DAMSL タグ [7]，JAIST タグ付き自由対話コーパスに用いられたタグ [8]，そして聞き役対話の分析に用いられたタグ [9] を参考に表 2 に示す 12 種類のタグを設定した．また，対話行為タグの付与は第一著者が行った．今後より詳細な分析を行うため，タグセットを精緻化する予定である．

2.4 表情ラベルの付与

各発話に対しては表情の推定ラベルを付与した．収集したデータにおいて参加者の表情の多くは笑顔もしくはニュートラルな表情であり，驚き・怒りのような表情はほとんど出現しなかった．そのため，話者の表情が笑顔もしくはその他の表情のどちらであるかを識別する表情認識器を実装し，表情のラベルを推定した．実装した表情認識器は OpenFace[10] によって検出された顔領域を Convolutional Neural Network (CNN) の入力として「笑顔」もしくは「その他」の二値分類を行う．CNN の事前学習には Facial Expression Recognition 2013 (FER2013) データセット [11] を用いた．学習には training set 及び public test set，評価には private test set を用いた．評価データに対する表情分類の再現率は「笑顔」で 71.3%，「その他」で 91.5%，平均して 81.4%であった．表情認識はフレームごとに行い，各発話の表情ラベルは発話区間内の表情認識結果の多数決を用いた．

表 3: 分析データの内訳

親密性	ペア数	対話数	発話数
低	8	40	7218
高	11	55	12085

3 親密性による対話コーパスの分析

面識のあるグループは概ね親密な関係であることが示唆されたため，「初対面」を親密性が低いグループ，「非初対面」を親密性が高いグループとして，この 2 グループ間における対話行為の出現頻度と韻律・表情の傾向を分析した．分析対象となる対話データの内訳を表 3 に示す．分析は対話単位で行った．

3.1 対話行為の出現確率

対話行為の出現確率について分析を行った．各グループにおいて式 (1) によって表される出現確率を計算した．

$$P(a_i) = \frac{C(a_i)}{N} \tag{1}$$

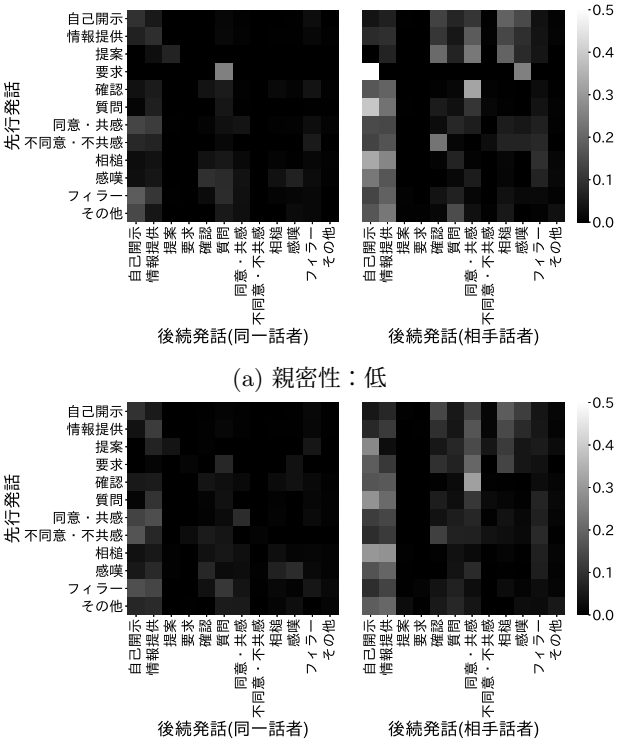
ここで a_i は各対話における i 番目の発話の対話行為ラベル，$C(a_i)$ は各対話における a_i の出現回数，N はそれぞれの対話内の総発話数を表している．各対話行為の出現確率をそれぞれのグループに対して計量した結果を表 4 に示す．表 4 にはそれぞれの対話行為の出現確率の平均値と標準誤差，2 群間での対応のない t 検定の結果を示した．結果として，自己開示・要求・質問において有意差が見られた．親密性の高いグループにおいては，親密性の低いグループに比べて自己開示の出現確率が小さいことが示された．社会浸透理論 [12] では二者間の親密性が向上するにつれて個人の内面的な自己開示が増えるとされているが，本稿で設定した話題には “好き，もしくは嫌いな食べ物・飲み物” などの表層的な内容が多いため，より親密なグループでは自己開示が少なかったと考えられる．また，親密性の

表 4: 各対話行為の出現確率 [%] (平均値 ± 標準誤差)

対話行為	親密性：低	親密性：高	p 値
自己開示	21.07±0.96	17.88±0.67	0.008^{**}
情報提供	19.42±1.11	22.13±0.99	$0.072^{\dagger}$
提案	0.27±0.08	0.38±0.08	0.315
要求	0.06±0.03	0.19±0.05	0.024^{*}
確認	7.28±0.52	7.47±0.42	0.776
質問	8.78±0.54	6.88±0.44	0.008^{**}
同意・共感	11.53±0.78	13.27±0.59	$0.079^{\dagger}$
不同意・不共感	0.45±0.10	1.04±0.28	$0.050^{\dagger}$
相槌	17.22±1.04	17.93±1.24	0.663
感嘆	6.55±0.56	5.33±0.37	$0.073^{\dagger}$
フィラー	6.20±0.60	5.93±0.37	0.709
その他	1.16±0.16	1.54±0.20	0.131

$^{\dagger}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01$

高いグループでは親密性の低いグループに比べて対話相手への要求や不同意がより多く出現した．関係が構築されると，比較的抵抗なく対話相手に対する要求や反対意見を述べやすくなるためだと考えられる．一方で親密性の低いグループにおいては対話相手に対する質問が増加した．相手と面識が無い場合は相手の情報が乏しいため積極的に対話相手の情報を得ようとするためであると考えられる．



(a) 親密性：低

(b) 親密性：高

図 3: 対話行為の遷移確率の分布

3.2 対話行為の遷移確率

対話行為の遷移確率について分析を行った．各グループにおいて式 (2) によって表される遷移確率を計算し，それぞれの遷移に関して比較した．

$$P(a_{i+1}|a_i) = \frac{C(a_i, a_{i+1})}{C(a_i)} \tag{2}$$

ここで $C(a_i, a_{i+1})$ はそれぞれ対話において出現する対話行為 a_i と，a_i から a_{i+1} への対話行為の遷移の出現回数を表している．

3.2.1 遷移確率の分布

図 3 に各グループの遷移確率の分布を示す．図 3 より親密性の低いグループと親密性の高いグループの遷移確率の分布は類似しており，どちらのグループも大まかには同様の対話を行っていることが示唆される．

3.2.2 頻出する遷移に関する比較

続いて，個々の対話行為タグの遷移に着目して分析を行った．本稿では，各グループにおいて出現しやすい遷移を比較するため遷移確率の大きい上位 10 項目を選出した．表 5 に選出した遷移確率の平均値と標準誤差及び 2 群間での対応のない t 検定の結果を示す．なお，選出された遷移は全て一方の話者から相手話者への遷移であった．結果より “質問から自己開示への遷移” において有意水準 1%で有意差が見られた．親密性の低いグループにおいては質問に対してそのまま自己開示をするといったような単純な構造の対話が多いことが示された．

3.3 交替潜時の比較

話者間の親密性が交替潜時に影響するかを調べるために，2 つのグループ間で交替潜時を比較した．本稿では一方の話者の発話が終了してから他方の話者の発話が開始されるまでの時間を交替潜時として扱った．ここで，発話がオーバーラップしている場合，交替潜時は負の値をとる．親密性の高いグループの対話の方がより円滑な対話が行われると考えられるため，親密性の高いグループの交替潜時の平均値は親密性の低いグループに比べ小さくなると予想できる．

各グループにおける交替潜時の平均値と標準誤差を表 6 に示す．表 6 より親密性の低いグループに比べて親密性の高いグループの交替潜時の平均値は小さかったが，対応のない t 検定の結果，有意差は見られなかった．

表 5: 上位 10 項目の対話行為遷移確率 (平均値 ± 標準誤差)

$a_i \rightarrow a_{i+1}$	親密性：低	親密性：高	p 値
要求 → 自己開示	0.50±0.29	0.18±0.09	0.361
質問 → 自己開示	0.39±0.03	0.29±0.02	0.009^{**}
確認 → 同意・共感	0.32±0.03	0.31±0.02	0.809
相槌 → 自己開示	0.33±0.03	0.30±0.02	0.265
相槌 → 情報提供	0.26±0.03	0.29±0.02	0.652
その他 → 情報提供	0.23±0.06	0.20±0.04	0.674
質問 → 情報提供	0.22±0.03	0.21±0.02	0.652
感嘆 → 自己開示	0.24±0.03	0.16±0.01	0.046^{*}
提案 → 同意・共感	0.24±0.11	0.15±0.07	0.505
自己開示 → 相槌	0.19±0.01	0.18±0.01	0.543

$^{\dagger}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01$

表 6: 交替潜時の分布 (平均値 ± 標準誤差)

親密性：低	127.8±12.5[ms]
親密性：高	109.6±8.9[ms]

3.4 発話のピッチ及びパワーの比較

一般に，対話が進むにつれて二者間の発話のパワーやピッチが同調することが知られている [13, 14]．話者間の親密性によってこの現象に違いが生じるかどうか分析した．河原らの研究 [15] を参考に，発話交換における発話間のパワーと対数基本周波数 (F0) の相関を分析した．ここでは，一方の発話の末尾 500[ms] 部分と，その発話に対する応答の先頭 500[ms] を分析区間とした．分析の手順を以下に示す．

Step 1 各分析区間で，フレーム幅 25[ms]，フレームシフト 10[ms] でパワー及び対数 F0 を算出

Step 2 各分析区間におけるパワー及び対数 F0 の平均値・最大値を算出し，分析特徴量とする

Step 3 ピッチやパワーの表出傾向は発話の末尾と先頭において異なると考えられるため，それぞれに関して話者ごとに標準化

Step 4 標準化された特徴量に関して，ペアごとに相関を計算

なお，F0 の抽出には Python のライブラリである pyworld[1] を使用し，分析区間において F0 が抽出されなかった部分は除いた．

表 7 に各グループの対話におけるパワーまたは対数 F0 の相関係数の平均値と標準誤差を示す．また，表中には 2 群間での対応のない t 検定の結果を示した．結果より，パワーの平均値と最大値の相関係数には有意差が得られ，対数 F0 の平均値の相関係数において有意傾向が得られた．いずれの分析特徴量においても親密性の高いグループの方が親密性の低いグループよりも相関係数の平均値が大きかったものの，各相関係数の値は相関が高いといえる程大きくなかった．音声のパワーに関しては，話者の姿勢や顔の向きの影響が考えられるが，一部の対話に関して動画像を確認したところ話者の姿勢や顔の向きのパワーに対する影響は小さかった．一方で，親密性の高いグループにおいては笑い声の同調現象が起こりやすかったと考えられるため，お互いに笑っている発話区間がパワーの相関に影響している可能性がある．そのため，今後はこのような笑い声の発生傾向の違いを調査し，音声区間と切り分けて分析を行う予定である．

[1]https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder

表 7: 話者間のパワー・対数 F0 の相関係数 (平均値 ± 標準誤差)

分析特徴量		親密性：低	親密性：高	p 値
パワー	平均値	0.010±0.071	0.112±0.080	0.013^{*}
	最大値	0.016±0.055	0.151±0.077	$<0.001^{**}$
対数 F0	平均値	0.116±0.131	0.239±0.123	$0.070^{\dagger}$
	最大値	0.079±0.205	0.129±0.125	0.575

$^{\dagger}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01$

3.5 表情の比較

表 8 に各対話行為での笑顔の出現割合の平均値と標準誤差を示す．また，表 8 には 2 群間での対応のない t 検定の結果も示した．結果より，多くの対話行為において親密性の高いグループで笑顔の出現割合が増加することがわかる．また相槌や不同意・不共感などの対話行為では有意差は見られなかったものの，多くの対話行為では有意差または有意傾向がみられる．また表 9 に一方の話者から他方の話者に発話が遷移した場合の表情の遷移確率を示す．結果より，笑顔の同調は親密性の高いグループにおいて頻繁に発生し，その遷移確率は親密性の低いグループの遷移確率の約 2 倍となることが分かった．このことから，表情の出現傾向は話者間の親密性に依存して変化することが示唆される．

4 考察

対話行為の出現頻度, 遷移確率において話者間の親密性の違いによって傾向が異なることが示唆された．分析結果より，ユーザの印象向上のためには初対面のユーザに対しては標準的な質問応答を行うが，ユーザがシステムに慣れるにしたがって質問応答の間に意見を述べるような逸脱のある対話を行うことが有用であると考えられる．また，話者の表情を分析した結果，対話行為によって表情の出現割合は異なり，親密な関係にお

表 8: 各対話行為における笑顔の発話の割合 (平均値 ± 標準誤差)

対話行為	親密性：低	親密性：高	p 値
自己開示	0.114±0.019	0.231±0.030	0.001^{**}
情報提供	0.098±0.019	0.224±0.030	$<0.001^{**}$
提案	0.038±0.028	0.150±0.048	0.045^{*}
要求	0.000±0.000	0.048±0.027	$0.073^{\dagger}$
確認	0.158±0.032	0.237±0.033	$0.088^{\dagger}$
質問	0.078±0.017	0.217±0.031	$<0.001^{**}$
同意・共感	0.150±0.025	0.252±0.035	0.018^{*}
不同意・不共感	0.146±0.054	0.174±0.040	0.681
相槌	0.193±0.028	0.233±0.035	0.372
感嘆	0.074±0.016	0.247±0.035	$<0.001^{**}$
フィラー	0.089±0.026	0.250±0.037	$<0.001^{**}$
その他	0.025±0.015	0.189±0.046	0.001^{**}

$^{\dagger}p < 0.10, ^{*}p < 0.05, ^{**}p < 0.01$

表 9: 表情の遷移確率 (平均値 ± 標準誤差)

表情の遷移 (先行 → 後続)	親密性：低	親密性：高
笑顔 → 笑顔	0.133±0.024	0.308±0.029
笑顔 → その他	0.867±0.024	0.692±0.029
その他 → 笑顔	0.119±0.016	0.202±0.026
その他 → その他	0.881±0.016	0.798±0.026

いては話者の表情は笑顔になりやすいことが分かった．したがって，ユーザとシステムの関係が親密である場合は対話エージェントの笑顔の割合を増加させ，対話行為に応じて表情の出現割合を変化させるといった制御が有効であると考えられる．

5 おわりに

本研究では，ユーザとの親密性を考慮した対話システムの実現を目標として，マルチモーダル雑談コーパスの構築を行った．分析によって，話者間の親密性によって対話行為や表情の出現傾向が異なることが明らかになった．今後は，コーパスの規模を拡大するため新たに 100 ペアのデータを収録する予定であり，現在 60 ペア分の収集が完了している．また，本稿で分析により有意差が得られなかった交替潜時や韻律情報に関して，データ数を拡大したコーパスを用いてあらためて分析を行う．同時に，分析から示唆された対話制御手法を対話システムに導入し，ユーザ評価によりその有効性を評価する予定である．

謝辞

本研究は，JST COI JPMJCE1303 及び科学研究費補助金 (課題番号 JP18K18136) の助成を得た．

参考文献

[1] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, “Building end-to-end dialogue systems using generative hierarchical neural network models,” in *Proc. Thirtieth AAAI Conference on Artificial Intelligence*, pp. 3776–3783, 2016.

[2] D. Bohus and E. Horvitz, “Managing human-robot engagement with forecasts and... um... hesitations,” in *Proc. the 16th International Conference on Multimodal Interaction*, pp. 2–9, 2014.

[3] Y. Chiba, T. Nose, and A. Ito, “Cluster-based approach to discriminate the user’s state whether a user is embarrassed or thinking to an answer to a prompt,” *Journal on Multimodal User Interfaces*, vol. 11, no. 2, pp. 185–196, 2017.

[4] Y. Kageyama, Y. Chiba, T. Nose, and A. Ito, “Improving user impression in spoken dialog system with gradual speech form control,” in *Proc. SIGDIAL*, pp. 235–240, 2018.

[5] Y. Kim, S. S. Kwak, and M. Kim, “Am I acceptable to you? Effect of a robot’s verbal language forms on people’s social distance from robots,” *Computers in Human Behavior*, vol. 29, pp. 1091–1101, 2012.

[6] S. M. Jourard and P. Lasakow, “Some factors in self-disclosure,” *The Journal of Abnormal and Social Psychology*, vol. 56, no. 1, pp. 91–98, 1958.

[7] D. Jurafsky, E. Shriberg, and D. Biasca, “Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual,” tech. rep., Institute of Cognitive Science, 1997.

[8] K. Shirai and T. Fukuoka, “JAIST annotated corpus of free conversation,” in *Proc. LREC*, pp. 741–748, 2018.

[9] T. Meguro, R. Higashinaka, K. Dohsaka, Y. Minami, and H. Isozaki, “Analysis of listening-oriented dialogue for building listening agents,” in *Proc. SIGDIAL*, pp. 124–127, 2009.

[10] T. Baltrušaitis, P. Robinson, and L.-P. Morency, “OpenFace: an open source facial behavior analysis toolkit,” in *Proc. WACV*, pp. 1–10, 2016.

[11] I. J. Goodfellow, D. Erhan, P. L. Carrier, *et al.*, “Challenges in representation learning: A report on three machine learning contests,” in *Proc. International Conference on Neural Information Processing*, pp. 117–124, 2013.

[12] I. Altman and D. A. Taylor, *Social penetration: The development of interpersonal relationships*. New York: Holt, Rinehart & Winston, 1973.

[13] M. Natale, “Convergence of mean vocal intensity in dyadic communication as a function of social desirability.,” *Journal of Personality and Social Psychology*, vol. 32, no. 5, pp. 790–804, 1975.

[14] S. Gregory, S. Webster, and G. Huang, “Voice pitch and amplitude convergence as a metric of quality in dyadic interviews,” *Language & Communication*, vol. 13, no. 3, pp. 195–217, 1993.

[15] T. Kawahara, T. Yamaguchi, M. Uesato, K. Yoshino, and K. Takanashi, “Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening,” in *Proc. APSIPA-ASC*, pp. 392–395, 2015.