＜ 修 士 論 文 ＞

コールセンターの応対音声品質の自動評価に向けた要素技術の研究

滋 賀 大 学 大 学 院
デ ー タ サ イ エ ン ス 研 究 科
デ ー タ サ イ エ ン ス 専 攻

修了年度：２０２０年度
学籍番号：6019111
氏　　名：高山　和明
指導教員：市川　治
提出年月日：２０２１年１月２０日

目次

第１章 はじめに

1.1 研究の背景

現代社会においてコールセンターの重要性は以前にも増して高まっている．大量生産・大量消費の時代が終わり，消費者ニーズの多様化と共に流行し定着した「カスタマーエクスペリエンス（顧客体験価値）の向上」というキーワードに象徴されるように，消費者は商品そのものの価値だけでなく購買体験や購入後のサポートも重視するようになっている．また公共サービスの電話応対業務の一部を民間に委託するケースも増えている．このような時代の中で，コールセンターは企業・団体と顧客・利用者との接点の最前線に位置しており，その需要は増え続けている．図１に示すように，国内のコールセンターサービスの市場規模が右肩上がりの成長を続けている事がそれを裏付けている．

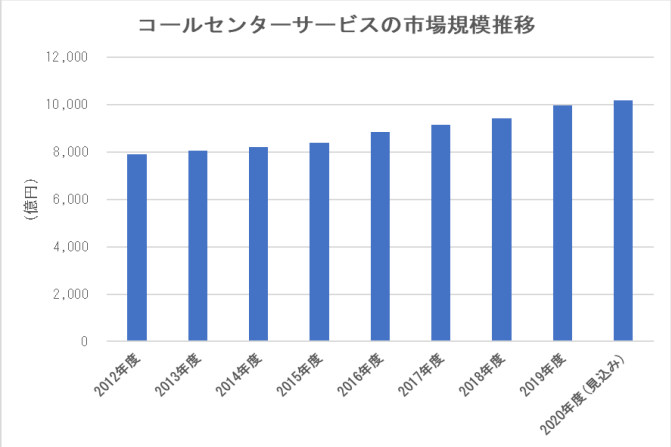

図１　コールセンターサービスの市場規模

（数字は株式会社 矢野経済研究所の調査による．）

そのような市場環境の中で，コールセンターを運営する企業では顧客対応の中心となる優秀なオペレーター（顧客対応を担当する要員）の確保と育成，各種業務の効率化が喫緊の課題となっている．それらの課題に対して IT 技術の活用は有効なソリューションとなりうる．例えば応対品質評価の自動化である．応対品質評価とは，オペレーターが顧客に対して適切な対応を行い，十分な顧客満足度を得られるレベルにあるかを評価する活動を指す．この活動を現在は主に各コールセンターの管理者（スーパーバイザーと呼ばれる，オペレーターの育成や管理を担うマネジメント職種）が人手で行っている．具体的にはコールセンターへの受電から通話終了までの数分から数十分にわたるオペレーターと顧客との対話が録音された音声を管理者が耳で聞き，オペレーターの発話内容が聞き取りやすいか，声の大きさや抑揚の強さが適切かどうか，相手に良い印象を与える声の表情で話している

かなど数十項目の評価項目に対して手作業でスコアを付けている．管理者は公正かつ客観的な評価を行うために専門の教育を受け，評価基準のブレを防ぐために定期的な校正（キャリブレーション）も受けている．この評価作業をコンピューターによって自動化すれば，人手では困難な，あるいは実現不可能な以下の事柄を実現できる．

① 全数評価・・・ひとつのコールセンターにおける１ヶ月あたりの応対通話時間は合計で数千時間から数万時間に及び，その全てに対して人手で評価を行う事は不可能である．そのため，人手による評価は全ての録音データの中から一部をピックアップして行われている．それに対して，十分な性能を持つコンピューターであれば全数評価が可能となる．全数評価を行う事により，たまたま応対が良かった／悪かったといった偶然性を排除して公平な評価ができる．
② 一定の基準による評価・・・定期的に校正（キャリブレーション）を受けたとしても，管理者は人間である以上，その日の体調等により評価にばらつきが出る可能性は否めない．それに対してコンピューターは常に一定の基準で評価を行う事ができる．
③ 人が行う作業の肩代わり・・・管理者による評価作業の一部または全部をコンピューターに肩代わりさせる事により，管理者は他の作業に時間を割く事ができる．もしくは今までよりも少ない人員で業務を遂行する事ができる．

これらを実現する事により，以下の効果が期待できる．
・オペレーター個々のスキルを正しく把握する事による適切な改善指導の実現．
・客観的で公平な評価によるオペレーターのモチベーションの維持，向上．
・評価を行う管理者の肉体的，精神的な負担の軽減．
・管理者が評価作業に掛ける時間を減らし，その代わりに現場でのオペレーターへの指導や援助を充実させる事によるサービスの品質向上．
・評価作業に掛かるコストの削減．

上記の通り，コールセンターの応対品質評価の自動化にはメリットが多い．そのため，これまでにも自動化の試みは行われ，製品やソリューションとして販売されているケースも存在する．例えば富士通株式会社や株式会社日立情報通信エンジニアリングの例では，発話速度，発話かぶりの回数や時間，必須ワードや NG ワードの使用回数を通話録音データから抽出し，コンピューターによる自動評価を行っている[1][2]．しかし，これらは主として音声データから音声認識技術によってテキスト情報を抽出し，音声とテキストのアラインメントを取る（対応づける）事によって実現していると考えられ，先に述べたような「相手に良い印象を与える声の表情で話しているか」といった，言語解析型の音声認識技術だけでは実現できない音響的な評価指標に対する自動評価が行われているものは現時点では見当たらない．

1.2　研究の目的

本研究では，コールセンターにおけるオペレーターの応対品質評価のうち，自動化が行われていない「相手に良い印象を与える声の表情で話しているか」の評価の自動化に向け，その実現可能性の検討を行う．実現手段としては音響解析型の音声感情認識の技術を応用し，人手による評価と比較してどの程度の精度（再現率，適合率）で自動評価ができるかを確かめる．その結果が人手による評価よりも劣る場合は，自動評価の実現に向けてどのような課題があり，どのような解決手段が考えられるかを検討する．

1.3　本論文の構成

本論文は全４章からなる．その構成は以下のようになっている．第１章では，本研究の背景としてコールセンターにおけるオペレーターの応対品質評価の現状と課題，自動化によって期待される効果をまとめ，本研究の目的を述べた．第２章では，応対品質評価の自動化のために利用する音声感情認識技術の概要を述べ，従来技術を用いた音声感情認識の実験結果を示す．また，認識率向上のために本研究で提案する発話末尾の音響特徴量を用いた感情認識の実験結果を示す．第３章では，音声感情認識の技術を使用した提案法によってコールセンターのオペレーターの「相手に良い印象を与える声の表情で話しているか」についての評点を推定した実験の結果を示す．第４章では，本研究のまとめと考察，今後の課題について述べる．

第２章 音声感情認識の実験

2.1 音声感情認識の技術を使用する理由

「相手に良い印象を与える声の表情で話しているか」（以降「声の表情」と略記する）の評価を自動化する手段として音声感情認識の技術を使用する理由について述べる．本研究では，日本国内でコールセンターを運営しているビーウィズ株式会社（以降「ビーウィズ社」と表記）から実際のコールセンターの音声データと人手で応対品質評価を行った結果のデータを提供して頂いた．その際にサンプルとして提供された「声の表情」の評価コメントには「口角が上がった表情で対応できている（いない）」という表現が頻繁に見られた．口角が上がった表情とは笑顔の事であり，ポジティブな感情を声で表現できているかどうかが評価のポイントとなっている事が伺えた．音声データを音声編集ソフトで再生して聞いた場合でも，「声の表情」について高い評点が与えられた音声からは喜びや受容といったポジティブな感情が感じ取られた．また，「声の表情」の評点が低い音声からは不機嫌さや拒絶といったネガティブな感情が感じ取られた．この事から，「声の表情」と感情音声（感情が込められた音声）には共通の音響的な特徴があるのではないかと考えた．音声感情認識は感情音声に特有の音響的特徴を抽出して識別を行う技術であるから，それを応用すれば「声の表情」の自動評価が可能になるのではないかと考え，本研究でその検証を行った．

2.2 音声感情認識について

この節では既存技術の音声感情認識について概観を述べる．

2.2.1 音声感情認識の種類（言語解析型と音響解析型）

音声感情認識は，人が発した音声から機械（コンピューター）によって感情を読み取ることを指す．そのアプローチは大きく言語解析型と音響解析型に分けられる．言語解析型の音声感情認識は音声認識により発話内容をテキスト化し，自然言語処理によって感情の推定を行う．音響解析型の音声感情認識は声の高さや大きさ等の音響的な特徴を元に感情の推定を行う[3]．言語解析型と音響解析型を組み合わせるアプローチ[4]，音声に加えて人の表情の画像解析結果を組み合わせるアプローチも存在する[5]．

本研究では，音響解析型の音声感情認識の技術を用いて「声の表情」の自動推定を試みる．その理由は，コールセンターにおいて管理者が行っている「声の表情」の評価が主としてオペレーターの音声の音響的な特徴を元に行われている為，その自動化を目指す本研究でも同様に音響的な特徴（音響特徴量）を使用する音響解析型のアプローチを取ることが妥当だと考えられる為である．

2.2.2　音声感情認識のアウトライン

音響解析型の音声感情認識において広く行われている手順を図 2 に示す．この手順は教師あり機械学習の枠組みの中で行われる．学習フェーズでは，まず初めに感情音声コーパスに収録された音声データから音響特徴量を抽出する．感情音声コーパスとは感情を込めて発話した音声のデータと感情ラベルを体系的に収集して構築された，音声感情認識のためのコーパスである[6]．感情音声コーパスの詳細は後述する．次に，抽出した音響特徴量を説明変数，感情ラベルを目的変数として教師データを作成し，分類器の学習を行う．推論フェーズでは，感情の推定を行いたい音声データから音響特徴量を抽出し，学習フェーズで作成した分類器を用いて感情の分類を行う．尚，後述する「感情の次元説」の立場に基づいて音声感情認識を行う場合は分類モデルではなく回帰モデルが使用される[3].

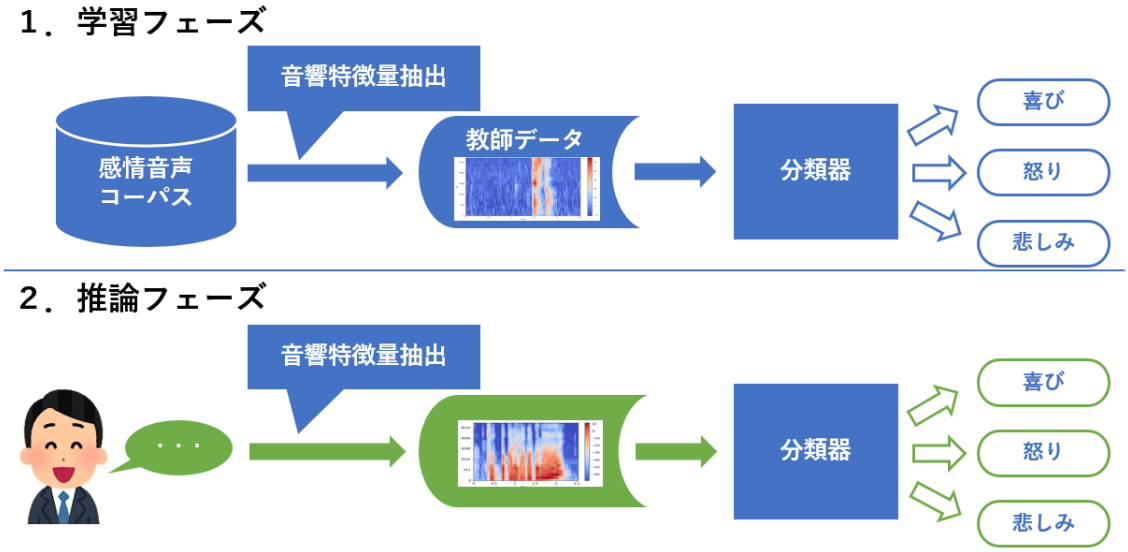

図 2　音声感情認識のアウトライン

2.2.3　感情の種類

音声感情認識が扱う感情の種類について述べる．最初に，感情は何種類存在するのかを考える．我々日本人が感情と聞いてまず思い浮かべるのは「喜」「怒」「哀」「楽」の 4 感情である事が一般的だと考えられる．中国の「五情」はこれに「怨（うらみ）」を加えた 5 感情を指す．米国の心理学者である Paul Ekman は表情認知の研究を元に，文化に依存しない人類共通の感情が「幸福感」「驚き」「恐れ」「悲しみ」「怒り」「嫌悪」の 6 つに分類されるとした．Ekman は後に「おもしろさ」「軽蔑」「満足」「困惑」「興奮」「罪悪感」「功績に基づく自負心」「安心」「納得感」「喜び」「恥」の 11 の感情を追加した．同じく米国の心理学者である Robert Plutchik は「感情の環」の理論で 32 種類以上の感情を定義している．日本の光吉俊二は日本語の感情表現を分類した結果，日本語には感情を表す単語が約 4,500 語あ

り，英語との対応付けが可能なものだけでも 223 種類の感情が存在するとしている[7].

このように感情は細かく分類すれば膨大な種類が存在するが，機械学習による音声感情認識では 3〜10 種類程度の感情を扱う事が多い．あまりに多くの感情を扱おうとすると学習・検証用の音声の収集やラベル付けのコストが増大する事，「幸福感」と「満足」のように似通った感情の識別は機械のみならずラベラーや話者本人にとっても困難である事などがその理由だと考えられる．実用上も，（音声対話システムなどで人間と同等のコミュニケーションができるシステムを構築するような場合は別として）多数の感情を一つのモデルで認識しなければならないケースは少ないと想定される．音声感情認識を社会実装する際には目的を定めて，どのようなシチュエーションでどのような感情を機械に認識させたいのかを明確にする事が重要である[6]. 例えば顧客応対で相手の怒りを検出する等である[3].

2.2.4 hot anger と cold anger

音声感情認識を行うにあたり意識しておくべき hot anger と cold anger について述べる. hot anger は「激しい怒り」，cold anger は「静かな怒り」を指す．これらはいずれも「怒り」の感情であるが，音響的な特徴は異なる事が報告されている[8]．この事は感情認識の目的に応じて適切な感情音声コーパスを選択する事の重要性を示している．例えば顧客の怒りを検出したい場合に cold anger が含まれない感情音声コーパスでモデルの学習を行うと，顧客の静かな怒りを検出できないモデルが出来上がってしまう．従って感情音声コーパスの選択は目的を踏まえて慎重に行う必要がある.

2.2.5 基本感情説と次元説

心理学において，感情の理論には基本感情説と次元説の 2 つの考え方がある[9]．基本感情説は，人間には喜怒哀楽などの基本的な感情がいくつかあり，それぞれが独立したものとして存在しているという説である．次元説は，感情は「快－不快」の軸と「覚醒－睡眠」の軸のような 2〜n 次元のベクトル空間上に配置されるという説である．どちらの説が正しいかという学問上の議論には決着がついていないが，音声感情認識を行う者は感情音声コーパスがこのいずれかの立場に基づいて作成されている事を認識しておく必要がある．基本感情説に基づいて作成されたコーパスは主に分類問題として音声感情認識を行うように設計されている．次元説に基づいて作成されたコーパスは主に回帰問題として音声感情認識を行うように設計されている.

2.2.6 感情音声コーパス

感情音声コーパスは，音声感情認識の機械学習モデルの作成と評価を行うために作成されたコーパスであり，音声データ，感情ラベル，発話内容のテキストから構成される[6]. 研究目的で使用できる英語ないし日本語の複数の感情音声コーパスが研究機関等によって構築され公開されている. 2.2.6.2 節では日本語の感情音声コーパスをいくつか紹介するが，

その前にこれら複数のコーパスの違いを理解する上で重要な「演技音声」と「自発音声」について述べる．

2.2.6.1　演技音声と自発音声

感情音声コーパスに収録される音声は演技音声と自発音声に分けられる．演技音声とは，話者に発話内容のテキストと感情を指示し，話者に演技させて音声を収録したものである．これに対して自発音声は，発話内容と感情を話者に指示する事なく自由に発話させた音声を収録し，発話内容のテキストと感情ラベルを後から付与したものである．これらの 2 種類のコーパスの特徴を表 1 に示す．音声感情認識を行う上では，それぞれの違いと長所・短所を意識して，実施するタスクやその目的に適したコーパスを選択する事が重要である．

自発音声コーパスの長所は，話者の自然な感情が音声に現れている点である．音声の収録時には話者に喜びや怒りなどの自然な感情を生じさせるためにゲームを用いるなどの工夫を行う事が多い．実世界の人間同士の対話から感情を推定するようなタスクでは自発音声からの感情認識が必要となり，自発音声コーパスはこの目的に適している[9]．短所としては感情ラベルの付与と書き起こしテキストを作成する際のコストの高さが挙げられる．感情ラベルの付与は正確を期すために 3 名またはそれ以上のラベラーによって行われ，複数のラベラーで一致したものが正解ラベルとして採用されることが一般的である．そのため，大規模なコーパスを構築する場合は高い費用と時間，十分なスキルを持ったラベラーの確保が必要となる．書き起こしテキストについては，作業負荷軽減と作業効率向上のために音声認識による自動書き起こしを使用したとしても，精度 100 %の音声認識はいまだ実現していないため人手による修正は必要になり，やはり高コストとなる．このようなコストの高さによって，自発音声コーパスは演技音声コーパスと比較して話者数，発話数が少ない小規模なものになりがちである．

演技音声コーパスの長所はラベル付与におけるコストの低さにあり，それによって大規模なコーパスの構築が自発音声コーパスと比較して容易な点にある．話者に読み上げさせるテキストと演じさせる感情をあらかじめ決めてから音声収録を行うため，自発音声コーパスで問題となるラベル付与の困難さが大幅に軽減される．その一方で，演技音声と自発音声では音響的な特徴が異なる場合があると報告されている[10]．話者の自発的な感情を捉えたい場合，演技音声コーパスで学習したモデルではそれが捉えられない可能性がある事を意味しており，これは演技音声コーパスの短所であると言える．

表 1　演技音声コーパスと自発音声コーパスの特徴

演技／自発	自然な感情が音声に現れているか	ラベルの正確さ	コスト	話者数	発話数
自発音声	○	△	高	少	少
演技音声	△	○	低	多	多

2.2.6.2　日本語の感情音声コーパスの例

日本語の感情音声コーパスのうち，研究用に使用できるものを数件ずつ表 2 および表 3 に示す．これらの表は前節までに述べた演技音声／自発音声，基本感情説／次元説に従って分類してある．

これらのコーパスは, 国立情報学研究所（NII）の情報学研究データリポジトリ（Informatics Research Data Repository : IDR）（URL: https://www.nii.ac.jp/dsc/idr/），音声資源コンソーシアム（SRC: Speech Resources Consortium）（URL: http://research.nii.ac.jp/src/organization.html）の Web サイトからの申し込み, あるいはこれらの Web サイトに無いものはコーパスの作成者・権利保有者に直接コンタクトを取る事により，研究目的であれば無償で入手可能である．

表 2　日本語の感情音声コーパス（演技音声）の例

名称	略称	演/自	立場	感情の種類	話者数など	特徴
Japanese Twitter -based Emotional Speech[11]	JTES	演技	基本感情説	怒り，喜び，悲しみ，平静（4 種類）	男性 50 名，女性 50 名 4 感情×50 文 計 20,000 発話	Twitter のつぶやきから感情表現語を含む口語的な文章を，音韻や韻律のバランスを考慮し選出
声優統計コーパス [12]	-	演技	基本感情説	怒り，喜び，ニュートラル（3 種類）	女性 3 名 3 感情 ×100 文 計 900 発話	日本語版 Wikipedia の本文から抽出した音素バランス文
Online gaming voice chat corpus with emotional label(Vol. 2)[13]	OGVC Vol.2	演技	基本感情説	喜び，受容，恐れ，驚き，悲しみ，嫌悪，怒り，期待，平静，その他（10 種類）	男性 2 名，女性 2 名 664 感情依存文 ×4 感情強度 計 2,656 発話	自然な対話で収録した発話の転記テキストについて，発話単位で感情とその強度を指定して発声

表 3　日本語の感情音声コーパス（自発音声）の例

名称	略称	演/自	立場	感情の種類	話者数など	特徴
宇都宮大学パラ言語情報研究向け音声対話データベース [14]	UUDB	自発	次元説	快-不快 覚醒-睡眠 支配-服従 信頼-不信 関心-無関心 肯定的-否定的 （6 次元）	男性 2 名，女性 12 名 (大学生 7 ペア) 計 4,737 発話	音声言語に付随して伝達されるパラ言語情報に主眼を置いて設計・構築された自然対話の音声コーパス．課題遂行を通して友人同士のいわゆる「ため口」を収録
Online gaming voice chat corpus with emotional label(Vol. 1) [13]	OGVC Vol.1	自発	基本感情説	喜び，受容，恐れ，驚き，悲しみ，嫌悪，怒り，期待，平静，その他 （10 種類） 感情強度ラベルあり	男性 9 名，女性 4 名 2 者対話 5 組，3 者対話 1 組 計 9,114 発話	MMORPG と呼ばれるオンラインゲームを対話収録に導入することで自発的で活き活きとした感情を誘発した音声を収録

2.2.7　音響特徴量

音声感情認識で用いられる特徴量（音響特徴量）について述べる．音声からテキスト情報を抽出する通常の音声認識では，音声ファイルから波形データが窓掛け操作によりフレーム毎に切り出され，離散フーリエ変換によって周波数スペクトルに変換される．その後，周波数毎のパワーに変換され，微細構造を落とすメルフィルタバンクを掛ける事で得られる対数メルスペクトルや，更に離散コサイン変換を行う事で得られる MFCC が音響特徴量として標準的に用いられる[15]．これらは人の声道特性（≒言葉を話している人間の喉頭から口蓋，鼻腔の形）を良く表す特徴量である．音声感情認識で用いられる音響特徴量は，この対数メルスペクトルや MFCC に加えて，声の高さを表す基本周波数，音量，音声波形の揺らぎを表すシマーやジッタなど，様々な特徴量が使用される．人の感情は声道の形だけに表れるのではないと考えられる為であり，国内外の多数の研究者により多様な特徴量が提案されている[9][16][17][18][19]．これらのフレーム毎に抽出される特徴量を LLD（Low Level Descriptor）と呼ぶ．

音声感情認識が通常の音声認識と異なるもう一つの点は，音声認識が LLD を直接 GMM-HMM や DNN などの機械学習モデルの入力として用いるのに対して，音声感情認識では各

LLD に対して発話全体の平均や分散，線形回帰直線の傾きや切片などの統計量を計算したものを機械学習モデルの入力として用いる点である．これは，音声に含まれる感情は音声波形の 1 フレーム（10〜数十マイクロ秒）や数フレーム程度の瞬間的な値では捉えることができず，発話全体の傾向を見ることによって捉えることができると考えられている為である[9][20]．複数の LLD のそれぞれに対して複数の統計量を計算して使用するため，音声感情認識で用いられる特徴量は数百〜数千次元の特徴量ベクトルとなる事が多い．音声感情認識では明確に必要な特徴量が未だ判明していないため，関係があると想定される特徴量を全て使用する事が一般的に行われている[20].

先程述べた通り音声感情認識で用いられる音響特徴量は多数の研究者により多様な特徴量が提案されているが，その中でもよく用いられている特徴量セットが INTERSPEECH 2009 Emotion Challenge（IS09）特徴量セット[16]と INTERSPEECH 2010 Paralinguistic Challenge（IS10）特徴量セット[17]である．これらは音声言語処理分野で一流の国際会議である INTERSPEECH で提案された音声感情認識のための特徴量セットで，性能が良い上に openSMILE ツールキット（音声信号から特徴量を抽出できるオープンソースのツールキット）によって簡便に特徴量を抽出できる事から広く用いられている．表 4 に IS09 特徴量セットの内容を，表 5 に IS10 特徴量セットの内容を記載する．本研究ではこれら 2 種類の特徴量セットを使用した．

表 4　INTERSPEECH 2009 Emotion Challenge（IS09）特徴量セット

LLD	音量，F0，MFCC（1-12 次）， その時点での音が声である確率， 波形のゼロ交差率 ※上記の静的特徴量および Δ（1 階差分）
統計量	算術平均，標準偏差，尖度，歪度， 最大値，最小値，最大値と最小値の差分， 最大値位置，最小値位置， 線形回帰直線の傾きと切片， 線形回帰直線からの二乗誤差，
次元数	384 次元

表 5　INTERSPEECH 2010 Paralinguistic Challenge（IS10）特徴量セット

LLD	メル周波数帯（0-7 次）の対数パワー，MFCC（0-14 次）， LSP（線スペクトル対）周波数（0-7 次）， F0，F0env（F0 の包絡），ラウドネス， シマー（振幅方向の波形の揺らぎ）， ジッタ（時間軸方向の波形の揺らぎ）， 差動フレーム間ジッタ（ジッタのジッタ）， 有声音らしさ ※上記の静的特徴量および Δ（1 階差分）
統計量	算術平均，標準偏差，尖度，歪度， 四分位数，四分位間の範囲， 最大値位置，最小値位置， 線形回帰直線の傾きと切片， 線形回帰直線からの線形誤差と二乗誤差， 1%，99%パーセンタイル， 99%パーセンタイルと 1%パーセンタイルの幅， レンジの 75%を超えている時間の割合， レンジの 90%を超えている時間の割合， 入力の総継続時間 F0 のオンセット数（疑似的な音節数）
次元数	1,582 次元

2.3 使用したデータ

ここまで音声感情認識の概観について述べた．ここからは本研究における音声感情認識の実験について述べる．

本研究において感情音声コーパスは JTES（Japanese Twitter-based Emotional Speech）を使用した．JTES は Twitter の呟きから抽出された喜び，怒り，悲しみ，平常の 4 感情を含む音韻・韻律バランス感情文 200 文（各感情 50 文ずつ）について，男女各 50 名の話者に自分の意図する感情をロボット（機械）に伝えるよう指示して発話させた，合計 20,000 発話からなるコーパスである[11]．2.2.6 節で挙げた複数の感情音声コーパスの中から JTES を選んだ理由は，話者数が多いため話者依存性の低い感情識別器の学習ができることを期待した為である．JTES に収録された音声は自発音声でなく演技音声であるが，本研究においてこの点は問題にならないと考えた．本研究の目的はコールセンターのオペレーターの応対品質のうち「声の表情」の自動評価であるが，オペレーターが電話越しに応対する時の声については，たとえ演技であっても「表情がある，感じの良い声」だと通話相手に感じさせる声であれば適切な応対品質として評価されてよいと考えられる為である．

音響特徴量は音声感情認識の先行研究において実績がある IS09 および IS10 を使用した．音声データからの音響特徴量の抽出には openSMILE を使用した．

2.4 従来技術を用いた音声感情認識の実験

コールセンター音声の分析に進む前に，従来技術を用いた音声感情認識の実験を行った．本節ではその結果について述べる．

2.4.1 SVM による音声感情認識

まず初めに，JTES を使った先行研究[21]と同じ音響特徴量（IS09）および分類器（SVM）を用いて音声感情認識の実験を行い，先行研究と同様の結果が得られるかどうかを確認した．実験条件を表 6 に示す．話者についてはテスト条件が open となるように学習データとテストデータを振り分けた．尚，話者についてテスト条件が open であるとは学習データとテストデータに同一の話者の音声が含まれない事を指す．発話内容のテキストについてはテスト条件を close とした（学習データとテストデータに同じテキストを使用した）．JTES の音声データより openSMILE を用いて IS09 特徴量を抽出し，SVM の多クラス分類器を学習させた．SVM のカーネルは線形カーネルを使用した．テストデータに対する分類結果を表 7 に示す．正解率（Accuracy）は 64.5 %となった．適合率と再現率は 0.49～0.72 となった．感情別に見ると「喜び」の識別精度（再現率，適合率，F 値）がやや低い結果になった．

結果の比較のために，先行研究[21]の結果を表 8 に示す．学習データとテストデータの振り分け方と件数が異なる（先行研究[21]の実験では学習データ 1,440 件，テストデータ 40 件である）ため単純な比較はできないが，F 値で見たときに概ね同等以上の識別精度が得られている事が確認できた．

表 6　実験条件（SVM，線形カーネル）

ハイパーパラメーター	
カーネル	線形カーネル
正則化	L2
C(コスト)	1.0
使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	48kHz
特徴量	IS09
テスト条件	話者 open，発話内容 close
学習	JTES 2,800 発話(50 文×4 感情 ×(男性 7 話者＋女性 7 話者))
テスト	JTES 1,200 発話(50 文×4 感情 ×(男性 3 話者＋女性 3 話者))

表 7　感情認識結果（SVM，線形カーネル）

		推定結果			
		喜び	怒り	悲しみ	平静
正解	喜び	148	84	28	40
	怒り	52	207	19	22
	悲しみ	24	11	216	49
	平静	32	14	51	203

発話数：　1,200

	適合率	再現率	F値
喜び	0.58	0.49	0.53
怒り	0.66	0.69	0.67
悲しみ	0.69	0.72	0.70
平静	0.65	0.68	0.66
平均	0.64	0.65	0.64

正解率
64.5%

表 8　先行研究における感情認識結果

先行研究[21]より抜粋

Table 6: openSMILE による感情認識結果

(a) 話者 closed

	適合率 [%]	再現率 [%]	F 値
喜び	75.7	69.0	0.72
怒り	79.5	86.0	0.83
悲しみ	88.1	77.6	0.83
平常	71.9	81.0	0.76

(b) 話者 open

	適合率 [%]	再現率 [%]	F 値
喜び	49.9	35.2	0.41
怒り	50.7	95.2	0.66
悲しみ	96.4	31.8	0.47
平常	49.4	53.6	0.51

SVM のカーネルをガウシアンカーネルに変更して同様の実験を行った．実験条件を表 9 に，実験結果を表 10 に示す．正解率は 64.5 %から 66.6 %に向上し，適合率と再現率も概ね線形カーネルでの結果を上回った．

表 9　実験条件（SVM，ガウシアンカーネル）

ハイパーパラメーター	
カーネル	ガウシアンカーネル
正則化	L2
C(コスト)	1.0
γ	0.0025
使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	48kHz
特徴量	IS09
テスト条件	話者 open，発話内容 close
学習	JTES 2,800 発話(50 文×4 感情 ×(男性 7 話者＋女性 7 話者))
テスト	JTES 1,200 発話(50 文×4 感情 ×(男性 3 話者＋女性 3 話者))

表 10　感情認識結果（SVM，ガウシアンカーネル）

		推定結果			
		喜び	怒り	悲しみ	平静
正解	喜び	156	80	21	43
	怒り	52	203	7	38
	悲しみ	19	19	210	52
	平静	21	17	32	230

発話数：　1,200

	適合率	再現率	F値
喜び	0.63	0.52	0.57
怒り	0.64	0.68	0.66
悲しみ	0.78	0.70	0.74
平静	0.63	0.77	0.69
平均	0.67	0.67	0.66

正解率
66.6%

ここで，先行研究[21]では話者だけでなく発話内容についてもテスト条件が open である事に気付いたため，同じ条件（話者 open，発話内容 open）となるように学習データとテストデータを振り分けて実験を行った．条件を表 11 に，結果を表 12 に示す．正解率は大幅に下がって 58.9 %となり，60 %を下回る結果となった．適合率と再現率も発話内容 close の場合と比較して大幅に低下した．原因としては，発話内容 close の場合には発話内容のテキストの音素に由来する特徴量の値をモデルが学習して不当にスコアが高くなっている事が考えられる．音響解析型の音声感情認識およびコールセンターの応対品質評価における「声の表情」の自動評価においては，発話内容のテキストによらない音響特徴量を用いる必要があると考えられるため，これ以降の実験は全て発話内容 open の条件で行う事とした．

表 11　実験条件（SVM，ガウシアンカーネル，発話内容 open）

ハイパーパラメーター	
カーネル	ガウシアンカーネル
正則化	L2
C(コスト)	1.0
γ	0.0025
使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	48kHz
特徴量	IS09
テスト条件	話者 open，発話内容 open
学習	JTES 6,240 発話(26 文×4 感情 ×(男性 30 話者＋女性 30 話者))
テスト	JTES 960 発話(12 文×4 感情 ×(男性 10 話者＋女性 10 話者))

表 12　感情認識結果（SVM，ガウシアンカーネル，発話内容 open）

		推定結果			
		喜び	怒り	悲しみ	平静
正解	喜び	137	58	15	30
	怒り	69	138	9	24
	悲しみ	40	10	163	27
	平静	22	16	79	123

発話数：　960

	適合率	再現率	F値
喜び	0.51	0.57	0.54
怒り	0.62	0.58	0.60
悲しみ	0.61	0.68	0.64
平静	0.60	0.51	0.55
平均	0.59	0.58	0.58

正解率
58.4%

発話内容についてテスト条件を open にした事による正解率，適合率，再現率の低下を補うため，音響特徴量として IS10 の使用を試みた．実験条件を表 13 に，実験結果を表 14 に示す．正解率は 64.2 %，再現率と適合率の平均は 0.64 となり，発話内容 close の場合には及ばないものの，いずれも 60 %を超える結果となった．

表 13　実験条件（SVM，ガウシアンカーネル，IS10 特徴量）

ハイパーパラメーター	
カーネル	ガウシアンカーネル
正則化	L2
C(コスト)	1.0
γ	sklearn.svm.SVC のデフォルト 1 / (n_features * X.var())
使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	48kHz
特徴量	IS10
テスト条件	話者 open，発話内容 open
学習	JTES 6,240 発話(26 文×4 感情 ×(男性 30 話者＋女性 30 話者))
テスト	JTES 960 発話(12 文×4 感情 ×(男性 10 話者＋女性 10 話者))

表 14　感情認識結果（SVM，ガウシアンカーネル，IS10 特徴量）

		推定結果			
		喜び	怒り	悲しみ	平静
正解	喜び	138	52	15	35
	怒り	42	166	13	19
	悲しみ	20	10	174	36
	平静	23	6	73	138

発話数：　960

	適合率	再現率	F値
喜び	0.62	0.58	0.60
怒り	0.71	0.69	0.70
悲しみ	0.63	0.73	0.68
平静	0.61	0.58	0.59
平均	0.64	0.64	0.64

正解率
64.2%

2.4.2 DNN による音声感情認識

コールセンターの応対品質評価の自動化のため，DNN を特徴量抽出器として利用する事を考えた．詳細は第 3 章で述べるが，本研究において利用できるコールセンターの音声データは数が限られるため，JTES の大量データを用いて DNN で感情分類器を作成し，その中間層の出力を新たな特徴量として応対品質評価用の分類器を作成する事が目的である．実験条件を表 15 に，実験結果を表 16 に示す．正解率は 64.5 %，再現率と適合率の平均は 0.64 となり，SVM と同等の性能の分類器を DNN でも作成できる事が確認できた．

表 15　実験条件（DNN）

基本構造	
中間層	3 層 4096, 2048, 512 ユニット
出力層	4 (喜び, 怒り, 悲しみ, 平静)
ハイパーパラメーター	
活性化関数	ReLU
drop out	0.5 (中間層)
正則化	なし
学習法	Adam
batch size	200
エポック数	10

使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	48kHz
特徴量	IS10
テスト条件	話者 open, 発話内容 open
学習	JTES 6,240 発話 (26 文×4 感情×(男性 30 話者＋女性 30 話者))
評価	JTES 960 発話 (12 文×4 感情×(男性 10 話者＋女性 10 話者))
テスト	JTES 960 発話 (12 文×4 感情×(男性 10 話者＋女性 10 話者))

表 16　感情認識結果（DNN）

		推定結果			
		喜び	怒り	悲しみ	平静
正解	喜び	136	51	13	40
	怒り	42	165	13	20
	悲しみ	17	9	186	28
	平静	29	8	71	132

発話数：　960

	適合率	再現率	F値
喜び	0.61	0.57	0.59
怒り	0.71	0.69	0.70
悲しみ	0.66	0.78	0.71
平静	0.60	0.55	0.57
平均	0.64	0.64	0.64

正解率
64.5%

次に，性別を表すフラグ（0：男性, 1：女性）を特徴量に追加した．JTES の音声データは男女の比率が 1 : 1 であるのに対して, コールセンターのオペレーターは殆どが女性である．男性の声と女性の声はピッチ（声の高さ）などが異なるため，感情認識を行う時にモデルの内部で使用される特徴量は男女で異なっている可能性がある．JTES データとコールセンターデータの男女比の違いがモデルの性能に及ぼす悪影響を回避するため，特徴量に性別を追加した．実験条件を表 17 に，実験結果を表 18 に示す．正解率は 65.2 %，再現率と適合率の平均は 0.65 となり，特徴量に性別を追加する前とほぼ同等の結果となった．尚，2.5 節で詳しく述べるが，DNN モデルでは全く同じ基本構造とハイパーパラメーターを用いてもモデルを再作成する度に正解率がばらつく結果となった（標準偏差 1.5 %程度）．そのため，表 16 と表 18 の細かい違いはここでは無視して，ほぼ同等の結果であると見なす．

表 17　実験条件（DNN，性別あり）

基本構造	
中間層	3 層 4096, 2048, 512 ユニット
出力層	4 (喜び, 怒り, 悲しみ, 平静)
ハイパーパラメーター	
活性化関数	ReLU
drop out	0.5 (中間層)
正則化	なし
学習法	Adam
batch size	200
エポック数	10

使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	48kHz
特徴量	IS10，性別
テスト条件	話者 open，発話内容 open
学習	JTES 6,240 発話 (26 文×4 感情×(男性 30 話者 ＋女性 30 話者))
評価	JTES 960 発話 (12 文×4 感情×(男性 10 話者 ＋女性 10 話者))
テスト	JTES 960 発話 (12 文×4 感情×(男性 10 話者 ＋女性 10 話者))

表 18　感情認識結果（DNN，性別あり）

		推定結果			
		喜び	怒り	悲しみ	平静
正解	喜び	145	58	14	23
	怒り	42	170	9	19
	悲しみ	21	14	175	30
	平静	34	6	64	136

発話数：　960

	適合率	再現率	F値
喜び	0.60	0.60	0.60
怒り	0.69	0.71	0.70
悲しみ	0.67	0.73	0.70
平静	0.65	0.57	0.61
平均	0.65	0.65	0.65

正解率
65.2%

次に，音声データのサンプリングレートを JTES の元データの 48kHz から 8kH にダウンサンプリングした．コールセンターの音声は電話回線と同じ 8kHz で録音されている為である．8kHz で録音された人の声は，元の音声と比較してこもったような音声になり，人の耳で聞いてもはっきり分かるほど違う音となる．そのため感情認識を行う分類器は 8kHz の音声データから抽出した特徴量で学習を行う必要がある．更に，録音ゲインが異なる実際的な録音環境を想定してゲイン正規化を行った．人は激しい怒りの感情（hot anger）を表現する時に声が大きくなる等，音量の大きさが感情認識の際の重要な特徴量になる可能性があるが，録音時のゲインが異なる場合は音声データ全体の音量レベルが変わるため音量を手掛かりにする事ができない．そのため，VAD で非音声区間を除去した上でゲイン正規化を行うという方法で音量レベルを揃えて実験を行った．実験条件を表 19 に，実験結果を表 20 に示す．正解率は 62.9 %，再現率と適合率の平均は 0.63 となり，前項の結果よりそれぞれ 2 %程度低下した．

表 19　実験条件（DNN，8kHz）

基本構造	
中間層	3 層 4096, 2048, 512 ユニット
出力層	4 (喜び，怒り，悲しみ，平静)
ハイパーパラメーター	
活性化関数	ReLU
drop out	0.5 (中間層)
正則化	なし
学習法	Adam
batch size	200
エポック数	10

使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	8kHz
特徴量	IS10，性別
テスト条件	話者 open，発話内容 open
学習	JTES 6,240 発話 (26 文×4 感情×(男性 30 話者＋女性 30 話者))
評価	JTES 960 発話 (12 文×4 感情×(男性 10 話者＋女性 10 話者))
テスト	JTES 960 発話 (12 文×4 感情×(男性 10 話者＋女性 10 話者))

表 20　感情認識結果（DNN，8kHz）

		推定結果			
		喜び	怒り	悲しみ	平静
正解	喜び	146	34	17	43
	怒り	60	151	14	15
	悲しみ	18	7	173	42
	平静	34	16	56	134

発話数：　960

	適合率	再現率	F値
喜び	0.57	0.61	0.59
怒り	0.73	0.63	0.67
悲しみ	0.67	0.72	0.69
平静	0.57	0.56	0.57
平均	0.63	0.63	0.63

正解率
62.9%

2.5　提案法：発話末尾の音響特徴量を用いた音声感情認識の実験

前節までは音声感情認識の既存の技術を用いて分類器の作成を行った．本節では本研究における提案法について述べる．

音声データのダウンサンプリングとゲイン正規化によって生じた性能低下を補う為に，発話末尾の音響特徴量を使用する事にした．その理由は人が感情を込めて話した時に発話末尾の声の調子が特徴的に変化すると考えた為である．具体的にはピッチ（声の高さ）の上がり下がり，震え，掠れなどが現れるのではないかと考えた．コールセンターの評価者が発話の末尾を注意して聞いているというビーウィズ社の担当者の方の言葉もヒントとなった．

提案法を図 3 に示す．発話末尾の効果を強化するために，発話全体の wav ファイル（音声ファイル）と，同じファイルから発話末尾 0.5 秒分を切り出した wav ファイルの両方から IS10 特徴量を抽出し，それらを連結して特徴量ベクトルとした．この特徴量を用いて DNN または LightGBM で分類器を作成した．LightGBM は決定木アルゴリズムに基づいた勾配ブースティングの機械学習フレームワークの一つであり，変数重要度の出力が可能である．この後で示すように発話末尾の音響特徴量を用いた結果，正解率の向上が認められた為，変数重要度を見て発話末尾のどのような要素が正解率の向上に寄与しているかを確認する為に LightGBM を使用した．実験条件を表 21 と表 22 に示す．

尚，DNN モデルでは図 4 に示すように学習を行うたびにテスト結果のばらつきが発生した．そこで，各条件で 100 回ずつ試行して正解率の平均を取り，これを評価指標として結果の比較を行った．LightGBM は同一条件であればテスト結果がばらつく事はなかった．

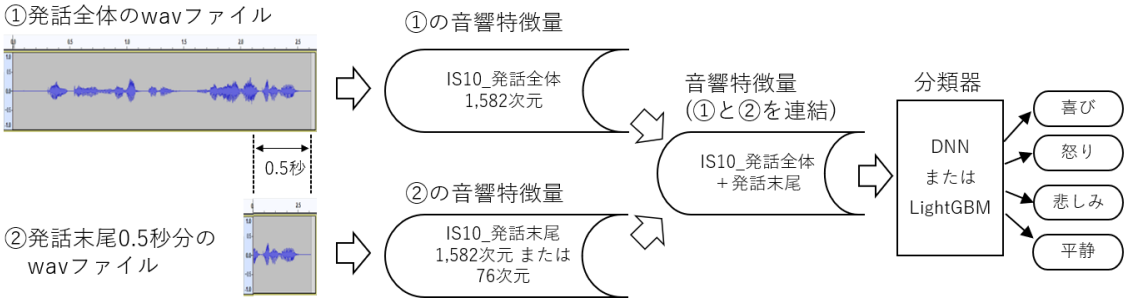

図 3　提案法（発話末尾の音響特徴量を使用した音声感情認識）

表 21　実験条件（DNN）

基本構造	
中間層	3 層 4096, 2048, 512 ユニット
出力層	4 (喜び，怒り，悲しみ，平静)
ハイパーパラメーター	
活性化関数	ReLU
drop out	0.5 (中間層)
正則化	なし
学習法	Adam
batch size	200
エポック数	10

使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	8kHz
特徴量	IS10(発話全体+末尾)，性別
テスト条件	話者 open，発話内容 open
学習	JTES 6,240 発話 (26 文×4 感情×(男性 30 話者＋女性 30 話者))
評価	JTES 960 発話 (12 文×4 感情×(男性 10 話者＋女性 10 話者))
テスト	JTES 960 発話 (12 文×4 感情×(男性 10 話者＋女性 10 話者))

表 22　実験条件（LightGBM）

ハイパーパラメーター	
num_leaves：ノード(葉)の数	31
min_data_in_leaf： 各ノードの最小データ数	20
max_depth：決定木の深さ	制限なし
使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	8kHz
特徴量	IS10(発話全体＋末尾)，性別
テスト条件	話者 open，発話内容 close
学習，評価，テスト	DNN と同じ

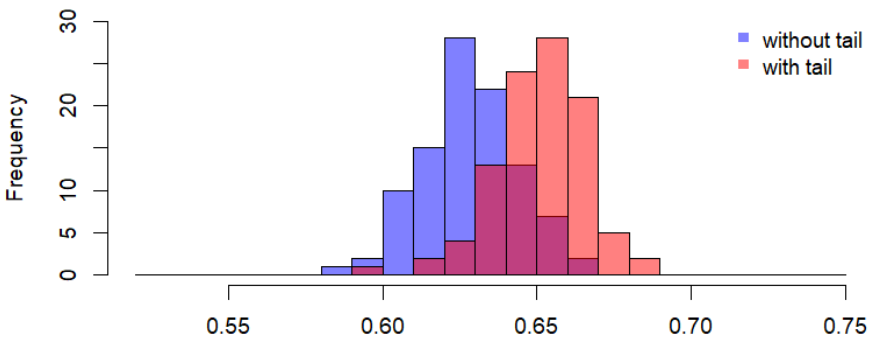

図 4　DNN モデルの正解率のヒストグラム

（青：発話末尾なし，赤：発話末尾あり）

実験の結果を表 23 に示す．モデル A は DNN を使用したモデルである．発話全体から抽出した IS10 特徴量および性別からなる 1,583 次元のベクトルを入力とした場合（末尾なし）の正解率は 62.8 %だった. これに対して発話末尾 0.5 秒から抽出した IS10 特徴量 1,582 次元を追加した 3,165 次元のベクトルを入力とした場合（末尾あり）の正解率は 65.1 %となった. 先にも述べた通り，これらの正解率の値は各条件で 100 回ずつモデル作成とテストを行った結果の平均値だが，Welch の t 検定により p 値 ＜2.2e-16 で有意差が認められ，発話末尾の音響特徴量を追加する事で認識精度が向上する事が確認できた.

モデル B は LightGBM を使用し，モデル A と同じ特徴量を入力としたモデルである．正解率は末尾なしの場合に 63.8 %，末尾ありの場合に 67.4 %となった．LightGBM でも DNN と同様に発話末尾の特徴量を加える事で認識精度が向上する事が確認できた.

モデル C は LightGBM を使用し，認識精度の向上に寄与している特徴量を確認する為に変数重要度の高い特徴量を入力としたモデルである．末尾なしの条件ではモデル B と同じ特徴量ベクトルを使用した．末尾ありの条件ではモデル B から変数重要度が一定値（0.001）を超える発話末尾の特徴量を 76 個抽出し，発話全体の IS10 特徴量と性別からなる 1,583 次元のベクトルにこの発話末尾の 76 次元を加えた 1,659 次元のベクトルを入力とした．正解率は発話末尾なしの場合に 63.8 %，発話末尾ありの場合に 66.1 %となり，発話末尾の特徴量を 1,582 個から 76 個へ大幅に減らしても精度向上の効果がある事が確認できた.

モデル D は DNN を使用し，モデル C と同じ特徴量を入力としたモデルである．正解率は末尾なしの場合に 62.8 %，末尾ありの場合に 63.2 %となった．その差は 0.4 %であるが，Welch の t 検定により p 値 = 0.033 で有意差が認められ，DNN でも発話末尾の少数の音響特徴量を追加する事で認識精度の効果がある事が確認できた.

表 24 にはモデル C，D で用いた発話末尾の特徴量の内訳を示す．メル周波数帯毎(0-7)のパワーは周波数毎の音の強さ，MFCC(0-4)と LSP(0-4)は声道特性，F0(基本周波数)は声の高さ，ラウドネスは人が感じる音の大きさを表している．発話末尾の特徴量のうち，人の声の様々な要素を表現している変数が識別の手掛かりになっている事が確認できた.

表 25 にはモデル C，D で使用した特徴量と未使用の特徴量を示す．声の微細な揺らぎを表すジッタとシマーは未使用であり，発話末尾の特徴量としては重要でない事が確認できた.

表 23　感情認識結果（発話末尾なし／あり）

モデル	手法	正解率			末尾の変数の個数
		末尾なし	末尾あり	差	
A	DNN	62.8 %	65.1 %	+ 2.3 %	1,582 個
B	LightGBM	63.8 %	67.4 %	+ 3.6 %	1,582 個
C	LightGBM	63.8 %	66.1 %	+ 2.3 %	76 個
D	DNN	62.8 %	63.2 %	+ 0.4 %	76 個

表 24　発話末尾の 76 個の特徴量の内訳

特徴量	個数	変数重要度*2
メル周波数帯(0-7)の対数パワー	36	0.112
MFCC(0-4)	17	0.050
LSP Frequency(0-4)	16	0.032
F0	2	0.006
有声音らしさ	2	0.004
ラウドネス	2	0.004
F0env	1	0.002
合計	76	0.211

表 25　モデル C，D で未使用の特徴量

特徴量	発話末尾の特徴量の個数		
	使用	未使用	合計
メル周波数帯(0-7)の対数パワー	36	300	336
MFCC (0-4)	17	193	210
LSP(線ｽﾍﾟｸﾄﾙ対)周波数(0-4)	16	194	210
F0	2	38	40
有声音らしさ	2	40	42
ラウドネス	2	40	42
F0env	1	41	42
MFCC(5-9)	0	210	210
MFCC(10-14)	0	210	210
LSP(線ｽﾍﾟｸﾄﾙ対)周波数(5-7)	0	126	126
ｼﾞｯﾀ (時間軸方向の波形の揺らぎ)	0	38	38
差動ﾌﾚｰﾑ間ｼﾞｯﾀ (ｼﾞｯﾀのｼﾞｯﾀ)	0	38	38
ｼﾏｰ (振幅方向の波形の揺らぎ)	0	38	38
合計	76	1,506	1,582

2.6 本章のまとめ

本章では JTES 音声コーパスのデータを使用し，音響解析型の音声感情認識の従来技術および提案法を用いて SVM，DNN，LightGBM で感情の識別が出来る事を確認した．コールセンターの通話品質と録音環境を想定したサンプリングレート 8kHz の低品質な音声へのダウンサンプリング，音量を感情認識の手掛かりとしないゲイン正規化の操作を音声データに対して行った場合でも感情の識別が可能である事が確認できた．また，これらの操作による識別精度の低下を補う手段として，提案法の発話末尾の音響特徴量を使用する方法が有効である事が確認できた．

第３章　コールセンター音声の分析

3.1　コールセンターの応対品質評価について

本研究の主題であるコールセンターにおけるオペレーターの応対品質評価について，現在人手で行われている評価の内容を説明し，本研究で何を目指すかを述べる．本研究で使用した音声データの提供元であるビーウィズ社では，「声の印象」に関して表 26 に示す 20 項目の評価を実施している．評価者はオペレーターと通話相手のやり取りが録音された音声を耳で聞き，これらの評価項目一つひとつに対して以下の 3 段階の評点を付与している．

・評点「1」：相手の心情を害するおそれがある．
・評点「2」：改善ポイントあり．
・評点「3」：適切な応対範囲．

表 26　コールセンターの応対品質評価の項目

No.	分類	評価項目
1	大きさ	全体的に大きすぎないか
2		全体的に小さすぎないか
3		音量の変化が不自然ではないか
4	語頭	語頭が弱くないか
5	語尾	語尾の跳ね
6		語尾消え（語尾の明瞭さ）
7		語尾伸び
8		語尾上がり
9		語尾下がり
10		語尾の強さ
11	滑舌	全体的に滑舌は悪くないか
12		特定の箇所のみ滑舌が悪い
13	抑揚	抑揚が極端でないか
14		抑揚が弱すぎないか
15		抑揚の場面が適切か
16	スピード	全体のスピードの速さ
17		全体のスピードの遅さ
18		スピードの変化
19	表情	一部表情がミスマッチな箇所が無いか
20		全体として表情があり，感じが良いか

これらの評点を項目毎に重みづけして合計した点数が，音声の音響的な特徴から通話相手が受ける印象を表す「声の印象」の評価点となる．更に，対話の内容（オープニングの名乗りからクロージングまで適切な対応が出来ているか），オペレーターとして求められている問題解決力を発揮できているかなどの複数の評価項目について評点が付与され，オペレーターの総合的な評価点が決定される．

本研究では，表 26 の No.20「全体として表情があり，感じが良いか」（＝「声の表情」）の評価の自動化を目指す．音声感情認識の技術を使用して機械学習により評点を推定する分類器を作成し，人手による評価と比較してどの程度の精度（再現率，適合率）で推定が行われるかを確かめる．その結果が人手による評価よりも劣る場合は，自動評価の実現に向けてどのような課題があり，どのような解決手段が考えられるかを検討する．

3.2　使用したデータ

3.2.1　受領データ

ビーウィズ社から受領したデータについて説明する．以下の 2 種類のデータを受領した．

① コールセンターの応対を，受電から通話終了まで録音した wav 形式の音声ファイル．8kHz のサンプリングレートで収録されている．（受託業務の音声は含まない．）

② Excel 形式のラベルデータ．以下の情報が記載されている．

- 発話の開始時刻と終了時刻．単位はミリ秒．
- 発話内容（音声認識により自動書き起こしされたテキスト情報）．
- 話者がオペレーターと通話相手のどちらであるかを表すフラグ．
- 20 項目の評価項目毎に人手で付与された評点．

尚，オペレーターと通話相手の音声が重なっている発話区間については「評価対象外」のラベルが付与されている．これは音声が重なっている場合に機械学習による推定が困難になると想定されたためビーウィズ社にて付与して頂いた．音声ファイルとラベルデータからは予め人名，住所，電話番号等の個人情報を削除された状態でデータの提供を受けた．

					1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		総合			大きさ			語頭	語尾						滑舌		抑揚			
		72		評価項目	全体的に大きすぎないか	全体的に小さすぎないか	音量の変化が不自然ではないか	語頭が弱くないか	語尾の跳ね	語尾消え（語尾の明瞭さ）	語尾伸び	語尾上がり	語尾下がり	語尾の強さ	全体的に滑舌は悪くないか	し特定の箇所のみ滑舌が悪い	抑揚が極端でないか	抑揚が苦すぎないか	抑揚の場面が適切か
				通話計	87	87	86	85	86	87	82	87	86	87	83	81	87	86	76
				平均	3.0	3.0	3.0	2.9	3.0	3.0	2.8	3.0	3.0	3.0	2.9	2.8	3.0	3.0	2.6
開始時間	終了時間	お客様のI	I	RecognitionText	合計	全体的(	全体的(	音量のう	語頭がも	語尾のり	語尾消え	語尾伸び	語尾上が	語尾下が	語尾のう	全体的(	し特定の	抑揚がも	抑揚がも	抑揚のも
0:00:00.906	0:00:03.367		1	お電話ありがとうございます無事でございます	59	3	3	3	3	3	3	3	3	3	3	3	2	3	3	3
0:00:04.618	0:00:10.845		2	[redacted]																
0:00:11.105	0:00:12.051		1	かしこまりました	58	3	3	3	3	3	3	3	3	3	3	2	3	3	2
0:00:12.355	0:00:17.662		1	今確認いたしますので恐れ入りますがお名前をフルネームでお	57	3	3	3	2	3	3	2	3	3	3	3	2	3	3	3
0:00:18.447	0:00:18.947		2	[redacted]																
0:00:19.342	0:00:20.297		1	かしこまりました	56	3	3	3	3	3	3	3	3	3	3	2	3	3	3	2
0:00:20.617	0:00:23.670		1	ご応募いただいた際のお電話番号よろしいでしょうか	57	3	3	3	3	3	2	3	3	3	3	2	2	3	3	3
0:00:25.414	0:00:27.550		2	[redacted]																
0:00:27.880	0:00:28.373		1	あっはい	60	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
0:00:29.247	0:00:29.375		2	[redacted]																
0:00:29.694	0:00:29.911		1	はい	60	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
			2	[redacted]																
0:00:30.893	0:00:31.297		1	はい	60	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
0:00:31.363	0:00:31.482		2	[redacted]																
0:00:31.783	0:00:33.499		1	はい、【フルネーム】様でいらっしゃいま	60	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
0:00:33.499	0:00:33.681	重なあり	1																	
0:00:33.499	0:00:33.681		2	[redacted]																

図 5　ラベルデータ（Excel ファイル）のイメージ

3.2.2　データの前処理

受領した音声データは受電から通話終了までが 1 つの wav ファイルに記録されていた．これを発話区間ごとに 1 つの wav ファイルとするため，ラベルデータに記載された発話区間の開始時刻と終了時刻の情報を元に音声加工・変換ソフトの sox を用いて wav ファイルの分割を行った．得られた全ての wav ファイルに対して音声波形の振幅の RMS（二乗平均平方根）が一定値となるようにゲイン正規化を行い音量レベルを揃えた．

3.3　提案法：応対品質（声の表情）の自動推定

3.3.1　提案法のアウトライン

「声の表情」を自動推定するための提案法を 2 つ述べる．

提案法 1 では音声感情認識と同じ手法を用いる．その概要を図 6 に示す．学習フェーズではコールセンターの音声データから音響特徴量を抽出し，人手で付与された評点「1」「2」「3」を正解ラベルとして分類器の学習を行う．音響特徴量には IS10 を，分類器には Light-GBM を使用する．これらは第 2 章の実験で良い性能を示していた事から選定した．推論フェーズでは学習済みの分類器を使ってテストデータの評点の推定を行う．

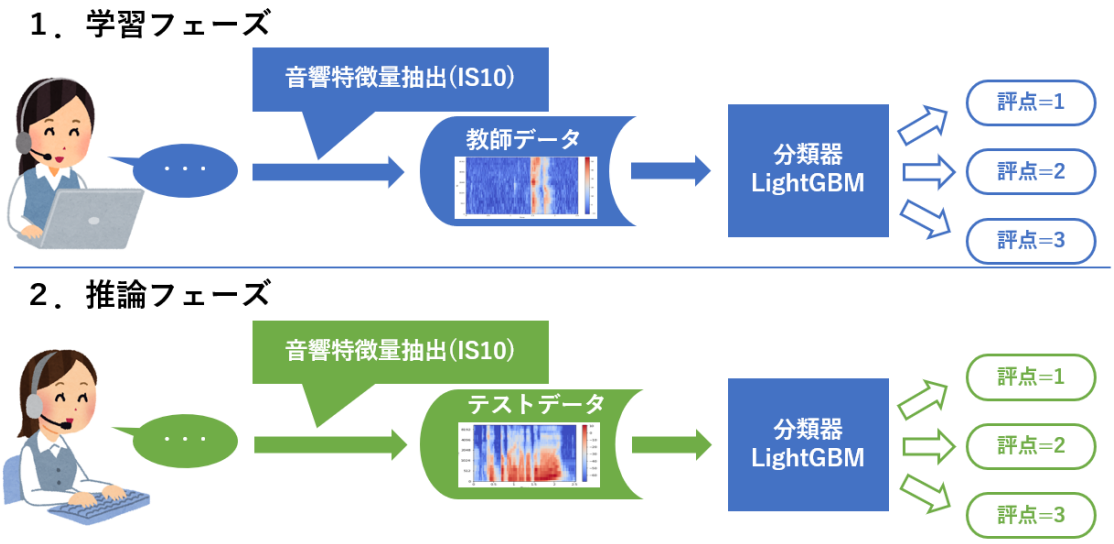

図 6　提案法 1

提案法 2 では，コールセンターの音声データの少なさを補うために JTES データで事前学習した DNN の感情分類器を特徴量抽出器として使用する．

まず，図 7 に示すように JTES 感情音声コーパスから抽出した IS10 特徴量を使用して感情認識の分類器を DNN で作成する．この学習済みの DNN が特徴量抽出器となり，その中

間層の出力は感情（≒声の表情）を捉えた，話者依存性の少ない特徴量となる（3.3.2 節で実験的に述べる）．

次に，図 8 に示すように応対品質の評点を推定する分類器の学習を行う．この時に用いる特徴量として，IS10 特徴量に加えて図 7 で学習させた感情認識の分類器（DNN）の中間層の出力を使用する．評点を推定する分類器には LightGBM を使用する．

最後に，推論フェーズでは図 9 に示すように学習済みの DNN と LightGBM の分類器を使ってテストデータの評点の推定を行う．

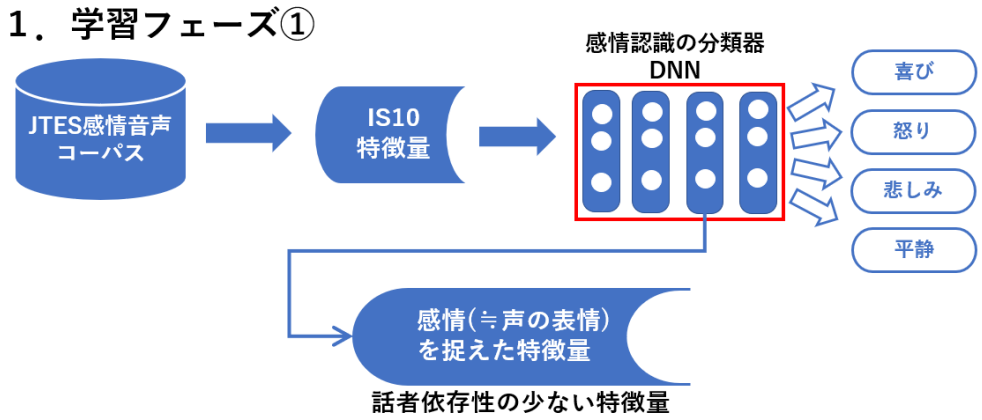

図 7　提案法 2 －特徴量抽出器（DNN）の学習

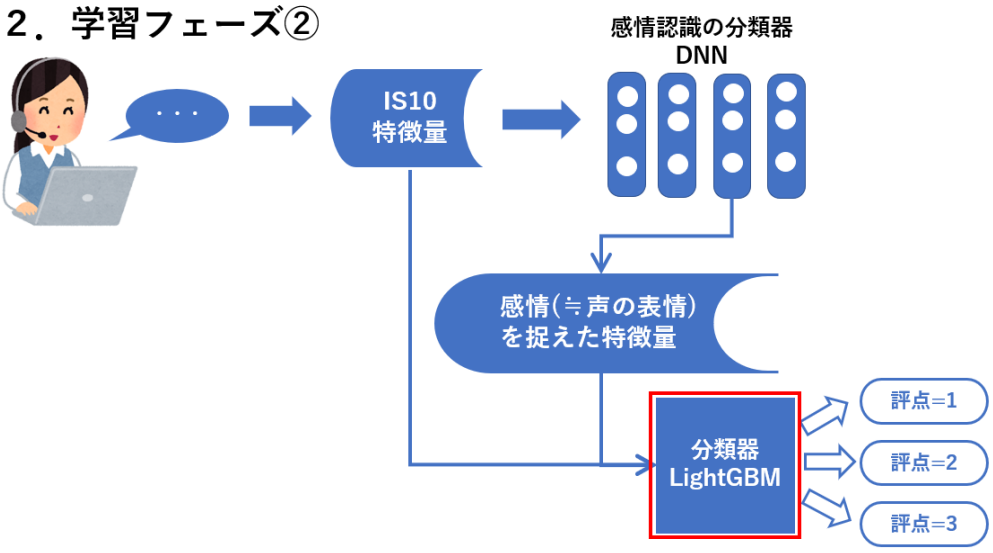

図 8　提案法 2 －応対品質（評点）の分類器の学習

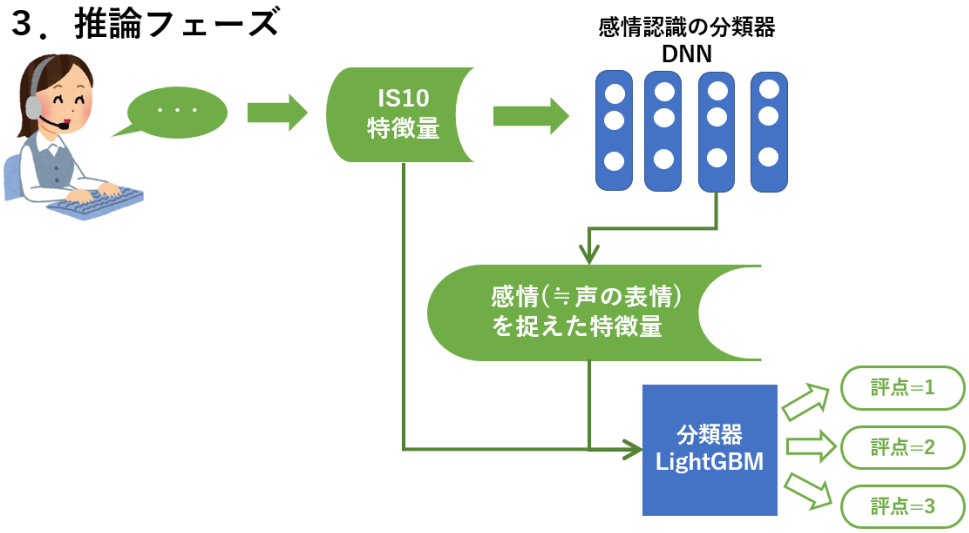

図 9　提案法２－応対品質（評点）の推定

3.3.2　話者依存性の少ない特徴量への DNN を用いた変換

前節で，感情音声コーパスで学習させた DNN の中間層の出力は感情（≒声の表情）を捉えた話者依存性の少ない特徴量となる事が期待できると述べた．その根拠を以下に示す．

図 10 は t-SNE を用いた次元削減により IS10 特徴量を 2 次元にプロットして可視化したものである．JTES 感情音声コーパスから 1 名の話者を無作為に抽出して，喜び，怒り，悲しみ，平静の 4 感情について各 50 発話，合計 200 発話分の特徴量を可視化した．プロットの 4 種類の数字と色が 4 つの感情を表している．この図から，同一の話者であれば感情毎におおむね集まった形で特徴量空間が形成されている事が分かる．

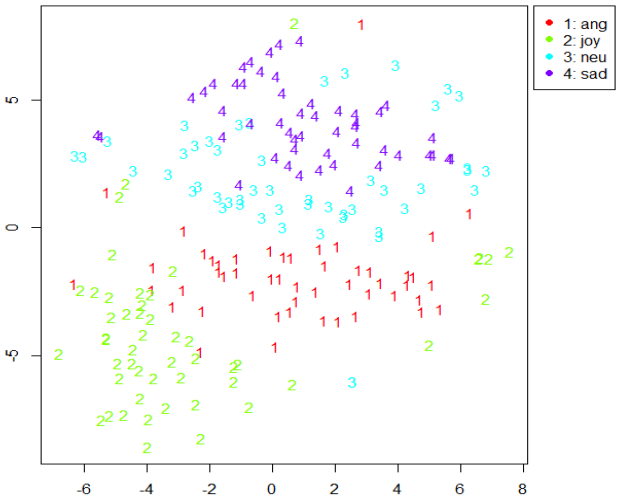

図 10　t-SNE による IS10 特徴量の可視化（1 名分）

次に，3 名分の IS10 特徴量を t-SNE で可視化した結果を図 11 に示す．凡例の「A_xxx」「B_xxx」「C_xxx」がそれぞれ 1 名分の話者を表しており，xxx の部分が感情を表している（ang：怒り，joy：喜び，neu：平静，sad：悲しみ）．この結果を見ると，特徴量空間の中で話者ごとに集まって島を形成している事が分かる．すなわち，IS10 特徴量のままでは話者依存性があるという事である．DNN を用いる事で，これを話者依存性の少ない特徴量に変換できる事を次に示す．

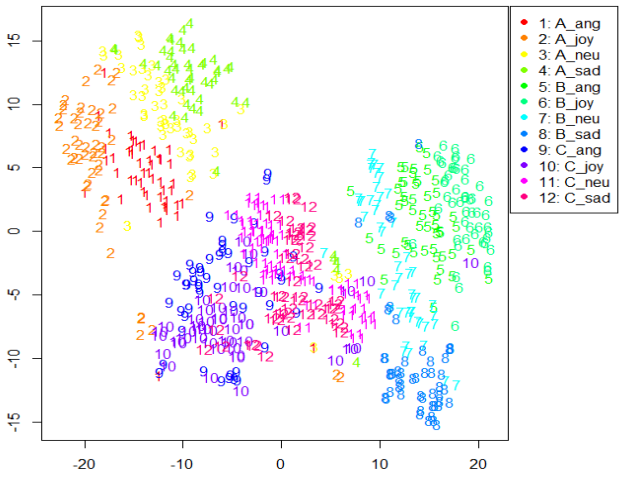

図 11　t-SNE による IS10 特徴量の可視化（3 名分）

図 12 は男女各 10 名，計 20 名の IS10 特徴量を可視化したものである．4 感情を表すプロットが混在しており，単純な境界線では分離できない状態になっている事が分かる．この特徴量を入力として 4 感情を分類する DNN の分類器の学習を行い，その中間層の出力を抽出して t-SNE により可視化を行った．図 13 に中間層 3 層の構造を持つ DNN の 3 層目の出力を，図 14 に中間層 7 層の構造を持つ DNN の 7 層目の出力を示す．IS10 の特徴量ベクトルを DNN に入力して処理する事で特徴量が変換され，感情毎に集まった形になり，単純な境界線で分離できる状態になっている事が分かる．この結果は DNN を用いる事で話者依存性のある IS10 特徴量を話者依存性の少ない特徴量に変換できる事を示している．

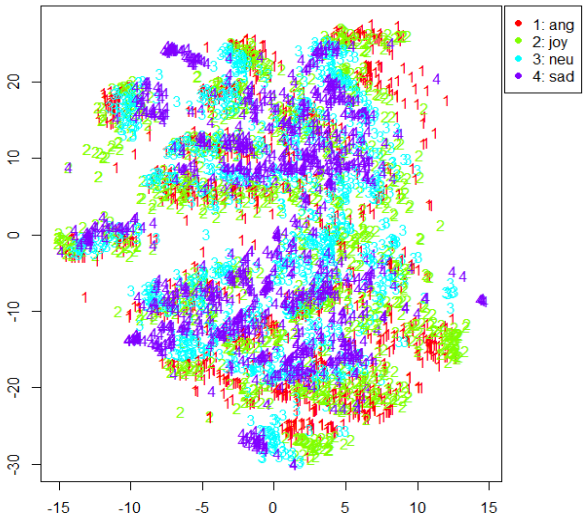

図 12　t-SNE による IS10 特徴量の可視化（20 名分）

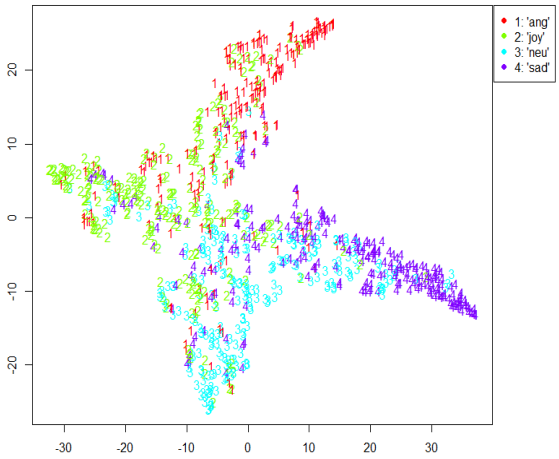

図 13　中間層 3 層の DNN の 3 層目の出力

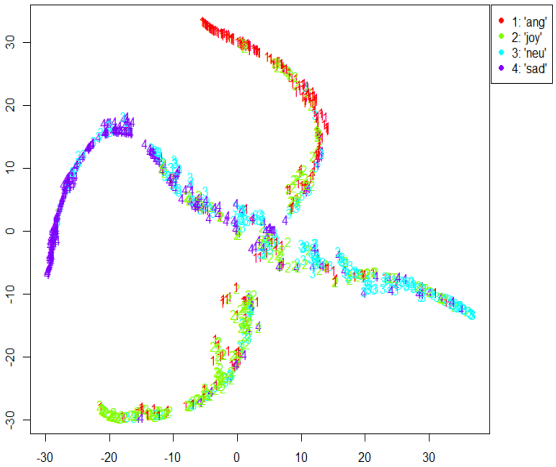

図 14　中間層 7 層の DNN の 7 層目の出力

3.4　音声感情認識の技術を用いた応対品質評価の実験

実際のコールセンターの音声データを使用して提案法の検証を行った．

まず，3.2.2 節に記載した方法で前処理を行った音声データを学習データとテストデータに分割した．分割方法は hold-out 法で行った．テスト条件が話者 open となるようにした上で，できるだけデータに偏りが生じないよう，オペレーターのスキルレベルと性別が均等になるように学習データとテストデータを分割した（表 27）．オペレーターのスキルレベルは評点の合計で判断した．ここで注意するべき事として，評点毎の音声データの件数は評点「2」が最も多く，評点「1」と「3」が少ない不均衡データとなっている．そのまま学習を行うと推定結果がデータ件数の多い「2」に偏りやすいモデルとなる．これを防ぐために評点「1」と「3」のデータを単純コピーで水増しする事でオーバーサンプリングを行い，評点毎のデータ件数が均等になるようにした．尚，オーバーサンプリングの方法には SMOTE（Synthetic Minority Over-sampling TEchnique）や拡張版 SMOTE 等の手法がある．今回，単純コピーの他に SMOTE も試したが，単純コピーの方が正解率，適合率，再現率において概ね良い結果となった為，本稿には単純コピーによるオーバーサンプリングで得られた結果のみを記載する．ダウンサンプリングは本研究では試さなかった．

表 27　学習データとテストデータの分割（hold-out 法）

評点の合計	順位	性別	1の個数	2の個数	3の個数	TRAIN/TEST
58.67	1	女性	0	11	35	01_TRAIN
58.57	2	女性	0	3	25	02_TEST
58.16	3	女性	0	27	35	02_TEST
58.13	4	女性	0	25	39	01_TRAIN
57.48	5	女性	0	30	33	01_TRAIN
57.26	6	女性	0	15	8	01_TRAIN
57.14	7	女性	0	32	4	02_TEST
56.89	8	女性	0	29	25	01_TRAIN
56.72	9	女性	0	23	46	01_TRAIN
56.60	10	女性	0	32	21	02_TEST
56.43	11	女性	0	37	21	01_TRAIN
56.39	12	女性	0	79	23	01_TRAIN
54.55	13	女性	0	16	6	02_TEST
54.48	14	男性	0	39	1	02_TEST
53.96	15	女性	4	19	1	02_TEST
53.95	16	女性	3	18	0	01_TRAIN
53.91	17	女性	0	46	8	01_TRAIN
53.82	18	女性	0	40	5	02_TEST
53.67	19	男性	3	37	2	01_TRAIN
53.60	20	女性	0	50	0	01_TRAIN
53.56	21	女性	0	25	9	02_TEST
53.53	22	女性	0	35	3	01_TRAIN
53.24	23	男性	0	17	0	01_TRAIN
52.96	24	女性	5	47	0	02_TEST
52.38	25	女性	0	52	0	01_TRAIN
51.90	26	女性	2	19	0	01_TRAIN
51.78	27	女性	2	33	1	01_TRAIN
50.71	28	男性	1	34	0	02_TEST
50.52	29	女性	1	22	0	02_TEST
48.96	30	女性	18	29	0	01_TRAIN

提案法 1 の検証結果について述べる．実験条件を表 28 に示す．特徴量については第 2 章の音声感情認識の実験と同様に IS10 特徴量と性別を表すフラグを使用した．結果を表 29 に示す．テストに用いたデータは前節で述べたように不均衡データである為，正解率よりも評点毎の適合率と再現率を重視して結果の評価を行った．評点「2」の推定結果については再現率, 適合率ともに 0.8 以上となった. 人手には及ばないが比較的高い値である. また, 評点「3」の適合率と再現率は約 0.4 と低いものの，大外れはしていない（評点「1」を「3」に誤分類したり「3」を「1」に誤分類しているものは少ない）．これらは，提案法である音声感情認識の特徴量を使用した機械学習で「声の評価」ができる可能性を示す結果である．一方，評点「1」の再現率は 0.18 と低い値になった．この評点「1」の推定精度の改善が課題となる事が分かった.

表 28　実験条件（提案法 1，hold-out 法）

ハイパーパラメーター	
num_leaves：ノード(葉)の数	31
min_data_in_leaf： 各ノードの最小データ数	20
max_depth：決定木の深さ	制限なし
使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	8kHz
特徴量	IS10，性別（提案法 1）
テスト条件	話者 open
テスト方法	hold-out 法
学習	857 発話 (男性 2 話者＋女性 16 話者)
テスト	454 発話 (男性 2 話者＋女性 10 話者)

表 29　「声の表情」の推定結果（提案法 1，hold-out 法）

		推定結果		
		1	2	3
正解	1	2	9	0
	2	1	271	64
	3	4	58	45

（発話総数：454）

	適合率	再現率	F値
1	0.29	0.18	0.22
2	0.80	0.81	0.80
3	0.41	0.42	0.42
平均	0.50	0.47	0.48

正解率
70.0%

前項の結果について，学習データとテストデータの分割のしかたによって偶然低い（もしくは高い）スコアになっている可能性がある．そこで，leave-one-out 法の考え方に基づき，1 回の試行では 1 人の話者の音声データをテストデータとし，残り全ての話者の音声データを学習データとしてモデルの学習を行い，これを話者の人数分（30 回）繰り返す方法を取る事にした．表 30 にその概念を示す．

表 30　学習データとテストデータの分割（leave-one-out 法）

評点の合計	順位	性別	1の個数	2の個数	3の個数	TRAIN/TEST
58.67	1	女性	0	11	35	02_TEST
58.57	2	女性	0	3	25	01_TRAIN
58.16	3	女性	0	27	35	01_TRAIN
58.13	4	女性	0	25	39	01_TRAIN
57.48	5	女性	0	30	33	01_TRAIN
57.26	6	女性	0	15	8	01_TRAIN
57.14	7	女性	0	32	4	01_TRAIN
56.89	8	女性	0	29	25	01_TRAIN
56.72	9	女性	0	23	46	01_TRAIN
56.60	10	女性	0	32	21	01_TRAIN
56.43	11	女性	0	37	21	01_TRAIN
56.39	12	女性	0	79	23	01_TRAIN
54.55	13	女性	0	16	6	01_TRAIN
54.48	14	男性	0	39	1	01_TRAIN
53.96	15	女性	4	19	1	01_TRAIN
53.95	16	女性	3	18	0	01_TRAIN
53.91	17	女性	0	46	8	01_TRAIN
53.82	18	女性	0	40	5	01_TRAIN
53.67	19	男性	3	37	2	01_TRAIN
53.60	20	女性	0	50	0	01_TRAIN
53.56	21	女性	0	25	9	01_TRAIN
53.53	22	女性	0	35	3	01_TRAIN
53.24	23	男性	0	17	0	01_TRAIN
52.96	24	女性	5	47	0	01_TRAIN
52.38	25	女性	0	52	0	01_TRAIN
51.90	26	女性	2	19	0	01_TRAIN
51.78	27	女性	2	33	1	01_TRAIN
50.71	28	男性	1	34	0	01_TRAIN
50.52	29	女性	1	22	0	01_TRAIN
48.96	30	女性	18	29	0	01_TRAIN

話者を変えて
30回繰返し試行

leave-one-out 法での実験条件を表 31 に，実験結果を表 32 に示す．適合率，再現率ともに hold-out 法の場合とあまり変わらない結果となった．F 値の平均と正解率は同じになった．再現率の平均は 0.47 から 0.48 に向上し，適合率の平均は 0.50 から 0.49 に低下したが，その差はわずかである．適合率と再現率を評点別に見ても大きな変化は無かった．各指標の値が大きく変わらなかった事から，hold-out 法での結果は学習データとテストデータの分割のしかたによって偶然低く（もしくは高く）なったものではないと考えられる．尚，これ以降の実験については全て leave-one-out 法を使用した．

表 31　実験条件（提案法 1，leave-one-out 法）

ハイパーパラメーター	
num_leaves：ノード(葉)の数	31
min_data_in_leaf： 各ノードの最小データ数	20
max_depth：決定木の深さ	制限なし
使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	8kHz
特徴量	IS10，性別（提案法 1）
テスト条件	話者 open
テスト方法	leave-one-out 法
学習	1,311 発話 (男性 4 話者＋女性 26 話者)
テスト	1,311 発話 (男性 4 話者＋女性 26 話者)

表 32　「声の表情」の推定結果（提案法 1，leave-one-out 法）

		推定結果		
		1	2	3
正解	1	7	31	1
	2	23	765	133
	3	7	197	147

（発話総数：1,311）

	適合率	再現率	F値
1	0.19	0.18	0.18
2	0.77	0.83	0.80
3	0.52	0.42	0.47
平均	0.49	0.48	0.48

正解率
70.1%

次に提案法２の検証結果について述べる．実験条件を表 33 に，実験結果を表 34 に示す．提案法１と比較して評点「1」の適合率が 0.19 から 0.29 に向上した．これは，評点「2」の音声が評点「1」に誤分類された件数が 23 件から 10 件に減少した事が寄与している．一方，評点「1」の再現率は 0.18 から 0.15 へとわずかに低下した．それ以外の結果はあまり変わらず，強いて挙げれば評点「2」と「3」の再現率がそれぞれわずかに向上した．この結果から，DNN を使用して話者依存性が少なくなるように変換した特徴量を用いる事は「声の表情」の推定に対して一定の効果がある事が分かった．適合率と再現率を指標とした場合に一部が改善され，それ以外の指標への悪影響はほとんど見られなかった．

表 33　実験条件（提案法２）

ハイパーパラメーター	
num_leaves：ノード(葉)の数	31
min_data_in_leaf： 各ノードの最小データ数	20
max_depth：決定木の深さ	制限なし
使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	8kHz
特徴量	IS10，DNN の中間層の出力， 性別（提案法２）
テスト条件	話者 open
テスト方法	leave-one-out 法
学習	1,311 発話 (男性 4 話者＋女性 26 話者)
テスト	1,311 発話 (男性 4 話者＋女性 26 話者)

表 34　「声の表情」の推定結果（提案法２）

		推定結果		
		1	2	3
正解	1	6	31	2
	2	10	775	136
	3	5	196	150

（発話総数：1,311）

	適合率	再現率	F値
1	0.29	0.15	0.20
2	0.77	0.84	0.81
3	0.52	0.43	0.47
平均	0.53	0.47	0.49

正解率
71.0%

ここまで提案法の検証結果について述べた．「声の表情」の自動評価の実現に向けて可能性を示すことができたものの，精度において課題が残る結果となった．精度を改善する方法を探るため，最後に，「声の表情」以外の評点を特徴量に加えた結果について述べる．

3.1 節の表 26 で示した応対品質評価の項目 No.1～18 は，声の大きさ，語頭，語尾，滑舌，抑揚，スピードに関する評価項目であるが，ビーウィズ社の評価担当者によるとこれらの項目は「声の表情」と密接な関わりがある．従って，これらの評価項目について自動推定を行い，その結果を補助特徴量として用いれば「声の表情」の推定結果が改善される事が期待できる．その可能性を探るため，人手によって付与された評価項目 No.1～18 の評点を特徴量に加えて実験を行った．

実験条件を表 35 に，実験結果を表 36 に示す．評点「1」の再現率，評点「2」の適合率，評点「3」の適合率と再現率がわずかに改善されたが，期待したほどの大きな効果は見られなかった．従って，精度改善のためには評価項目 No.1～18 の評点以外の特徴量を使用する必要がある事が分かった．

表 35　実験条件（「声の表情」以外の評点を使用）

ハイパーパラメーター	
num_leaves：ノード(葉)の数	31
min_data_in_leaf： 各ノードの最小データ数	20
max_depth：決定木の深さ	制限なし
使用データ	
ｻﾝﾌﾟﾘﾝｸﾞﾚｰﾄ	8kHz
特徴量	IS10，DNN の中間層の出力，性別，人手によって付与された評価項目 No.1～18 の評点
テスト条件	話者 open
テスト方法	leave-one-out 法
学習	1,311 発話 (男性 4 話者＋女性 26 話者)
テスト	1,311 発話 (男性 4 話者＋女性 26 話者)

表 36　「声の表情」の推定結果（「声の表情」以外の評点を使用）

		推定結果		
		1	2	3
正解	1	8	30	1
	2	21	778	122
	3	3	194	154

（発話総数：1,311）

	適合率	再現率	F値
1	0.25	0.21	0.23
2	0.78	0.84	0.81
3	0.56	0.44	0.49
平均	0.53	0.50	0.51

正解率
71.7%

3.5　本章のまとめ・考察

提案法について実際のコールセンターの音声を用いて「声の表情」の自動評価の検証を行った．その結果，評点「2」の音声については再現率，適合率ともに 0.8 前後と比較的高い精度で推定される事が確認できた．

評点「3」の適合率と再現率はおよそ 0.4～0.5 であり高くはないが，表 37 に示す通り評点「1」「2」「3」をそれぞれランダムに推定した場合の評点「3」の適合率の期待値は 0.27，再現率の期待値は 0.33 であり，それに比べれば高い数値となっている．また，評点「3」の推定結果が外れたケースの大半は評点「2」から「3」への誤分類か「3」から「2」への誤分類であり，「1」から「3」への誤分類や「3」から「1」への誤分類は少数である．つまり評定「3」に関しては推定結果を大きく外したものは少ないため，今後の改善に向けて希望が持てる結果であると言える．ここまでの結果で音声感情認識に用いられる特徴量およびそれを話者依存性が少なくなるように変換した特徴量は「声の表情」を自動推定するのに有効な特徴量であることを示す事ができた．

表 37　ランダムに推定した場合の期待値

		推定結果		
		1	2	3
正解	1	13	13	13
	2	307	307	307
	3	117	117	117

（発話総数：1,311）

	適合率	再現率	F値
1	0.03	0.33	0.05
2	0.70	0.33	0.45
3	0.27	0.33	0.30
平均	0.33	0.33	0.27

正解率
33.3%

一方，評点「1」の再現率はランダムに推定した場合の期待値の 0.33 を下回る低い値であり，今後に向けての課題である．評点「1」の声の特徴を捉える音響特徴量が必要である．

推定に失敗した評点「1」の音声を耳で聞いた時，ぶっきらぼうで強い調子に聞こえる声が複数あった．したがって語調の強さ，とげとげしさを捉える特徴量を考える必要がある．具体的にはパワーの分散もしくはΔ（1 階差分）の統計を取った時に評点「1」の音声では分散やΔが大きいといった特徴があるかもしれない．あるいはその分布に特徴があるかもしれない．もし，ヒストグラムを描いた時に評点「1」の音声のパワーの分散やΔは値の大きな方に偏った（左に裾野の広い）分布になっていたり，あるいはピークが立っていたりする一方，評点「3」の音声のヒストグラムは左右対称で丸みを帯びているなどの特徴が見られれば，パワーの分散やΔの統計量が有効な特徴量になる可能性がある．また，評点「1」の音声には平板で無表情に感じられるものもある．分散やΔが極端に小さい場合も「声の表情」が適切でないという評価を下すようにモデルを学習させると良いかもしれない．このように，今後は音声の個別の特徴量の分布を詳細に見ていくアプローチが必要になると考えられる．

また，評点「1」と「3」の音声データは評点「2」の音声データと比較して件数が少なく，特に評点「1」の音声データは 39 件であり評点「2」の音声データ 921 件と比較すると極端に少ない．今後ビーウィズ社のご協力のもとでデータ件数を増やす事ができれば精度が向上する可能性があると考えられる．

第４章　結論

本研究ではコールセンターにおけるオペレーターの応対品質評価のうち「声の表情」の評価の自動化を目標として，音声感情認識の技術を応用してその実現可能性を検討した．

第２章では感情音声コーパスの音声データを使い，既存の音声感情認識の技術を使用して音声感情認識の実験を行った．その結果，電話回線のサンプリングレートである 8kHz の低品質な音声データで，かつ録音後にゲイン正規化を行った音声でも音声感情認識ができる事を確認した．また，本研究における提案法である発話末尾の音響特徴量を使用する方法により感情認識の正解率が向上する事を確認した．

第３章ではコールセンターの音声データを使い，音声感情認識の技術を応用した提案法によって「声の表情」の自動評価を行った．その結果，人による評価には及ばないものの，データ件数を多く確保できた評点「2」の音声については人による評価結果に対して 8 割程度の再現率と適合率で推定される事を確認した．音声感情認識で用いられる特徴量および本研究における提案法で変換された特徴量は「声の表情」の評価の自動化に向けて有効な特徴量である事が確認できた．

一方，評点「3」と評点「1」の推定精度には課題が残る結果となった．特に評点「1」の再現率が低く，今後は精度向上に向けてこれらの音声の特徴を捉えることのできる音響特徴量を検討していく必要がある事が分かった．

謝辞

本研究を進めるにあたり，多くの方にご協力を賜りました．ここに，心より感謝の意を表します．特に，滋賀大学大学院データサイエンス研究科 市川治教授には，指導教員として多大なご指導とご支援を頂きました．研究を進めるための貴重なアドバイスを頂き，思うような進捗が出せなかった時にも温かい励ましの言葉を常に掛けて頂きました．心から感謝申し上げます．また，本研究を進める為に不可欠なコールセンターの音声データをご提供頂き，その全てに機械学習用の詳細なラベルデータを付与して下さり，現場における品質評価の実際や評価のポイントをご教示下さりましたビーウィズ株式会社のご担当者の皆様に深く感謝申し上げます．同じく本研究を進める為に不可欠であった JTES 感情音声コーパスは東北大学大学院工学研究科伊藤・能勢研究室よりご提供頂きました．深く感謝申し上げます．

参考文献

[1] 富士通株式会社, “応対自動評価システム”, URL:https://www.fujitsu.com/jp/services/application-services/enterprise-applications/crm/voicetracking/qualitymanager/

[2] 株式会社日立情報通信エンジニアリング, “音声分析サービス for コンタクトセンター”, URL:https://www.hitachi-ite.co.jp/products/voice_analysis/index.html

[3] 鈴木基之, “音声に含まれる感情の認識”, 日本音響学会誌 71 巻 9 号(2015), pp.484-489, 2015.

[4] 有本泰子 *et al*, “「怒り」の発話を対象とした話者の感情の程度推定法”, 自然言語処理 Vol.14 No.3, pp147-163, 2007.

[5] 岡田敦志 *et al*, “表情･音響情報･テキスト情報からのリアルタイム感情推定システム”, *The 31st Annual Conference of the Japanese Society for Artificial Intelligence* , 2017.

[6] 森大毅, “感情音声の研究を始める人のための音声コーパス入門”, 日本音響学会 2019 年春季研究発表会スペシャルセッション[音声 A／音声 B], URL:https://speakerdeck.com/hirokisince1998/gan-qing-yin-sheng-falseyan-jiu-woshi-meruren-falsetamefalseyin-sheng-kopasuru-men

[7] 株式会社 AGI, “定量精神分析研究の動向”, URL: https://www.agi-web.co.jp/technology/trend.html

[8] 池本真知子 *et al*, “感情判別における声質の影響”, 感情心理学研究 2009 年 第 16 巻 第 3 号, pp.209-219, 2009.

[9] 森大毅, “音声からの感情・態度の理解”, 電気情報通信学会誌 Vol.101 No.9, 2018.

[10] 有本泰子 *et al*, “音声チャットを利用したオンラインゲーム感情音声コーパス”, 日本音響学会講演論文集 2013 年 9 月, pp.385-388, 2013.

[11] Emika Takeishi, "Construction and Analysis of Phonetically and Prosodically Balanced Emotional Speech Database", *Proceedings of Oriental COCOSDA*, pp.16-21, 2016.
[12] 日本声優統計学会, “声優統計コーパス”, URL:https://voice-statistics.github.io/
[13] "感情評定値付きオンラインゲーム音声チャットコーパス（OGVC）", URL: https://sites.google.com/site/ogcorpus/
[14] “UUDB 宇都宮大学 パラ言語情報研究向け音声対話データベース”, URL: http://uudb.speech-lab.org/index.html
[15] 西川仁, 佐藤智和, 市川治, 清水昌平, “テキスト・画像・音声データ分析”, pp.139-178, 講談社, 2020.
[16] B.Schuller *et al*, "The Interspeech 2009 emotion challenge", *Proc.INTERSPEECH*, pp.312-315, 2009.
[17] B.Schuller *et al*., “The INTERSPEECH 2010 Paralinguistic Challenge”, *Proc. Interspeech*, pp. 2794-2797, 2010
[18] 千吉良好紀 *et al*, “漸次的な感情認識法における excitation pattern と基本周波数の利用法の検討”, 日本音響学会講演論文集 2014 年 9 月, pp.387-388, 2014.
[19] 竹部真晃 *et al*, “音声感情認識における声門特性に基づく特徴量の検討”, 日本音響学会講演論文集 2015 年 9 月, pp.113-116, 2015.
[20] 羽田優花 *et al*, “日本語感情音声コーパス JTES を対象とした感情認識の基礎検討”, 情報処理学会東北支部研究報告 Vol.2019 No.A3-1, 2019.
[21] 武石笑歌 *et al*, “感情音声データベース構築に向けた音韻・韻律バランス感情音声の収録と分析”, 日本音響学会講演論文集, pp.355-358, 2016.