Vol.2018-SLP-123 No.5
2018/7/27

会話におけるニュース記事伝達のための発話意図の分類と認識

横山 勝矢[1]　高津 弘明[1]　本田 裕[2]　藤江 真也[1,3]　小林 哲則[1]

概要：人同士の会話において，聞き手は反射的な短い応答によるフィードバックで話し手に明示的/暗示的に様々な意図を伝達することが多い．話し手はこれらの意図を認識し，コミュニケーションをより効率的にするために発話計画を変更する．これらの機能は人ーシステム間の会話にも役立つことが期待されるが，人が実際のシステムと対面したときに同様のフィードバックを常に返すとは限らない．そこで，我々の設計したシステムに対するユーザのフィードバック現象を調査し，それらの意図を認識する枠組みを提案した．まず，システムの動作に影響を与えるフィードバックの意図を，ユーザの関心の有無や理解状態，話者交代などの観点から分類した．次に，実システムを用いてユーザの短い応答を収集し，第三者の聞き取りによる意図のラベル付けを行った．さらに，収集したデータをもとに意図の認識モデルの学習を行った．結果として，現状のシステムに対しては質問を意図したフィードバックが多いことが分かり，またその認識に関して提案モデルの有効性が確認できた．

キーワード：音声対話システム，発話意図，音声

Utterance Intention Classification and Recognition for Convasational News Delivery System

Katsuya Yokoyama[1]　Hiroaki Takatsu[1]　Hiroshi Honda[2]　Shinya Fujie[1,3]
Tetsunori Kobayashi[1]

1. はじめに

実会話システム利用時における，システム発話に対するユーザのフィードバックの現象を分析するためのコーパスについて検討・収集し，収集したコーパスを用いて，発話意図理解のための深層学習に基づく韻律特徴量の自動抽出を試みた．

人同士の会話において，聞き手は，話し手の発話に対し，しばしば反射的な応答を行う．これらの応答は，発話の理解状態や発話内容への興味の程度など，聞き手の様々な状態を，話し手にフィードバックする．話し手は，これらのフィードバックをもとに，相手の理解が乏しければ情報を補足し，興味がなさそうなら話題を変えるなど，自らの発話計画を適宜修正し，聞き手の状態に合わせた発話を行うことになる．

こうした一連の処理を，人とシステムとの会話にも適用することで，従来の一問一答の枠組みを超えた，円滑で利便性の高い会話システムを実現できることが期待される [1] [2]．我々はニュース記事のようなまとまった量の情報を音声対話によって効率的に伝達するシステムの開発を行っている [9] が，ここでも，ユーザからのフィードバックを常に観測し，これに応じて伝達する情報を制御することが必要になる。

ここで，聞き手からのフィードバックは，必ずしも言語的に明示的な形でのみ表現されるとは限らず，抑揚や声色などに表れるパラ言語を伴った複雑な形で表現される [3]．このため，こうした機能の実現には，自然なフィードバッ

[1] 早稲田大学，Waseda University
[2] 本田技術研究所，Honda R&D Co.,Ltd
[3] 千葉工業大学，Chiba Institute of Technology

クを含む大規模な会話音声コーパスが必要となる．

そこで，本研究ではシステムにとって有用なユーザのフィードバック現象の分類について検討を行い，システムとの会話音声コーパスを収集した．さらに，韻律情報を用いて逐次的に認識する手法を提案し，収集したデータを用いて認識実験を行った．

以下本稿では，2 で関連研究について述べる．3 でユーザが会話時に行うフィードバックの分類について説明し，4 でデータセットの収集方法やアノテーションについてとその結果について述べる．5 で提案する発話意図認識モデルについて説明し，6 で収集したデータを用いた認識実験とその結果についてを述べる．

2. 関連研究

2.1 コーパス収集

会話における聞き手からのフィードバックを扱うためのコーパスの研究には，既に様々なものがある．牧野らは，人―人の会話における話者交代の現象を分析対象として，話者交代の際に起こる韻律情報と視線情報を調査するためのコーパスを作成した [4]．荒木らは，雑談における会話内容に対するユーザの興味の有無を分析対象として [5]，MMDAgent を用い，WoZ（Wizard of Oz）方式により対話データの収集を行った．井上らは，エンゲージメント（興味や意欲による対話継続欲求度）を分析対象として，自立型アンドロイド ERICA[7] を用い，WoZ 方式により，会話時の相槌，笑い，うなずき，視線などのデータを収集した [6]．

これら，人―人の会話や，WoZ 方式のように，応答時に人が介在して収集したコーパスには，極めて自然なフィードバックが含まれ，研究対象としては興味深い．しかし，現状，実際の会話システムを使用したときのユーザの振る舞いとは乖離している可能性があり，システムへ組み込む機能を実現するための学習データとしては必ずしも適切ではない．

このような観点から，実際の自律システムとの対話を通じてデータ収録した研究もある．長澤らは，インタビューを行うシステムにおいて，ユーザの反応によって話題を掘り下げるか，話題を転換するかなどの対話戦略を決めるために，実際にロボットを用いた対話を行ってデータを収録し，ユーザの発話への積極性を 5 段階でアノテーションしている [8]．

本研究においても，長澤らの研究と同様に，実システムの利用時におけるユーザの応答を調査対象とすることを試みる．現在我々が作成しているニュース伝達のための会話システム [9] を対象として，ユーザが実際にシステムとの会話時にどのようなフィードバックを行い，システムはどのようにそれを利用しうるかを検討する．システムからの提示情報に対するユーザの興味の有無，発話交代に関する働きかけ（発話権を取りたいか，継続したいかなど），システム発話の理解状態など，ユーザがフィードバックの発話に込めた，システムの振る舞いに影響を与えうる情報を広くとりあげ，これらの機能を整理・分類するとともに，これらが，現状のシステムとの会話におけるユーザの応答にどの程度含まれるかを調査する．

2.2 フィードバックの認識

従来，発話意図の認識・理解は韻律情報を用いて行なわれてきた．例えば，藤江らは，システムに対する利用者の発話態度（肯定的か否定的か）を推定するために，第 1 モーラの基本周波数（F0）の傾き，発話全体の F0 レンジ，最終モーラの継続長からなる 3 次元の特徴量を使用している [11]．Ando らは，対話データにおいてその発話が肯定的であるか否定的であるかを推定するために，音響特徴量として単語ごとに算出した F0，パワーの最大・最小，継続長，間などの情報を，言語特徴量としてバイグラム言語モデルのパープレキシティを使用している [12]．Nishimura らは，収集した子供の音声に対して喜んでいるか嫌悪を示しているかの推定を行う際，F0 値，パワーから求められる 16 次元の特徴量の中から，因子分析により寄与の高い因子を選択することで，高い識別率を達成している [13]．林らは，発話タグ（疑問文，平叙文，相槌，同意，笑い）の推定を行う際，F0 値やパワー，話速などの 23 次元の音響特徴量を使用している [14]．

これらの研究は，韻律情報を詳細に観察・分析することで意図理解のための特徴パラメータを設計している．しかしながら，人手で設計可能な特徴パラメータは，推定対象の発話意図の違いが文章や音声を観察することで見いだせる場合に限られる．つまり，観察・分析だけでは容易に特徴が見つけられないような微妙なニュアンスの意図については特徴量の設計が困難になる．加えて，発話意図ごとに特徴パラメータの設計が必要となる．

本研究では，発話意図理解の韻律特徴量をニューラルネットワークを用いて自動で抽出する枠組みを提案する．具体的には，音響情報としてスペクトログラムを CNN を含む AutoEncoder に入力し，その中間層出力を韻律特徴量として用いる．これにより，発話意図の種類に依らず同一の仕組みで韻律特徴量の獲得が可能となる．

3. 発話意図分類

システム会話におけるユーザのフィードバック現象を分析するコーパスの収集に先立ち，フィードバックがシステムの振る舞いに影響を与えうる情報を広くとりあげ，その機能を整理・分類した．

フィードバックの役割は，システムにある種の要求を伝えることにあるが，これらは，ユーザの要求を直接明示的に伝えるものと，ユーザ状態等を伝達することで間接的，

あるいは暗示的に伝えるものがある．本稿ではこれらをまとめて発話意図（intention）と呼ぶ．

情報伝達効率の高い会話を実現するための機能として，システムが伝える情報量の観点から，その増減を要求するフィードバックがあると考えた．また，円滑な会話を実現するための機能として，人の発話とシステムの発話が同時に起きるべきではないと考えた（発話の衝突回避）．そこで，本研究では，フィードバックの役割を 3 つに分類し，それぞれの役割を持つ 6 つの発話意図を提案する．

伝達情報増加を求める役割
・質問
・補足要求
・反復要求
伝達情報減少を求める役割
・無関心
・既知
発話の衝突を避ける役割
・待機要求

3.1 伝達情報増加を要求

「質問」「補足要求」「反復要求」は伝達する情報を増やすことを要求する意図である．まとまった量の情報を伝達するシステムにおいてユーザの理解状態に応じて伝える情報の量を調整する必要があるが，中でもユーザの疑問を解消しながら必要十分な情報を伝達する機能は不可欠である．これらの意図を認識できれば, 伝達する情報を増やす調整を行うことができる．「質問」はユーザが疑問をもち陽に質問を行ったもので，このフィードバックを受けたシステムは質問の内容に沿った受け答えをすることが考えられる．（例：U「サービスの開始はいつだっけ？」，S「5 月 28 日だよ」）「補足要求」は陽に質問は行っていないが，システム発話に疑問点を感じているもしくはシステム発話に理解が追いついていないため補足説明がほしいという意図である．このフィードバックに対しては，直前の発話に関する補足説明をすることが考えられる．（例：S「平成 25 年から ICT 成長戦略会議を行っているみたい」，U「ふーん？」，S「ICT 成長戦略っていうのはね・・・」）「反復要求」はシステム発話を聞き取れなかった場合や受け取れなかった場合にもう一度同じ発話を繰り返してほしいという意図である．このフィードバックに対しては，該当発話を繰り返すことが考えられる．（例：U「ん？もう一回言って？」，S「(直前発話を繰り返す)」）

3.2 伝達情報減少を要求

「無関心」「既知」は，伝達する情報を減らすことを要求する意図である．興味のないトピックや既に詳細を知ってる情報を長々と伝えるシステムは利便性のあるものではない．そういった情報をユーザの反応に応じて省くことで利便性があり，継続して利用したくなるようなシステムになると考えられる．「無関心」は対話トピックに対してユーザが興味を持っていないという意図である．ユーザの興味の無い情報を伝達することは無駄だと考え，この意図を認識した場合は他のトピックへ遷移することが考えられる．「既知」はシステムが伝える情報がすでにユーザが有する情報であることを示す意図である．既知情報を含むトピックに関しては詳細な情報を省き伝達することで，情報伝達効率を上げることができると考えられる．

3.3 発話衝突を避けることを要求

「待機要求」はユーザがある発話の後少し間を空けても話し続けることを示し，システムの発話は待ってほしいという意図である．従来の音声対話システムでは VAD(Voice Activity Detection) により発話の区切りが検出されたら次の発話を開始するというものであった．そのためユーザはまだ発話を継続するつもりにも関わらず，システムが発話を開始し発話の衝突が起こっていた．この「待機要求」を認識することでシステムは発話を待機し，衝突を陽に回避することができると考えられる．

4. データセット構築

3 で提案した発話意図に関するラベルを付与したデータセットの構築について述べる．発話意図のあらわれ方は，人同士で対話した場合と，システムと対話した場合とでは異なることが考えられる．本研究ではシステムと対話した際の現象をとらえるため，実際に [9] のシステムを用い，24 人の被験者が対話を行った際の音声データを収集した．さらに．収集したデータを 10 人のアノテータにより 3 で提案した発話意図に関するアノテーションを行った．

4.1 音声収録方法

人がシステムと対話した際の現象をとらえるため，実際に [9] のシステムを用い，被験者が対話を行った際の音声データを収集した．現在 2060 対話の収録が完了している．音声データ収集には 24 人（男：12 人，女：12 人）の被験者がシステムとの対話を行い，その際の音声データを随時収録した．

4.2 アノテーション

アノテーションは，図 1 に示すアノテーションツールを用い，10 人のアノテータで行った．このツールで用いるラベルは 3 で提案した発話意図を示すことができるかどうかと，ラベルのつけやすさを考慮し設計した．具体的には，ラベルを現象，行為，内部状態の 3 つのレイヤに分け内部状態には 5 つの軸を設定した．

現象では，「発話の完了」を示す終了，「発話が途切れていたり，遮られている」ことを示す継続，「音声のない区

図 1　作成したアノテーションツール

表 1　各ラベルに対応する発話意図とそれに対して想定されるシステムアクション

ラベル		発話意図	システムアクション
質問		質問	質問応答
意外		補足要求	補足説明
相槌	＆懐疑		
or	＆不受理	反復要求	繰り返し
リアクション	＆無関心	無関心	他トピックへ移行
当然		既知	詳細説明省略
継続		待機要求	傾聴
ターンをとる			

間」を示す無音・雑音，「システム対話と関係のない音声」を示す無効，の 4 ラベルから選択する．

行為では，発話行為として分類される相槌，ターンをとる，返答，質問，情報開示，リアクションの 6 ラベルについて当てはまるものを選択する．

内部状態では「システム発話を受け取れているかどうか」「システム発話の内容に疑問を感じているかどうか」「対話内容に興味を抱いているかどうか」「システム発話の内容に好意的かどうか」「システム発話の内容が予期しえたかどうか」の 5 つの軸についてそれぞれ 3 段階で評価を行う．

各レイヤでの選択は独立に行われ，内部状態のレイヤでは軸ごとにも独立に評価が行われる．各ラベルに対応するユーザ意図とそれに対して想定されるシステムアクションを表 1 に示す．

対象音声の分割は WebRTC vad (aggressiveness=2)[*1] を用い，反射的な反応を対象とするため 1.5 秒以下の音声を対象とした．また [10] において前の文脈が対象発話の評価に影響を及ぼすことが示されている事を考慮し，評価の際の材料として「対象音声の 3 秒前から再生」「対象音声の 5 秒前から再生」を作成した．さらに 5 秒前の音声からのみ

[*1] https://github.com/wiseman/py-webrtcvad

表 2　各ラベルの数

ラベル名		数	ラベル名		数
現象	終了	11671	内部状態	受理	11476
	継続	445		不受理	1158
	無音・雑音	1286		納得	7279
	無効	379		懐疑	3571
行為	相槌	5753		興味	6300
	ターンをとる	191		無関心	2644
	返答	429		当然	509
	質問	2933		意外	3819
	情報開示	811		好感	1566
	リアクション	7977		嫌悪	1498

では評価できなかった場合を考慮して，「対話全体」を聞けるようにし，対象発話が全体の中で何秒から始まるかの記述も行った．

他にも，アノテータ間の意見の相違をなくすためアノテーションの試用期間を設け，不一致が多かったものに対する意見のすり合わせ等も行った．

アノテーションの手順としては，初めに「ラベルをつける発話」を再生し，対象発話がどのようなものか聞く．次に，「3 秒前から再生」，「5 秒前から再生」を聞き，対象発話の前の文脈がどのようなものであったかを確認する．この時，前の文脈を確認できなかった場合には「対話全体」から対象発話の前の文脈を聞くこととする．その後，「現象」「行為」「内部状態」の順にラベルを付与する．「現象」のレイヤにおいて「終了」を選択した場合には必ず「行為」「内部状態」の各レイヤについてもラベルを付与することとし，「終了」以外を選択した場合には「行為」「内部状態」のラベルを振る必要はないものとした．

アノテーションを行う際には，アノテータ同士の一致度を求めるため．200 発話分に対しては全員がラベルを付与し，その一致度を求めた．一致度の計算にはフライスのカッパ係数（Fleiss' Kappa）を用い，「現象」「行為」ではそれぞれどのラベルを選択したか，「内部状態」では各軸毎にどのラベルを選択したかで一致度を求めた．

4.3　結果と考察

現在，15604 発話に対するラベル付与が済んでいる．集まったデータについて各ラベルの数は表 2 の通りである．この数とは，アノテータが重複している場合は一人でも付与しているラベルがあればカウントしたものである．

この中から表 1 で示した発話意図の分類に当てはまるラベルの数は表 3 の通りである．

また，アノテータ全員がラベルを付与した 200 発話に対するラベルの一致度 (Fleiss' Kappa) を図 2 に，提案した発話意図毎の一致度を図 3 に示す．3 で提案した発話意図において，最も表出頻度が高かったものは「補足要求」で 4620 発話，次いで「質問」が 2933 発話，「無関心」が 2542

表 3　各発話意図に対応するラベルとそのデータ数

ラベル		発話意図	データ数
質問		質問	2933
意外		補足要求	4620
相槌	＆懐疑		
or	＆不受理	反復要求	515
リアクション	＆無関心	無関心	2542
当然		既知	509
継続		待機要求	517
ターンをとる			

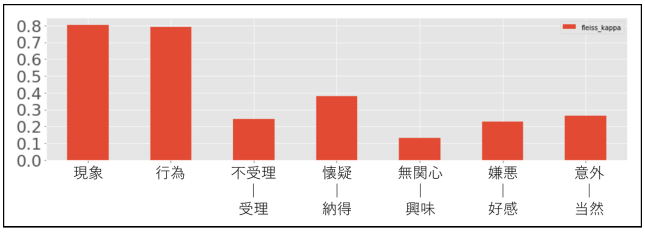

図 2　アノテータ同士のラベル一致度（Fleiss' Kappa）

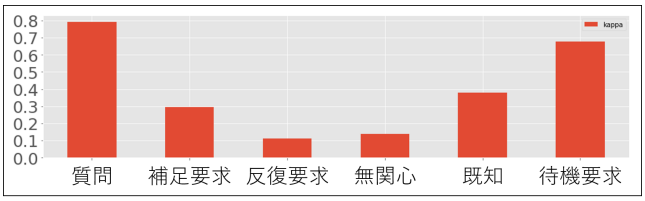

図 3　発話意図毎の一致度（Fleiss' Kappa）

発話であり，他の意図に関しては約 500 発話と表出頻度が少なかった．また，ラベル一致度についてみると現象や行為のラベルについては κ 係数が約 0.8 とほぼ完全な一致に近い値であったが，内部状態のラベルについてはどれも 0.4 を下回り，アノテータ間における評価の一致度は低かった．発話意図毎の一致度についてみると，「質問」については κ 係数が約 0.8 とほぼ完全な一致を，「待機要求」については κ 係数が 0.6 を上回り実質的な一致を示していたが，その他の意図については 0.4 を下回り，アノテータ間の評価一致度は低かった．

内部状態のラベル数については，「システム発話を受け取れているかどうか」の軸に関するラベルの付いた発話数が 6463，「システム発話の内容に疑問を感じているかどうか」に関するラベルの付いた発話数は 5289 と「終了」のラベルが付いた発話の半数以上にラベルがついていたが，「対話内容に興味を抱いているかどうか」に関するラベルが付いた発話数は 3309，「システム発話の内容に好意的かどうか」に関するラベルが付いた発話数が 892，「システム発話の内容が予期しえたかどうか」に関するラベルが付いた発話数が 332 と「システム発話を受け取れているかどうか」「システム発話の内容に疑問を感じているかどうか」の軸以外は内部状態のラベルの付いた発話が少ないことがわかる．このことから，現状のシステムとの対話において話者はそれほど内部状態の表出を行っていないことが分かる．

使用したシステムのニュースを伝えるという特性上，「質問」「補足説明」といった発話意図の表出が増えることは容易に想像でき，それは表出頻度にも表れている．中でも「質問」に関しては評価の一致度も高いため，データセットとしての質は高く，システムでも高精度に認識できることが期待できる．このため，システム実運用においても有用な意図として利用可能と考えられる．しかし，「補足要求」については一致度が低く，システム実運用の際に判断のブレが強く出てくることが考えられる．「無関心」については，話者の内部状態表出が少ないことから，棒読みに聞こえることが多く，第三者からすると無関心に聞こえることが多かったものと考えられる．また，ラベルの一致度については「補足説明」と同様のことが言える．「待機要求」については，ラベラー同士の一致度は高く，有用なデータであるように見えるが，そのデータ数は少ない．他の意図に関しては表出頻度が低く，構成するラベルの一致度も低い．

以降，発話意図の認識実験においては，データ数が多く，評価の一致度も高い「質問」と，データ数の多い「無関心」，「補足要求」を対象として，深層学習に基づく韻律特徴量の自動抽出について述べる．

5.　発話意図認識モデル

時間方向に可変な音声の入力を，短い時間幅で逐次処理し，その都度過去の履歴を考慮して現時点で予測される発話意図ラベル（e.g. 質問，待機要求，反復要求）を出力するモデルを構築した．まず，短い時間幅で切り出した音声の断片からスペクトログラムを生成する．次に，得られたスペクトログラムを CNN を含む AutoEncoder（CNN-AutoEncder）に入力し，その中間層ベクトルを圧縮された韻律特徴量として時系列に沿って LSTM に入力する．LSTM は逐次発話意図ラベルを出力するが，音声認識結果が得られた段階で，LSTM の最終層と音声認識結果から得られる発話の言語特徴量を識別器に与え，最終的な発話意図ラベルの推定結果を得る．モデルの全体像を図 4 に示す．以下，特徴抽出部と識別部について説明する．

5.1　特徴抽出部の設計

一般に発話意図の推定では，特徴量として基本周波数（F0）が用いられる．しかしながら，音声波形の準周期性や周辺雑音，有声音中の基本周波数の変化が広域に渡るなどの理由により，基本周波数を正確に抽出するのは難しい．そこで，F0 の推定を介さずに音声の時間・周波数スペクトルから直接特徴量を抽出する方法を提案する．

音韻や声の高さに関する特徴はスペクトログラムの模様として表れる．そこで，このスペクトログラムの模様を二次元の画像と見なし，5 層の畳み込み層と 3 層の全結合層からなるネットワークを折り返した全 16 層で構成される

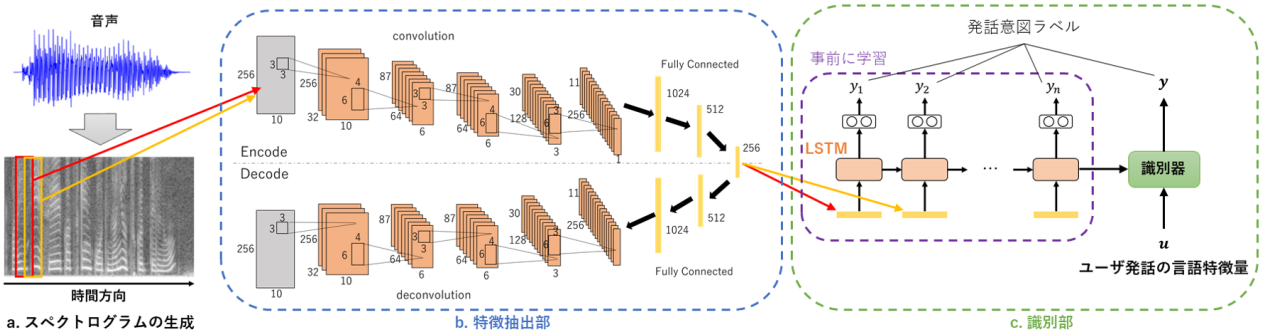

図 4　発話意図認識モデル

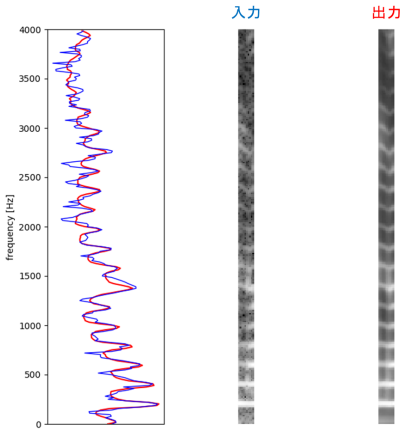

図 5　CNN-AutoEncoder の出力例（左:各スペクトログラムの 5 番目のスペクトルの周波数強度（青：入力，赤：出力），中：入力したスペクトログラム，右：出力したスペクトログラム）

CNN-AutoEncoder（図 4.b）を学習する．このとき, 中間層は韻律をコンパクトに表現する特徴量と考えることができる. そこで，この中間層を発話意図の識別に用いた．

5.2　識別部の設計

音声は時間方向に可変長であり，時間方向の長さに頑健なモデルであることが望まれる．同じ文字列の音声でも人の違いや発話する条件や状態によってその継続長は異なる．また，発話末のピッチが上昇すると「質問」と捉えやすくなるなど，発話意図認識において韻律の時間方向の変化は有用な情報である．

そこで，本研究では，RNN の中でも長期のデータ列に係る特徴抽出能力に長けた LSTM を用いた．また，マルチラベルの学習ではなく，各発話意図毎にモデルを学習し，それぞれが該当する意図か否かを出力するものとした（図 4.c）．

さらに，音声認識結果が得られた段階で，LSTM の最終層と音声認識結果から得られる発話の言語特徴量を併用して再度識別することも試みた（図 4.c）．

表 4　実験で用いたデータセットの統計

	「質問」		「無関心」		「補足要求」	
	正例	負例	正例	負例	正例	負例
訓練セット	2042	2042	1776	1763	3228	3197
テストセット	584	584	508	504	924	915
開発セット	292	276	254	252	461	457
総数	5827		5057		9182	

6.　発話意図識別実験

6.1　特徴抽出部の学習と結果

特徴抽出部における CNN-AutoEncder の学習には大規模な『日本語話し言葉コーパス（CSJ）』[*2] を用いた．

CNN-AutoEncoder の入力は，フレームサイズ 800（50ms），フレームシフト 160（10ms），チャンクサイズ 1024 で切り出した音声から生成したスペクトログラムを時系列に並べたものとし，そのサイズは 10×256 とした．このデータをもとにネットワークの学習を行い，特徴抽出器を構築した．

学習した CNN-AutoEncoder の出力例を図 5 に示す．ここで, 中央のスペクトログラムが入力として与えたスペクトログラムで, 右側のスペクトログラムが CNN-AutoEncoder が出力したスペクトログラムである．左側のグラフは各スペクトログラムの 5 番目のスペクトルの周波数強度を表している．これらの結果からモデルが入力スペクトログラムの模様パターンを復元できていることが分かった．

6.2　識別部の学習と結果

4 で作成したデータセットのうち，データ数の多い発話意図「質問」,「無関心」,「補足要求」に関して識別実験を行った．本実験で用いたデータセットの統計を表 4 に示す．

6.2.1　LSTM の学習と識別結果

まず，6.1 節で学習した CNN-AutoEncoder の中間層の値（256 次元）を入力として LSTM の学習を行った．ここで，CNN-AutoEncoder への入力には，100ms のスペクトログラムを 50ms ずつシフトさせながら与えた．つまり，

*2　http://pj.ninjal.ac.jp/corpus_center/csj/

表 5　各発話意図毎の識別結果

	Accuracy	正例→正例	負例→負例
質問（韻律のみ）	0.903	0.928	0.878
質問（韻律＋言語）	0.965	0.957	0.973
無関心（韻律のみ）	0.763	0.756	0.770
無関心（韻律＋言語）	0.765	0.783	0.746
補足要求（韻律のみ）	0.650	0.955	0.342
補足要求（韻律＋言語）	0.758	0.831	0.684

現時刻の入力と次の時刻の入力で 50ms のオーバーラップが存在する．これは，LSTM に時間変化の情報を陽に学習させるためである．モデルのパラメータは,LSTM のユニット数を 256, 出力層のユニット数を 2 に設定した.

LSTM の最終層の出力ラベルと正解ラベルの一致率を求めたところ，発話意図「質問」に関する Accuracy は 0.903，正例を正例と識別する精度は 0.928，負例を負例と識別する精度は 0.878 であった．また,「無関心」に関する Accuracy は 0.763，正例を正例と識別する精度は 0.756，負例を負例と識別する精度は 0.770.「補足要求」に関する Accuracy は 0.650, 正例を正例と識別する精度は 0.955, 負例を負例とする精度は 0.342 であった．「質問」についてエラー分析を行ったところ，「何を」や「なにそれ」,「なんでその日なの」等，言語表記を見れば「質問」と判別できるような発話を負例と判定してしまう例が多くあった．このような誤りを解消すべく，次節では発話の言語情報も組み合わせた発話意図の識別を行った．

6.2.2　韻律情報と言語情報を用いた識別器の学習と識別結果

6.2.1 節において学習した識別器では，韻律特徴量しか用いていないため，言語表記を見れば誤った判定をしないような例についても誤った判定がなされていた．そこで，LSTM の最終層と言語特徴量を用いて再度識別を行うことで，発話内容にも頑健な認識モデルを構築した．用いる言語特徴量としては，ユーザー発話の文字または単語の Bag-of-Words（BoW）と，JUMAN（Ver.7.01）[*3]を適用して得られる形態素情報のうち形容詞を使用した．識別器には scikit-learn[*4] における線形カーネルの SCM を用いた．その結果を表 5 に示す.

7.　おわりに

実会話システム利用時における，システム発話に対するユーザのフィードバック現象に関するコーパス収集を行い，収集したデータを用いて認識実験を行った.

現在我々が作成しているニュース伝達のための会話システム [9] を用いるうえで，ユーザがシステムとの会話時にどのようなフィードバックを行い，システムはそれをどのように利用しうるかを調査した結果，現状のシステムでは発話意図「質問」「無関心」「補足要求」においては出現頻度は高く，更に「質問」についてはアノテータ間の一致度も高いことが分かった.

一方で，その他の意図は表出頻度が低く, かつ，その程度も弱いためアノテータ間の一致度は低かった．これは，現状のシステムでは，聞き手からのフィードバックが少ないことを意味する．提案する会話システムは，聞き手からのフィードバックをもとに情報伝達の効率化を目指すものであるから，このような状態だとそもそもシステムの特徴を活かすことができない．このため，内部状態の自然な表出をユーザに促すシステムであることが求められる．このとき, 何を改善すべきか，どのような機能を加えるべきか，それとも，そもそもシステムとの会話では人は内部状態は表出しないのか，などについて検討が必要である.

また，分類・整理した発話意図について，韻律情報を元に逐次的に認識する手法を提案した．提案手法の特徴は，従来意図の種類ごとに観察に基づいて行われてきた特徴パラメータの設計を深層学習を用いることで，寄与率の高い韻律特徴量を自動で抽出しているところにあり，その識別精度は「質問」において 96.5%,「無関心」において 76.5%,「補足要求」においては 75.8%の識別率を達成した．今後は，他の発話意図に関してもデータを増やし，識別実験を行うとともに, 感情認識等のタスクにおける認識手法との比較実験を行う.

[*3] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[*4] http://scikit-learn.org/stable/

参考文献

[1] Shinya Fujie, Riho Miyake, Tetsunori Kobayashi: Spoken Dialogue System Using Recognition of User's Feedback for Rhythmic Dialogue, in *Proceedings of Speech Prosody (2006)*
[2] Tetsunori Kobayashi, Shinya Fujie: Conversational Robots: An Approach to conversation protocol issues that utilizes the paralinguistic information available in a robot-human setting, in *Acoustical Science and Technology, Vol. 34,* pp64-72 (2013)
[3] 益永祐吾, 川端豪: 発話意図判定における文型・韻律・音色の相補的効果, 情報処理学会研究報告音声言語情報処理（SLP）, Vol.2010-SLP-80, No.14 (2014)
[4] 牧野孝史, 三浦寛也, 竹川佳成, 平田圭二: 複数人対話における話者交替現象分析のためのコーパスの作成, 人工知能学会言語・音声理解と対話処理研究会（SIG-SLUD）, pp.3-8 (2017)
[5] 荒木雅弘, 冨増紗也華, 中野幹生, 駒谷和範, 岡田将吾, 藤江真也, 杉山弘晃: マルチモーダル対話データの収集と興味判定アノテーションの分析, 人工知能学会言語・音声理解と対話処理研究会（SIG-SLUD）, pp.20-25 (2017)
[6] 井上昂治, Divesh Lala Pierrick Milhorat, 高梨克也, 河原達也: 潜在キャラクタモデルによるリアルタイム対話エンゲージメント推定, 人工知能学会言語・音声理解と対話処理研究会（S IG-SLUD）, pp.78-83 (2017)
[7] K. Inoue, P. Milhorat, T. Zhao, T. Kawahara: Talking with ERICA, an autonomous android, in Processings of SIGDIAL Conference, pp.212-215 (2016)
[8] 長澤史記, 石原卓弥, 岡田将吾, 新田克己: ユーザーの態度推定に基づき適応的なインタビューを行うロボット対話

システムの開発, 人工知能学会言語・音声理解と対話処理研究会, pp.72-77 (2017)
[9] 高津弘明, 福岡維新, 藤江真也, 林良彦, 小林哲則: 意図性の異なる多様な情報行動を可能とする音声対話システム, 人工知能学会論文誌, Vol.33, No.1 (2018)
[10] 森大毅, 相澤宏, 粕谷英樹: 対話音声のパラ言語情報ラベリングの安定性, 日本音響学会誌 61 巻 12 号, pp.690-697 (2005)
[11] 藤江真也, 江尻康, 菊池英明, 小林哲則: パラ言語の理解能力を有する対話ロボット, 情報処理学会研究報告音声言語情報処理（SLP）, Vol. 2003, No. 104(2003-SLP-048), pp. 13-20 (2003)
[12] Ando, A., Asami, T., Okamoto, M., Masataki, H., and Sakauchi, S.: Agreement and disagreement utterance detection in conversational speech by extracting and integrating local features, in *Proceedings of the 16th Annual Conference of the International Speech Communication Association*, pp. 2494-2498 (2015)
[13] Nisimura, R., Omae, S., Kawahawa, H., and Irino, T.: Analyzing dialogue data for real-world emotional speech classification, in *Proceedings of the 7th Annual Conference of the International Speech Communication Association*, pp. 1822-1825 (2006)
[14] 林佑樹, 大佛駿介, 中野有紀子: 協調学習における韻律特徴を用いた発話タグ推定モデル, 教育システム情報学会第 39 回全国大会, pp. 441-442 (2014)
[15] 田中真詞 , 川端豪: 文型と音調によるユーザの発話意図の推定, 情報処理学会研究報告音声言語情報処理（SLP）, Vol. 1998, No.68(1998-SLP-022), pp. 55-60 (1998)
[16] 岩田和彦, 小林哲則: 終助詞とその音調とによって聞き手に伝わる発話意図の分析, 電子情報通信学会技術研究報告, Vol. 112, No. 281, pp. 31-36 (2012)
[17] Schuller, B., Rigoll, G., and Lang, M.: Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 577-580 (2004)
[18] Han, K., Yu, C., and Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine, in *Proceedings of the 15th Annual Conference of the International Speech Communication Association*, pp. 223-227 (2014)
[19] Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks, in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097-1105 (2012)