人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B303-02

発話者の音声に対応する動作生成と遠隔操作ロボットへの動作の付加効果

Online speech-driven head motion generation system and evaluation on a tele-operated robot

○境　くりま [*1,2], 石井　カルロス寿憲 [*2], 港　隆史 [*2], 石黒　浩 [*1,2]
Kurima SAKAI[*1,2], Carlos Toshinori ISHI[*2], Takashi MINATO[*2], Hiroshi ISHIGURO[*1,2]
ATR[*1], 大阪大学大学院 基礎工学研究科 [*2]
sakai.kurima@irl.sys.es.osaka-u.ac.jp, carlos@atr.jp, minato@atr.jp, ishiguro@sys.es.osaka-u.ac.jp

Abstract

本論文では，遠隔操作対話ロボットの頭部動作を操作者の音声情報のみから自動生成するシステムを提案する．遠隔対話では発話音声と一致した頭部動作の表現が必要となるため，発話の意味（相槌や発話の保持などの談話機能）を言語情報と韻律情報を用いてをリアルタイムで推定し，推定した談話機能に基づき頭部動作を生成する．提案システムには推定誤りが含まれ，対話に適さない動作が生成される場合がある．そのため，提案システムを用いた対話時の動作の印象を被験者実験により評価した．主観評価から，提案システムによる動作を付加することで，ロボットの動作がより対話に適したものになることが示された．

1　はじめに

電話やインターネットなどの通信技術の発達により，遠隔地にいる人といつでもどこでも対話することが容易になってきたが，そのようなコミュニケーションメディアを介した対話では，対話相手と対面しているように感じられず，円滑な対話が阻害される．円滑な遠隔対話の実現は，円滑な人間関係の構築につながる重要な課題である．我々は遠隔操作ロボットを用いて場の共有感や身体動作といった非言語情報を伝達することで，円滑な遠隔対話の実現を目指している．遠隔操作ロボットによる対話システムでは，操作者が遠隔地にいる人型ロボットを操作することで，ロボットが操作者の音声と身体動作を表現する．対話相手はそのロボットと対話することで，操作者と対面しているように感じながら対話を行うことができる（図 1）．

ロボットの身体動作の生成手法は，操作者の動きを計測しロボットの関節角にマッピングする手法が一般的である．しかし，この手法では「対話コンテキストに一致しない動作」と「ロボットの身体的制限による不完全な動作」の 2 つの動作が問題となる．例えば，前者の問題は，操作者が対話中に操作者の部屋の時計を見ると，ロボットも同じ動作をするが，遠隔地では操作者の部屋と同じ位置に時計があるわけではないため，対話者は動作の意味が分からなくなるという問題である．また，後者の問題は，ロボットの腕が可動範囲の制約で頭まで上がらないにもかかわらず，操作者が頭をかく動作を行うと，ロボットは不完全に動作を再現するため，対話者は意味が理解できなくなるという問題である．これら問題は，遠隔対話を阻害する要因となる．

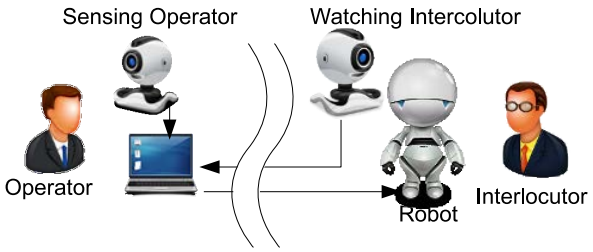


図 1:　Overview of the tele-communication by tele-operated robot.

解決策として，対話者が意味を理解できるようにロボットの動作を変換すればよいが，音声はそのまま伝達されるため，動作の変換が不適切で動作の意味が音声の意味と一致しないと，逆に対話者の理解を妨げることになる．音声との一致という制約を考慮すると，音声の意味を推定し，ロボットが表現可能な動作で表現すればよい．この方針の下，遠隔操作ロボットを用いた対話を行う際にシステムが満たすべき条件は，発話音声から発話の意味と一致する動作をリアルタイムで生成することである．本論文では，音声と身体動作の関係性の知見を基に，リアルタイムで発話の意味を推定し動作を生成するシステムを構築する．

音声情報から動作を生成するシステムはいくつか提案

されている．オフライン処理システムでは，ユーザの音声の韻律情報から，コンピュータグラフィックス上のアバタの動作を生成する試みがある．Sargin et al.[1] や Busso et al.[2] は，ユーザの音声の韻律情報から，アバタの頭部動作を生成するシステムを提案している．韻律情報に加え，Foster et al.[3] は対話コーパスを用いる手法を提案している．ロボットの動作生成では，Ishi et al.[4, 5, 6] は日常会話の頭部動作と談話機能（発話の意味）の関係性に基づく頷きと首傾げ動作生成モデルを提案している．

しかし，リアルタイムで音声から動作を生成する研究は少ない．Le et al.[7] は音声の韻律情報（ピッチとパワー）と頭部動作（首の角度，速度，加速度）を機械学習を用いて事前に学習し，リアルタイムで頭部動作を生成するシステムを提案している．Watanabe et al.[8, 9] は，音声の ON-OFF 情報を入力とした頷き動作の予測モデルを構築し，コンピュータグラフィックスのエージェントやヒューマノイドロボットの頷き動作をリアルタイムで生成するシステムを構築した．

これらの従来研究は，発話の意味に基づく動作生成とリアルタイム性を両立するものはない．Watanabe et al.[8] の手法では，頷き動作しか生成できず，発話者の否定の意味や困惑などを表現することができない．また，日本語（本論文の対象）では韻律特徴と頭部動作の相関が低いことが報告されている [10]．そのため，Le et al.[7] の手法のように，韻律情報を用いる手法は日本語音声から動作を生成する場合には適さないと考えられる．Ishi et al.[6] の手法を用いれば，発話の意味に合う動作が生成できると考えられる．彼らは，音声に手動で談話機能をラベリングしモデルを評価したが，リアルタイムで動作を生成するシステムは構築していない．そのため，本論文では発話者の音声からリアルタイムで談話機能を推定し，ロボットの頭部動作を生成するシステムを提案する．

2　頭部動作システムの構築

提案システムは，操作者の発話音声から談話機能を推定し，推定した談話機能に基づき頭部動作を生成する．本節では，まず Ishi et al.[4, 5, 6] の談話機能に基づく頭部動作生成モデルを説明する．次に，発話音声から談話機能を推定するための発話音声と談話機能の関係性，言語情報の抽出手法と韻律情報の抽出手法を説明する．最後に，言語情報と韻律情報を組み合わせ動作を生成するシステムを説明する．

2.1　談話機能と頭部動作の関係性

Ishi et al. はマルチモーダル対話音声データベースを用いて，頭部動作と談話機能の関係性を解析した [4, 5]．データベースには以下の Ishi et al. が提案した談話機能タグ [11, 12] が付与されている．

- k(keep): 発話権の保持（ポーズないしはっきりしたピッチのリセットが伴う強い句境界）
- k2(keep): 発話権の保持 2（発話文の中にある弱い句境界）
- k3(keep): 発話権の保持 3（話者が発話末の音節を伸ばし，考えていることや発話の途中であることを表現する場合）
- f(filler): 「えっとー」「あのー」などの感嘆詞を伴った，考え中であることの表現（フィラー）
- f2(conjunctions):「じゃ」などの感嘆詞を伴った，考え中であることの表現（短いフィラー）
- g(give): 発話権の譲渡（当話者の発話が終了し，発話権を対話相手への譲渡する場合）
- q(question): 発話権の譲渡 2（対話相手確認するなど応答を求める場合）
- bc(backchannels): 「うん」「はい」などの感嘆詞を伴った相槌の表現
- su(admiration/surprise/unexpectedness): 「へー」「うそ！」「ああ！」などの感嘆詞を伴った，驚きや感心の表現
- dn(denial, negation): 「いいえ」「ううん」などの感嘆詞を伴った，否定の表現

談話機能と頭部動作の関係性の分析によると，頷き動作が対話の中で最も多く生起し，特に頷き動作は相槌 (bc) や強い句境界 (k,g,q) で多く見られることが報告されている [4]．また，発話者が考えていたり，次の発話の準備をしているなどの場合では，語尾を延ばすことが多い．それら弱い句境界 (f,k3) では，首かしげ動作が最も多く出現したことも報告されている．さらに，驚きや感心を表す感嘆詞 (su) においても，顔上げ動作や首かしげ動作が頻繁に見られる．

2.2　談話機能と発話の関係性

日本語対話における談話機能と言語情報・韻律情報の関係性は分析されている．相槌 (bc) は「うん」「ええ」「ああ」「はい」のような感嘆詞を下降調トーンで発話することが多く，驚きや感心 (su) は「ええ」「へえ」「うん」を上昇調トーンで発話することが多く，フィラー (f) は「ええ」「へえ」「ううん」などを長く平坦調に発話することが多いことが報告されている [11, 12]．発話権を保持する強い句境界 (k) では,「で」「から」「けど」などの接続助詞を伴い，句末の音節のトーンが大きく下降する傾向にあることが報告されている [4]．これらの知見に基づき本論文では，“bc”，“su”，“f”，“k” の談話機能を推定し，頭部動作を生成するシステムを構築する．

2.3 言語情報の抽出

2.2 節で説明したように，談話機能の推定には言語情報と韻律情報が必要となる．本論文では，オープンソースである大語彙連続音声認識エンジン Julius[13] を用いて操作者の音声から言語情報を抽出する．Julius に付属する音響モデルは読み上げ音声を用いて作成されている．しかし，自然会話の「ああ」「ううん」「ええ」などの感嘆詞は，はっきり発音されないことが多いため，付属の音響モデルでは正しく認識することが困難である．感嘆詞の認識率を向上させるために，我々は自然対話データベースの音声データから感嘆詞を抽出し音響モデルを作成した．音響モデルの学習には，4406 フレーズ（男性:1903，女性:2503）の音声を用いた（「ああ」「ええ」「はい」「はあ」「へえ」「ほお」「うわあ」「わあ」「ううん」「いや」「いいえ」）．感嘆詞のモノフォン HMM（隠れマルコフモデル）モデルの作成には HTK(`http://htk.eng.cam.ac.uk/`) を用いた．韻律特徴は 12 MFCC（メル周波数ケプストラム），12 delta-MFCC，1 delta-power を用い，HMM の状態数は各感嘆詞の音素数にあわせ 8~16 とした．

遠隔対話における自由会話の言語モデルを作成することは困難であるため，本論文では言語モデルは検出したい相槌 (bc) や感心 (su) やフィラー (f) で見られる感嘆詞をキーワードとし，それ以外の音節を組み合わせる記述文法を用いた．また，漸次認識結果に音素アライメントを出力させるよう Julius のソースコードの改良も行った．

2.4 韻律情報の抽出

2.2 節で説明したように，談話機能を推定するには韻律情報（音調）も必要となる．そのため，基本周波数 (F0) の抽出を用いて，音調の識別を行った．

まず，F0 の値の抽出には，32 ms のフレーム幅で 10 ms 毎に LCP(Lear Predictive Coding) 逆フィルタによる残差波形の自己相関関数の最大ピークに基づいた処理を行う．さらに，人間のイントネーションの知覚特性と一致するよう，F0 の値を対数スケールに変換した．

$$F0[\text{semitone}] = 12 \times \log_2(F0[\text{Hz}]) \tag{1}$$

次に，音節内で F0 の変化量を表す $F0move$(人間の音調の知覚に基づくパラメータ [14]) を抽出した．$F0move$ は音節の後半の F0 の近似直線上の音節末の F0($F0tgt2b$) と前半部の F0 平均値 ($F0avg2a$) との差分を用いて計算する（式 2）．そして，音節の音調は式 3 に応じて，上昇調，下降調，平坦調に分類した．

$$F0move = F0tgt2b - F0avg2a \tag{2}$$

$$tone = \begin{cases} rising\ (Rs) & (F0move > 1 \text{ semitone}) \\ falling\ (Fa) & (F0move < -2 \text{ semitones}) \\ flat\ (Ft) & (\text{otherwise}) \end{cases} \tag{3}$$

2.5 リアルタイム音声駆動頭部動作生成システム

図 2 に実装したシステムの概要を示す．adintool（Julius に付属）はマイクロフォンから操作者の音声信号を取得し，音声のパワーと零交差に基づき音声区間のセグメンテーションを行う．音声情報は言語情報を取得するために Julius に送られ，また韻律情報を取得するために F0 抽出部へ送られる．リアルタイムで処理するために，Julius は 100 ms 毎に漸次認識結果を出力する．F0 値は 10 ms 毎に取得できる．動作生成部では，Julius からの音節アライメント情報に基づき，F0 情報を用いてキーワード区間の音調を識別する．抽出した言語情報と韻律情報に基づきロボットの頭部動作を生成し，ロボットにモータコマンドを送信する．すべてのモジュール間のデータ通信は TCP/IP を用いた．

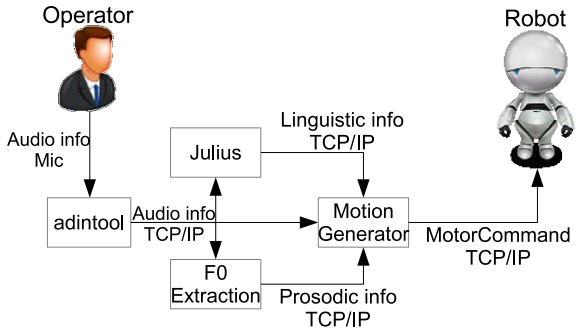


図 2: Overview of our online speech-driven head motion generation system.

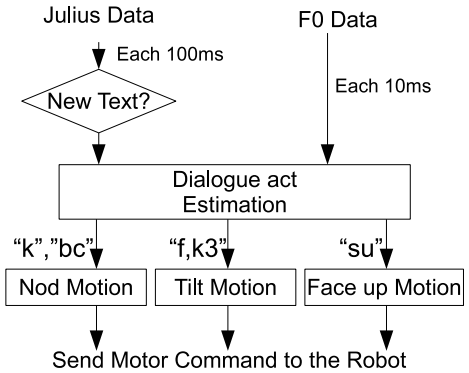


図 3: The system flow of the Motion Generator module.

図 3 に動作生成部のシステムフローを示す．Julius からの漸次認識結果は 100 ms 毎に取得できるため，新しく認

識された言語のみ処理する必要がある．そのため，漸次認識結果の最後の音節がすでに処理されたものかを確認するルールを設け，“bc”，“f”，“k3”，“su” は 100 ms 毎に処理される．一方で，“k” は Julius が発話終わりまたはショートポーズを認識した際に処理される．

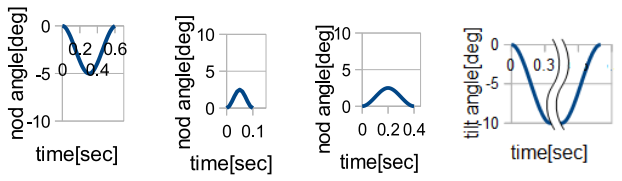
以下に各談話機能の推定と動作生成のルールを説明する．生成する動作の大きさと時間は従来研究の分析に基づいたものを用いた [4, 5]．

“bc” の推定は，「ああ」「あー」「ええ」「はい」「はあ」「へえ」「ほお」「ううん」「うん」といった感嘆詞が認識され，その発話区間の音調が下降調である場合とする．その際，図 4(a) に示す頷き動作を生成する．従来研究では，頷き動作には微かに頭部を上げてから下げる動きが観測されている [4]．しかし，首上げ動作は小さく，また動作生成の遅延をできる限り小さくするために，本論文では首上げ動作はないものとして首下ろし動作のみ実装した．

“k” の推定には，本来は「で」「から」「けど」などの接続助詞の抽出が必要になるが，現状の音声認識では接続助詞の抽出は現状困難である．そのため，句末やショートポーズの前の音節が下降調に発話された場合を “k” とした．

“su” の推定は，「ええ」「へえ」「ほお」といった感嘆詞が認識され，その発話区間の音調が上昇調である場合とする．その際に 2 種類の動作を生成する．1 つは図 4(b) に示す驚きを表す動作であり，他方が図 4(c) に示す感心を表す動作である．驚き動作は発話区間が短い場合に生成し，感心動作は発話区間が長い場合に生成する．

“f” と “k3” の推定では，一般的な会話での母音区間の長さは 200~300 ms であるため，1 音節の閾値を 350 ms とし，1 音節が閾値以上長く平坦調である場合を “f” と “k3” した．この際生成される首傾げ動作を図 4(d) に示す．従来研究 [6] 同様に，首傾げ状態は発話が終わるまで維持される．



(a) (b) (c) (d)

図 4: Head rotation angle shapes used in our head motion generation system: (a)Nod Motion; (b)Face up Motion (surprise); (c)Face up Motion (admire); (d)Tilt Motion.

3 遠隔操作ロボットを用いた提案システムの評価実験

3.1 実験目的

本節では提案システムの実用性を評価する．従来研究において，談話機能を手動でラベリングし，談話機能の認識に誤りがない場合，談話機能に基づき生成された頭部動作が自然と評価された [6]．しかし，提案システムは言語情報の制限や音声認識の誤りによる談話機能の推定に誤りがあるため，対話に不適切な動作が生成される可能性がある．それらは対話相手に対話コンテキストに適さない不自然な動作という悪い印象を与えてしまう．そのため，本実験では提案システムを用いた対話条件と用いない対話条件を比較し，提案システムにより生成される動作の対話への適切さを評価した．

3.2 実験設定

本実験では遠隔操作ロボットテレノイドを用いた（図 5）．テレノイドは首に 3 自由度（ピッチ軸，ロール軸，ヨー軸）と口に 1 自由度（上下開閉）のアクチュエータを持つ．また，テレノイドの外見は操作者に対する印象形成を乱さないような外見になっているため [15]，テレノイドを用いることで操作者の音声とロボットの外見との不適合による印象への影響を軽減することができる．

提案システムは操作者が発話しないとロボットが全く動かないため，正規雑音モデルに従って生成した指令値を首関節に常に入力し，人間の微小な不随意運動を表現した．500 ± 150N（N は標準偏差 0.1 の正規雑音）msec かけて 0.5 ± 0.5N（N は標準偏差 0.1 の正規雑音）度，首の 3 回転軸（ピッチ軸，ヨー軸，ロール軸）を移動させ，同じ時間をかけて元の位置に戻るというものである．不随意動作を頷き・首傾げ・首上げ動作に解釈されることを避けるため，約 1 秒のゆっくりした動きを用いた．また，テレノイドの口の動きは発話者音声に基づいて口唇動作を制御する手法 [16] を用いて，操作者の音声のみから口の開閉動作を生成する．

本実験では，操作者（実験協力者）が操作するテレノイドと対話する人が被験者となり，テレノイドの動作の印象を評価した．操作相手に対する印象を統制するために，操作者は 1 人（女性）とした．

被験者はテレノイドを通して操作者と対話し動作の適切さを評価した．比較条件は，テレノイドが口の開閉動作に加え上記の不随意運動を行う条件 (Noise) と，さらに提案手法による頭部動作を加えた条件 (Voice+Noise) である．被験者の話題に対する興味を統制するため，事前に被験者に「好きなスポーツ」「好きな映画」「好きな本」の 3 つから対話のテーマを 2 つ選ばせた．また，被験者自身が対話テーマを選ぶため，被験者に話し始めるよう教示した．この教示により被験者が話を率先することになり，被

験者はテレノイドから伝わる非言語情報から自分の話題に対する操作者の反応を把握しようとすると考えられる．このようにして，被験者にテレノイドから伝わる非言語情報に自然に注意を向けるように仕向けた．被験者には 2 条件とも体験してもらい被験者内比較を行った．また，操作者には自分がどちらの条件でテレノイドを操作しているのかわからないようにした．

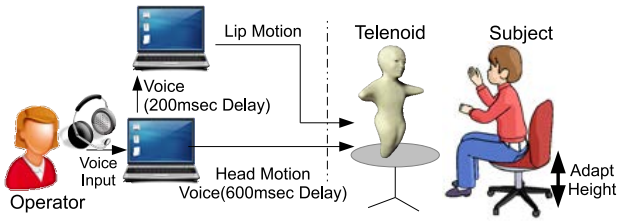


図 5: Experiment setup for evaluation of the proposed system using a tele-operated robot.

3.3 実験システム

図 5 に実験システムを示す．音声と動作を同期させるために 600 ms の音声遅延を設けテレノイド内のスピーカーから出力する．操作者の反応が対話者に 600 ms 送れて伝わるため，対話者の反応も 600 ms 程度遅れて操作者に伝わることになるが，予備的試行では，操作者はこの程度の遅れがあっても自然に振る舞うことができていた．また，Ishi et al.[16] の手法による口動作の生成の遅延は 400 ms の音声遅延が必要となる．そのため，頭部動作に合わせ口動作を生成するために，口動作を生成する PC には 200 ms の遅延をかけて操作者の音声を入力する．ロボットの見え方を統制するために，被験者は椅子の高さを調節しテレノイドと視線を合わせた（図 5）．

3.4 実験手順と評価指標

実験手順を以下に示す．

1. 対話ロボットの説明と遠隔操作の実演
2. 女性実験者が操作するロボットと自由に対話（3 分）
3. 1 回目テーマトーク（3~5 分）
4. 2 回目テーマトーク（3~5 分）
5. 1 回目と 2 回目のロボットの動作についての比較アンケート記入

1 回目のテーマトークと 2 回目のテーマトークでの動作条件はカウンタバランスを取った．被験者は 1 回目と 2 回目の動作のうちどちらが対話に適しているかを 7 段階（1: 1 回目の対話の方，4: どちらとも言えない，7: 2 回目の対話の方）で評価した．

3.5 主観評価実験結果

実験の被験者は 15 人（男 : 7 人，女 : 8 人，平均年齢 : 21.7，標準偏差 : 2.1）であった．動作音がうるさかったなどと，動作の評価に影響しそうなことに言及した被験者（2 人），実験中ロボットが故障した際の被験者（1 人）は解析から除いた．実験後のインタビューとアンケートの結果を比較し，Noise 条件の方が VoiseNoise 条件よりも動きが多かったと答えた被験者（2 人）も解析から除いた（被験者が順序を混同した可能性があるため）．解析に用いた被験者は 10 人（男 : 5 人，女 : 5 人，平均年齢 : 21.9，標準偏差 : 1.7）であった．

図 6 に主観評価の結果のヒストグラムを示す．アンケート結果の平均値が 4 より大きければ大きいほど，Voice+Noise 条件の動作がより適していることを表し，4 より小さければ小さいほど，Noise 条件の動作がより適していることを表す．そこで，平均値が 4 よりも大きいかどうかの検定を行った．解析データには Shapiro-Wilk 検定より正規性が認められたため，t 検定を行い，有意差が認められた ($t(9) = 3.10, p < 0.01$)．平均値は 5.4 であり，提案システム (Voice+Noise) の動作がより対話に適切であることが示された．

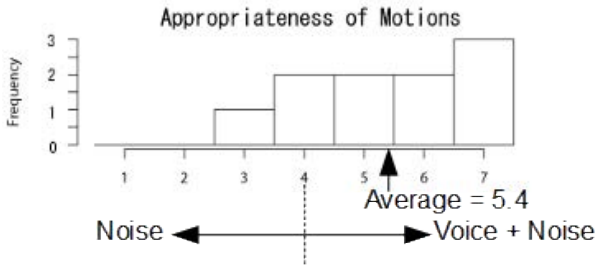


図 6: Subjective preference scores between the baseline motion (“Noise”) and the motion generated by the proposed system (“Voice+Noise”).

3.6 談話機能の誤認識分析

実験中の操作者の音声を録音し，Ishi et al.[5] の手法に従って別の実験協力者（1 人）が手動で談話機能タグを付けたものを正解データとした．この実験協力者はどの音声データがどちらの条件のものかは知らされていない．“bc” と “k” は同じ頷き動作を生成するため，“bc” と “k” の混同は対話者にとっての印象形成には影響しない．そのため，対話相手の印象形成への影響を考慮し，談話機能の推定の誤りではなく，生成された動作の誤りについて分析した．

表 1 に正しい談話機能を基にした動作生成タイミング (expected motions) において，システムが生成した動作 (generated motions) の個数を示す．頷き動作は 55%，首上げ動作は 50%で提案システムは正しく動作を生成でき

ていた．これは，頷き (bc) と驚き・感心 (su) についての感嘆詞とイントネーションを正しく検出できたためであると考えられる．しかし，首傾げ動作については 10%以下の正解率であった．これは，フィラー (f) や言いよどみの推定には音節の長さと音調のみ使用したため，音節の長さが短い場合の言いよどみが検出できなかったと考えられる．一方で，多くの頷きタイミングで首傾げ動作を生成していたり，どの動作も期待されないフレーズで頷き動作の生成も見られた．

表 1: Distributions of the generated motions for each of the expected motions (when the dialogue act tags are given).

	expected motion			
generated motion	nod	tilt	face up	insertions
nod	354	0	35	341
tilt	43	3	18	13
face up	113	0	83	36
no motion (deletions)	135	36	28	

4 考察

本実験により，提案システムによりリアルタイムで音声から動作が生成できることが確認できた．さらに，本システムにより動作を付加することで，誤り動作が含まれるにもかかわらず，より適切な印象を与えることが明らかになった．

提案手法は談話機能と頭部動作を一対一でマッピングしている．しかし，人間同士の対話では相槌の際に頷かない場合があるなど，提案システムのように談話機能と頭部動作が一対一に対応しているわけではない [11]．そのため，削除誤り（動作が生成されない場合）があっても評価が下がらなかったと考えられる．一方で，置換誤り（例：首傾げタイミングで頷いてしまう）や挿入誤り（例：フィラー以外で首をかしげる）は対話の意味と異なる意味を伝達するため，遠隔対話を行う上でより問題となる．しかし，頷き動作の意味は相槌のみならず，発話の強調や相手の頷きに対する反応や相手の応答を促すなど多岐にわたるため [17, 18]，頷き動作の挿入誤りや置換誤りが不自然と評価されなかったと考えられる．一方で，首傾げ動作は自然会話で生起するタイミングは頷き程多くはなく，その意味も頷き動作ほど多くはない．そのため，実験インタビューにおいても「頷き動作は変ではなくリアルで自然であった．首の動かし方は人それぞれなので不気味とかは感じなかったが,「首をこのタイミングで動かすか」と思うことはあった.」と答える被験者がいた．ただし，首傾げ動作とかわいらしい声という関係性から「声がかわいらしかったのでそれにマッチするかわいらしいしぐさ」と首傾げ動作を解釈する被験者もいた．前者の被験者は，誤認識動作を不適切と思っているが，後者の被験者はそれを操作者の癖と思っていると考えられる．そのため，誤認識があっても適切と判断される理由として，後者も一要因として考えられる．さらにこの結果から，ロボットの動作生成においては，音声との適合だけでなく，操作者のイメージとの適合も重要な要素であることが判った．

多少の誤り動作も許容されることが分かったが，置換誤り・挿入誤りは対話相手にとっては解釈できないものとなり得るため，将来研究としてそれら誤りを減らすよう制約を加えることが重要となる．また，ロボットの見た目によって求められる人間らしい動作の程度が変わると考えられるため（例：人らしいほど人間らしく動いてほしい），本論文で用いたテレノイド以外のロボットを用いた際の印象評価も行う．本実験では，操作者を 1 人に固定しているため，複数の操作者を用いた場合における提案システムの実用性の評価も重要である．さらに，頭部動作のみならず表情や視線といった他のモダリティの自動生成も行う．

5 まとめ

本論文では，円滑な対話を実現するために，遠隔操作ロボットの新たな動作自動生成手法を提案し，その有効性を心理実験により検証した．操作者音声の談話機能に適合した動作（頷きや首傾げなどの頭部動作）を自動生成したところ，操作者本人動作と同じ動作でないにも関わらず，対話において適切な動きと評価されることが判った．これはボディジェスチャや視線や表情といったモダリティを音声から生成できる一例となる．将来研究として，音声からそれら他のモダリティの生成が課題となる．

提案手法は操作者の音声情報のみ用いるため，操作者は携帯電話などの小型通信機器での遠隔操作も可能である．また，四肢が不自由なため身体動作による意思疎通が困難な人たちも，音声のみでロボットアバタに動作を表現させることで，円滑なコミュニケーションが可能である．このように，提案手法は新たな遠隔コミュニケーションサービスに役立つはずである．

参考文献

[1] Mehmet Emre Sargin, Oya Aran, Alexey Karpov, Ferda Ofli, Yelena Yasinnik, Stephen Wilson, Engin Erzin, Yücel Yemez, and Ahmet Murat Tekalp. Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis. *International Conference on Multimedia and Expo*, 2006.

[2] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. Rigid

Head Motion in Expressive Speech Animation: Analysis and Synthesis. *Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 3, 2007.

[3] Mary Ellen Foster and Jon Oberlander. Corpus-based generation of head and eyebrow motion for an embodied conversational agent. *Language Resources and Evaluation*, Vol. 41, No. 3-4, pp. 305–323, 2007.

[4] Carlos Toshinori Ishi, Judith Haas, Freerk P. Wilbers, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of head motions and speech, and head motion control in an android. In *IROS2007*, pp. 548–553, 2007.

[5] Carlos Toshinori Ishi, ChaoRan Liu ChaoRan Liu, H Ishiguro, and N Hagita. Head motion during dialogue speech and nod timing control in humanoid robots. In *HumanRobot Interaction*, pp. 293–300, 2010.

[6] Chaoran Liu, Carlos Toshinori Ishi, H Ishiguro, and N Hagita. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *HumanRobot Interaction*, pp. 285–292, 2012.

[7] Binh Huy Le, Xiaohan Ma, and Zhigang Deng. Live Speech Driven Head-and-Eye Motion Generators. *Transactions on Visualization and Computer Graphics*, Vol. 18, No. 11, pp. 1902–1914, 2012.

[8] Tomio Watanabe, Masashi Okubo, Mutsuhiro Nakashige, and Ryusei Danbara. InterActor: Speech-Driven Embodied Interactive Actor. *International Journal of Human-Computer Interaction*, Vol. 17, No. 1, pp. 43–60, 2004.

[9] Hiroki Ogawa and Tomio Watanabe. InterRobot: a speech driven embodied interaction robot. *RO-MAN2000*, pp. 322–327, 2000.

[10] Kevin G. Munhall, Jeffery A. Jones, Daniel E. Callan, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological science*, Vol. 15, No. 2, pp. 133–137, 2004.

[11] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts. *INTERSPEECH*, 2006.

[12] Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication*, Vol. 50, No. 6, pp. 531–543, 2008.

[13] Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano. Julius — an Open Source Real-Time Large Vocabulary Recognition Engine. In *EUROSPEECH*, pp. 1691–1694. ISCA, 2001.

[14] Carlos Toshinori Ishi. Perceptually-Related F0 Parameters for Automatic Classification of Phrase Final Tones. *IEICE transactions on information and systems*, Vol. 88, No. 3, pp. 481–488, March 2005.

[15] Kaiko Kuwamura, Takashi Minato, Shuichi Nishio, and Hiroshi Ishiguro. Personality distortion in communication through teleoperated robots. In *RO-MAN2012*, pp. 49–54, 2012.

[16] Carlos Toshinori Ishi, Chaoran Liu, Hiroshi Ishiguro, Norihiro Hagita, Intelligent Robotics, and Communication Labs. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. *IROS2012*, pp. 2377 – 2382, 2012.

[17] Joseph D. Matarazzo, Arthur N. Wiens, George Saslow, Bernadene V. Allen, and Morris Weitman. Interviewer mm-hmm and interviewee speech durations. *Psychotherapy: Theory, Research & Practice*, Vol. 1, No. 3, pp. 109–114, 1964.

[18] 庵原彩子, 堀内靖雄, 西田正史, 市川熹. 自然対話におけるうなずきの機能に関する考察 (分析、生成と評価)(音声とコミュニケーション及び一般). 電子情報通信学会技術研究報告. HCS, ヒューマンコミュニケーション基礎, Vol. 104, No. 445, pp. 13–18, 2004.