修士論文

発話者の意図の確信度を用いた柔軟な応答生成

2010 年 2 月 9 日

指導教員　広瀬 啓吉 教授
峯松 信明 准教授

東京大学情報理工学系研究科
電子情報学専攻　48-086415

高橋 琢己

内容梗概

人間と機械が音声対話を介してタスクを解決するようなシステムでは, 言語情報の理解だけでなくユーザの発話意図や感情をシステムが理解することがとても重要である．発話者の感情や意図を, システムが自動認識することで, ユーザの発話を誤認識する率を減らすだけでなく, よりヒューマンフレンドリーなシステムを構築することができる.

感情や意図の識別では, 音声における感情や意図の表出を特徴量の形で抽出し, それらが感情や意図の変化によってどのように変化するのかをモデリングするという方法が最も一般的である．しかし, 利用される特徴量は様々で, 収録時の条件（自然発声かそれとも演じてもらったものか, など）や、使用する言語など, 実験者によって実験環境に違いがあるため, どの特徴量が最適かという議論はしにくい．また, 意図や感情の識別（認識）についての研究は多くなされてきたが, その結果を用いてどのように対話管理を行うかという観点に踏み込んだ研究例は少なく, そのためか離散的なクラスとして意図や感情を定義した上で, どこまで識別率を上げられるかということに終始している研究が多い．人間の心的状況を離散的に扱うことは対話管理の面から見ると, 融通が利かずユーザにとって逆に使いづらいシステムになってしまう危険性をはらんでいる．クラスを連続的に設定するのは難しいので, 識別結果のみからは見えてこない情報をうまく処理して，クラスとクラスの間に落ち込んだ発声に柔軟に対応することが必要である.

本稿では, まず話者の意図の識別実験を行い, 音響特徴については，MFCC 及び ST 特徴が識別精度及び実用面での取り扱いも考慮して本研究で想定したタスクに対して有効であることを示した．韻律的特徴については，自動抽出時の誤差及び音素持続時間の自動抽出の際に音声認識結果である言語情報が必要であることなどから，本研究で試行した特徴量ベクトルの作成法については，不備があったといわざるを得ない．次いで, 対話管理の観点から, 話者の意図が定まっていない場合も含め, 話者の意図の決定度合いを識別して柔軟な応答生成に活かすべく，意図の “確信度” という考え方を導入し，対話戦略について検討を行った．その結果，2 値の識別においては確信度の概念を導入することで誤識別を減らし，実際のクラス分け以上の柔軟な応答が可能であることを示した．また，多値の場合においても，2-best の識別結果の分析を行うことで識別誤りを起こす確率の高い発話を 2 値の場合とは別に定義し，それらに対して柔軟な対応をとることで，精度の改善が見込めることを示した.

目次

図目次

表目次

第1章

序論

1.1　本研究の背景

人間と機械が音声対話を介してタスクを解決するようなシステムでは，言語情報の理解だけでなくユーザの発話意図や感情をシステムが理解することがとても重要である．発話者の感情や意図を，システムが自動認識することで，ユーザの発話を誤認識する率を減らすだけでなく，よりヒューマンフレンドリーなシステムを構築することができる．

感情や意図の識別では，音声における感情や意図の表出を特徴量の形で抽出し，それらが感情や意図の変化によってどのように変化するのかをモデリングするという方法が最も一般的である [1]．しかし，利用される特徴量は様々で，収録時の条件（自然発声かそれとも演じてもらったものか，など）や、使用する言語など，実験者によって実験環境に違いがあるため，どの特徴量が最適かという議論はしにくい．また，意図や感情の識別（認識）についての研究は多くなされてきたが，その結果を用いてどのように対話管理を行うかという観点に踏み込んだ研究例は少なく，そのためか離散的なクラスとして意図や感情を定義した上で，どこまで識別率を上げられるかということに終始している研究が多い [2, 3]．人間の心的状況を離散的に扱うことは対話管理の面から見ると，融通が利かずユーザにとって逆に使いづらいシステムになってしまう危険性をはらんでいる．クラスを連続的に設定するのは難しいので，識別結果のみからは見えてこない情報をうまく処理して，クラスとクラスの間に落ち込んだ発声に柔軟に対応することが必要である．

本稿では，まず話者の意図の識別実験を行い，タスク指向の音声対話システムにおける最適な識別の特徴について検討を行った．次いで，対話管理の観点から，話者の意図が定まっていない場合も含め，話者の意図の決定度合いを識別して柔軟な応答生成に活かすべく，意図の “確信度” という考え方を導入し，対話戦略について検討を行った．

1.2　本論文の構成

本論文は，以下のように 7 つの章より構成される．第 1 章（本章）では，本論文の背景・目的などを述べた．第 2 章では，音声言語処理の基礎に深く関わる音響特徴について述べる．第 3 章では，一般的な音声対話システムの構成について紹介し，音声対話システムの要である対話管理技術について先行研究を例にとって様々な方法論について述べる．尚，第 3 章にで，音声研究における本研究の位置づけを明確にする．第 4 章では，本研究における対話管理に深く関わる，音響特徴量を用いた話者の意図 · 感情識別手法について，先行研究を例に述べる．第 5 章では，本研究の話者意図識別に対するアプローチについて音響特徴の面から説明し，実際に行った実験について述べる．第 6 章では，第 5 章で行った話者の意図の識別の結果得られた情報を用いて応答生成に活かす方法について識別のクラスが 2 値の場合と多値の場合に分けて述べる．最後に，第 7 章では，本論文をまとめ，今後の課題について述べる．以降，次章より本論を進めていくこととする．

第2章

音響特徴量

2.1　はじめに

音声から得られる情報はその目的や分析手法によって様々である．本章では，本研究で利用するものを中心に，音声情報処理の各技術においてよく利用される音響特徴量について述べる．

2.2　韻律的特徴

音声における最小の単位は個々の音であるが，これを単音または，分節と呼び，その音響的特徴を分節的特徴という．対して，複数の分節にわたって発現する特徴を超分節的特徴といい，中でも音の高低 · 強弱 · リズム · テンポを特に韻律的特徴という [4]．韻律的特徴は話者の発話意図や発話時の感情などのパラ言語情報 · 非言語情報を伝達する担い手と言われており [5]，認識，合成，対話管理の各分野で多く利用されている特徴量である．以下では，韻律的特徴と対応する物理量について話者の感情や意図と関連が深いものについて説明していく．

2.2.1　基本周波数

韻律的特徴の中でも音の高低の時間変化は，音の強弱や，アクセント位置などの特徴とも関連が深く重要な特徴となっている．音の高低の時間変化の知覚量はピッチと呼ばれる．また，音の高低に対応付けられる物理量としては基本周波数がある．基本周波数は音声波形の基本周期の逆数として定義され，声帯の振動数に対応する．よって，声帯が振動しない無声音においては基本周波数は観測されず，その部分において基本周波数とピッチは一致しない．しかし，ピッチは感覚量であるがために観測される音声波形から直接得ることは難しい．そこで，通常は音の高低の時間変化を捉えるためには基本周波数パターンを観測する．

その情報量の多様さにより，音声分析の場面でよく利用される基本周波数であるが，基本周波数パターンの自動抽出は容易ではない [6]．声門励起音源は実際には準周期的であるにすぎず，一回ごとに周期も振幅も波形も変化する．そのため，どの時間区間を基本周期として採用すればよいのかは自明ではない．また，収録時の周辺雑音や，基本周波数の非定常性 · 非線形性も抽出の際には障害となる．これらの問題は現在も根本的には解決されておらず，基本周波数を特徴量として利用する場合には，抽出の段階で若干の誤差が生じることを理解した上で利用しなければならない．

また、基本周波数は音声合成技術においても非常に重要な役割を担っている．基本周波数パターンの生成過程は図 2.1 に示すように抽象化して比較的簡単なモデルで表現することが出来る [4]．輪状甲状筋の斜部の瞬間的な活動を理想化してインパルス関数で表したものをフレーズ指令といい，フレーズ指令の臨界制動の 2 次線形系の応答をフレーズ成分という．フレーズ成分は語調成分とも言い，文の構造や区切りを示す成分である．また，輪状甲状筋の直部の持続的な活動を理想化してステップ関数で表したものをアクセント指令

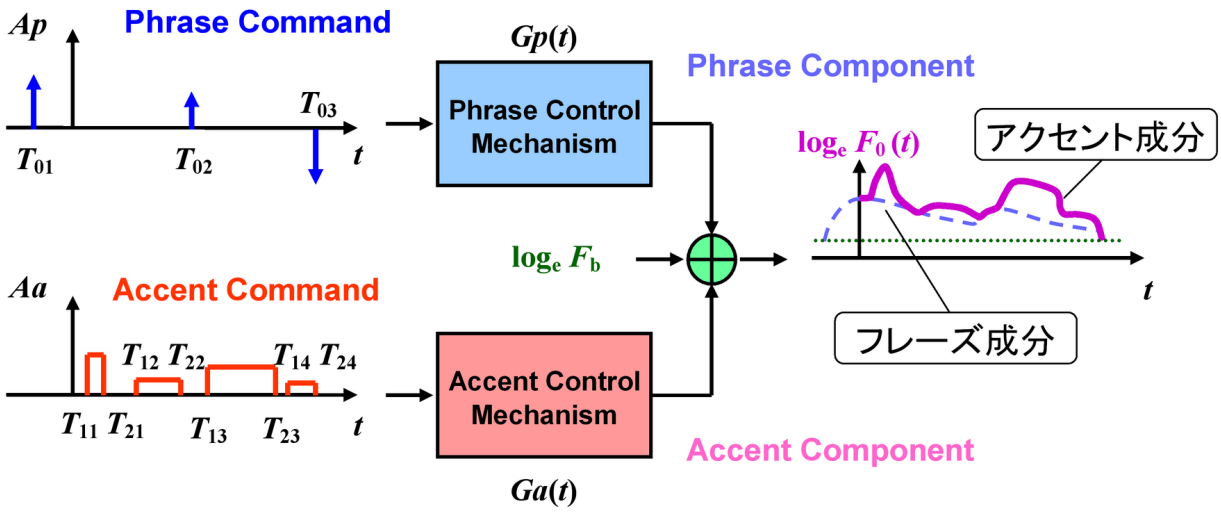

図 2.1: 基本周波数パターン生成過程モデル

といい，アクセント指令の臨界制動の 2 次線形系の応答をアクセント成分という．モデルの最終的な出力としての基本周波数パターンは，これら 2 種類の成分と固定項 $\log F_b$ との和で表すことが出来る．よって，フレーズ成分とアクセント成分を制御することによって任意の基本周波数パターンを生成できることから，従来の朗読調の合成音声に韻律的特徴を付加することで，より表現豊かで，ヒューマンフレンドリーな音声合成を実現しようという研究が近年進められている [7]．前に述べたように，声の高低の時間変化はパワーやアクセントなどのほかの情報とも関連が深く，そのため付加される特徴として基本周波数と音素の持続時間（この 2 つの特徴によって音の高低の時間変化＝音調が表される）がよく利用される．

2.2.2　その他の韻律的特徴

音源信号強度や，声質なども話者の感情や意図と関連が深い韻律的特徴である．

音源信号は，発声後声道を通って観測者が信号を受け取るまでの間に増幅 · 減衰するため，音源信号そのものの強度を観測するためには，発話者に観測器具を取り付けるなど特殊な処理を必要とする．そのため，一般的には収録条件が一定であるという仮定の下，観測された信号の強度を音源信号の強度として扱う．

声質も，非言語情報 · パラ言語情報を伝達する特徴である．一口に “声質” と言うと，話者特有の声の特徴や声道 · 鼻腔での特徴的な声の質などいろんな解釈が可能であるが，ここでは狭義での声帯振動のモード（発声様式）によって特徴付けられる声の質を指す [8]．石井らはこの声質に着目し，それぞれの発話様式と伝達されうるパラ言語情報（石井らはこれを発話行為と呼んでいる）との対応関係を調べている [9]．しかし，声質のラベリングを正確に行うのは，声質に関する知識と分類の経験がある者でなければ難しく，また，話者の地声の影響を色濃く受けてしまうなど，まだ解決されていない課題も多い．

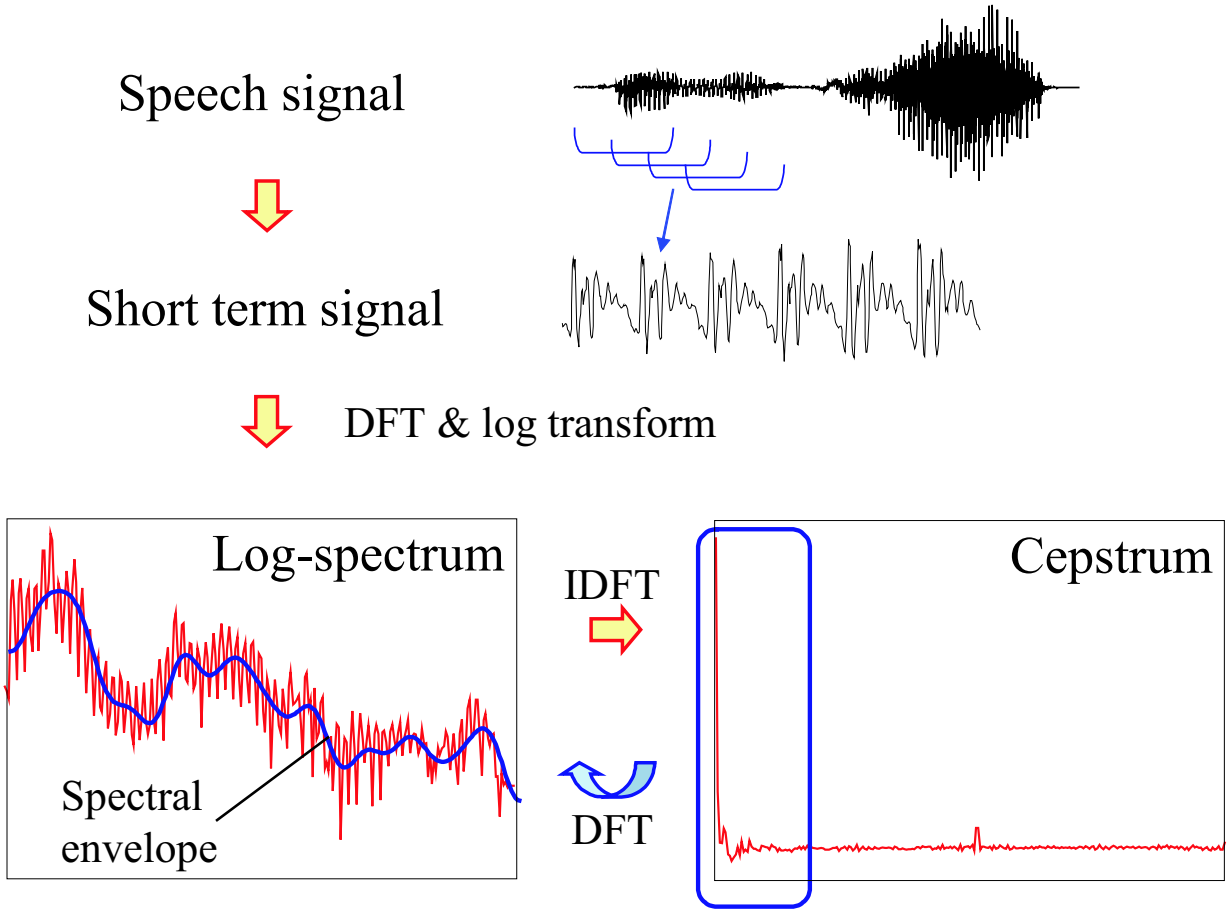

図 2.2: ケプストラム抽出

2.3　ケプストラム特徴

音声信号は，10 ミリ秒程度の分解能で見ると概ね定常であるとみなすことができることから，フレーミングを行い，各フレームに対して特徴量を求めるというやり方が主流となっている [10]．音声波形からケプストラムを抽出する過程を Fig.2.2 に示す．まず音声波形から数十ミリ秒程度を 1 つのフレームとして切り出し，その区間について短時間の離散フーリエ変換（Discreter Fourier Transform; DFT）を施し，その区間の周波数特性であるスペクトルを抽出する．その後，対数パワースペクトルに対して逆離散フーリエ変換（Inverse DFT; IDFT）を施して得られるのがケプストラムである．音声の特徴は，その周波数特性であるスペクトルによく現れるため，初期の音声認識システムにおいては，スペクトルの最も顕著な特徴であるフォルマントを利用していた．しかし，ケプストラム領域の特徴を用いることにより特徴量の各次元の独立性が高まり，より効率的な照合が行えるようになった．現在では，このケプストラムに人間の聴覚特性を反映させた特徴の一つであるメル周波数ケプストラム特徴（Mel-Frequency Cepstrum Coefficient; MFCC）[11] が音声認識において代表的な地位を確立したと言える．

2.3.1　聴覚特性に基づくケプストラム特徴

人間の音の高さに対する周波数分解能は低い周波数ほど細かく，高い周波数のど粗い．このような人間の知覚特性をよく近似する尺度としてメル尺度がある．これは，音の周波数に対してほぼ対数に近い特性を示す．MFCC は，このメル周波数（メル尺度化された周波数）軸上に等間隔に配置された三角窓によるフィルタバンク処理を，音声にかけることで得られる．

なお，メル周波数は以下の式で表される周波数ウォーピングにより求めることが出来る．

$$f_{mel}(f) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.1}$$

2.3.2　時間方向に動的なケプストラム特徴

スペクトルの時間軸に対する動的な特徴を捉えるために Δ ケプストラムを用いることも多い．ケプストラム自体は，各フレームごとの静的な情報だが，実際の発話においては，音素は周辺の音素に影響を受けてスペクトルは時間と共に連続的に変化する．これを調音結合という．そこで，信号系列を重み付き最小二乗法で直線近似した際の傾きとして定義される Δ ケプストラムを利用することで時間方向の動的な変化を捉えることが可能となる．Δ ケプストラムは差分に基づく特徴であるため，収録機器の伝達特性の変化に対して頑健で，音声認識や音声分析の場面で広く利用されている．

2.4　時間周波数分析によって得られる特徴

音声分析の手法として，ケプストラム分析以外に時間周波数分析が挙げられる．音声の特徴は，時間領域に現れる周期性の特徴（時間情報）と周波数領域に現れる調波性の特徴（周波数情報）に分けられるが，どちらか一方のみを利用した分析は，特徴の抽出が容易であるが他方を無視しているために起こる問題があった．例えば，基本周波数推定における雑音問題では，雑音が低周波数帯域に強い時には，音声波形が雑音の影響で歪み，音声波形の自己相関法のような時間情報のみを用いる手法では推定誤差が大きくなる．逆に，雑音が中 · 高周波数帯域に強い時には，音声スペクトルの調波構造が雑音の影響で歪み，ケプストラム法のような周波数情報のみを用いた手法では雑音の影響を大きく受けてしまう．実環境において人間がどんな雑音環境下でも音の高低の変化を知覚出来ることを考えれば，時間情報と周波数情報の両方を知覚していることは明らかである．そこで，時間情報と周波数情報の両方を統合した特徴を得るために時間周波数分析という分析手法が存在する．この分析手法は時間情報と周波数情報を両方利用する分析手法の総称で，その具体的な手段は多岐にわたる．主な例を挙げれば，時間方向と周波数方向の両方向にフィルタをかけて特徴量選択により必要な情報を抽出するもの [2] や，スペクトル平面の局所的な勾配に基づく特徴を抽出するもの [12] などがある．本研究では，前者の手法を用いて特徴抽出を行うが、その詳細は 4-5 章で述べる．

2.5　まとめ

本章では，本研究で利用するものを中心に音声言語処理の各場面でよく利用される特徴について述べた．

第3章

音声対話システムと従来の対話管理技術

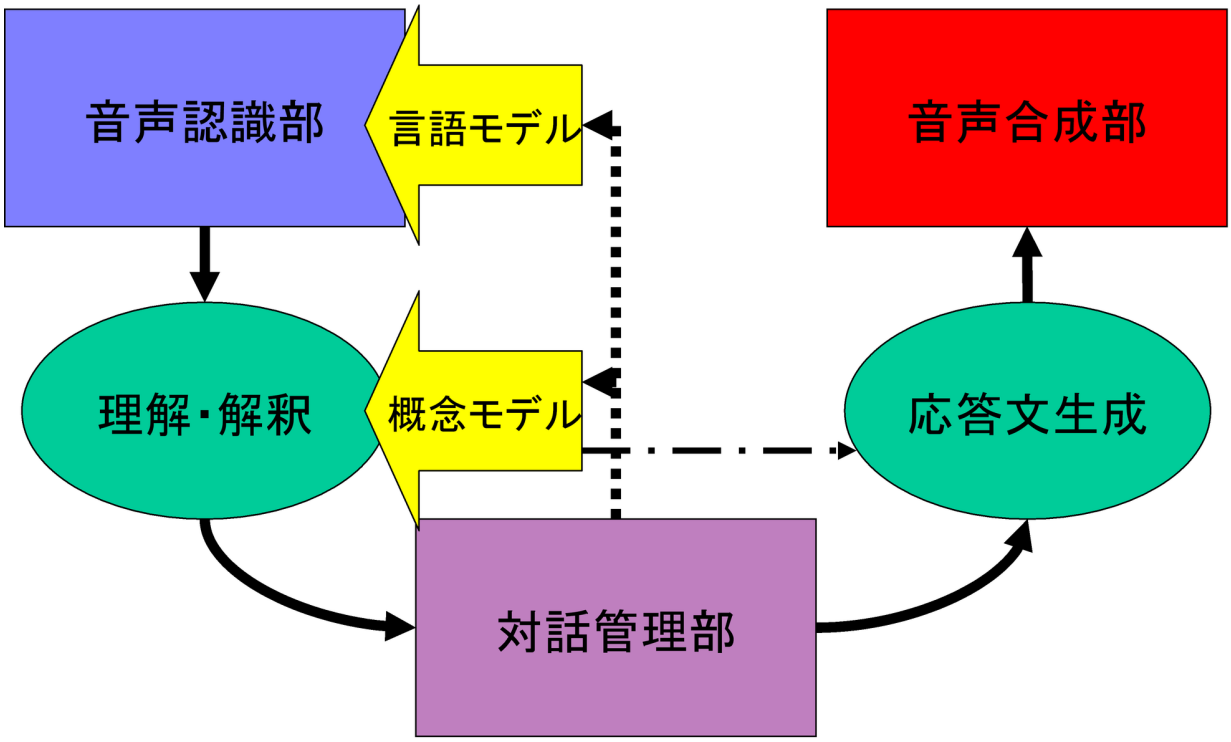

図 3.1: 一般的な音声対話システムの構成

3.1　はじめに

音声対話システムとは，音声をただ単に認識・するだけでなく，発話の意図を推定・理解して適切な応答をするという，複雑な処理をこなすシステムである．そのためには，音声認識・音声合成・対話管理などの技術の発展が不可欠であり，今もなおそれらの研究は各分野において進められている．

本章では，一般的な対話システムの構成と各部の詳細を簡単に述べた後，音声対話システムの要である対話管理技術について先行研究を例にとって説明する．

3.2　音声対話システム

音声対話システムの一般的な構成を図 3.1 に示す．音声対話システムは，大きく分けると 3 つの部分に分けて考えることが出来る．まず，ユーザの発話を入力として受理する音声認識部である．ここでは，ユーザの発話音声を音響特徴へ変換し，音響モデルや言語モデル・辞書ファイルなどと照合することで音声から文字情報への変換を行う．次の対話管理部では，ユーザの “発話内容” を理解し，それまでの対話履歴や，場合によっては事前に登録された，又は入力音声から推定しうる，話者の情報などを利用して，要求されている情報の検索などの処理を行う．そして，その結果を受けて，最後に音声合成部が応答発話を生成する．

音声認識部と対話管理部の間の情報のやり取りにおいて，入力音声の認識結果から対話処理に必要な情報へ変換する部分を特に言語理解部と言うことがある．同様に，対話管理部での処理結果から適切な応答文を生成し音声合成器に渡す部分を言語生成部と言うことがある．しかし，その境界は必ずしも定まってはおらず，研究の目的や対象によっては必ずしも 5 つに区分することが適当でない場合も存在する．本研究では，音声の非言語・パラ言語情報に着目し，対話システムの対話管理フェーズにおけるその活用を目標としているため，言語理解部・対話管理部・言語生成部までを連続して扱うこととなる．それゆえに，本稿ではこれら 3 つの部分をまとめて対話管理部とよぶこととする．

以下では，一般的な音声対話システムにおける音声認識部・音声合成部・対話管理部の各部での処理の詳細に触れた後，実用化されたものを中心に音声対話システムの例を挙げる．

3.2.1　音声認識部

音声認識は，一般的にその認識対象が孤立発声（単語）か連続発声（文章）か，語彙サイズが小規模か大規模かによって分類される．音声認識における最も簡単なタスクとは，数字や “はい／いいえ” などの単語認識であるが，このようにある程度語彙数が限られていて（数百語以下），認識対象が孤立発声であるような場合，隠れマルコフモデル（Hidden Marcov Model：HMM）を用いて比較的高い精度で認識することが可能である．HMM は時系列信号の確率モデルであり，複数の定常信号源の間を遷移することで非定常な時系列信号をモデル化する [13]．この HMM を音響特徴量系列に適用することで音響特徴をモデル化する．このとき状態遷移確率が発声の時間的な揺らぎを，出力スペクトルの確率がスペクトルの揺らぎをうまく表現していると言われる．図 3.2 に HMM の構造を示す．図 3.2 において，S_i は i 番目の状態を，a_{ij} が S_i から S_j への遷移確率を表している．各状態 S_i はベクトル x を出力する確率 $b_i(x)$ をパラメータとして持つ．HMM では，このパラメータ (a_i, b_i) を学習することでモデル化を行う．学習アルゴリズムには最尤推定が使われ，学習データから観測された音響特徴量の時系列データの尤度を最大化するパラメータを求める問題に帰着する．このようにして音素ごとに作成された HMM を連結することで単語の HMM を作ることが出来る．もちろん，単語発声から音響特徴を抽出し，単語 HMM を作ることも可能である．しかし，単語発声から単語 HMM を作成するよりも，音素 HMM の連結によって単語 HMM を作ったほうが，必要な HMM が少ないこと，未知語への対応が利くことなどの理由から音素 HMM から単語 HMM を作成することが多い．

地名検索のように，孤立発声が対象であっても語彙数が多い（数千語程度）場合は，単語認識とその結果の言語処理を完全に分離することができるためそんなに難しい問題とはならない．同様に，語彙サイズは小さいが，認識対象は連続発声という場合においても，文法ノードや語彙を性的に展開することが可能であり，最適解を見つけるのは比較的容易である．しかし，音声対話システムで入力されうる音声について考えてみると，ユーザは音声対話システムに登録されている語彙を考慮してしゃべってくれるわけではないので，同じ意味内容でも話者によって様々な言い方が存在するため，語彙サイズは大きくなる．また，認識対象となる音声は当然ながら連続発声である．このような大語彙連続音声認識においては，その探索空間は膨大なため，音響モデルと言語モデルが真に統合された認識機

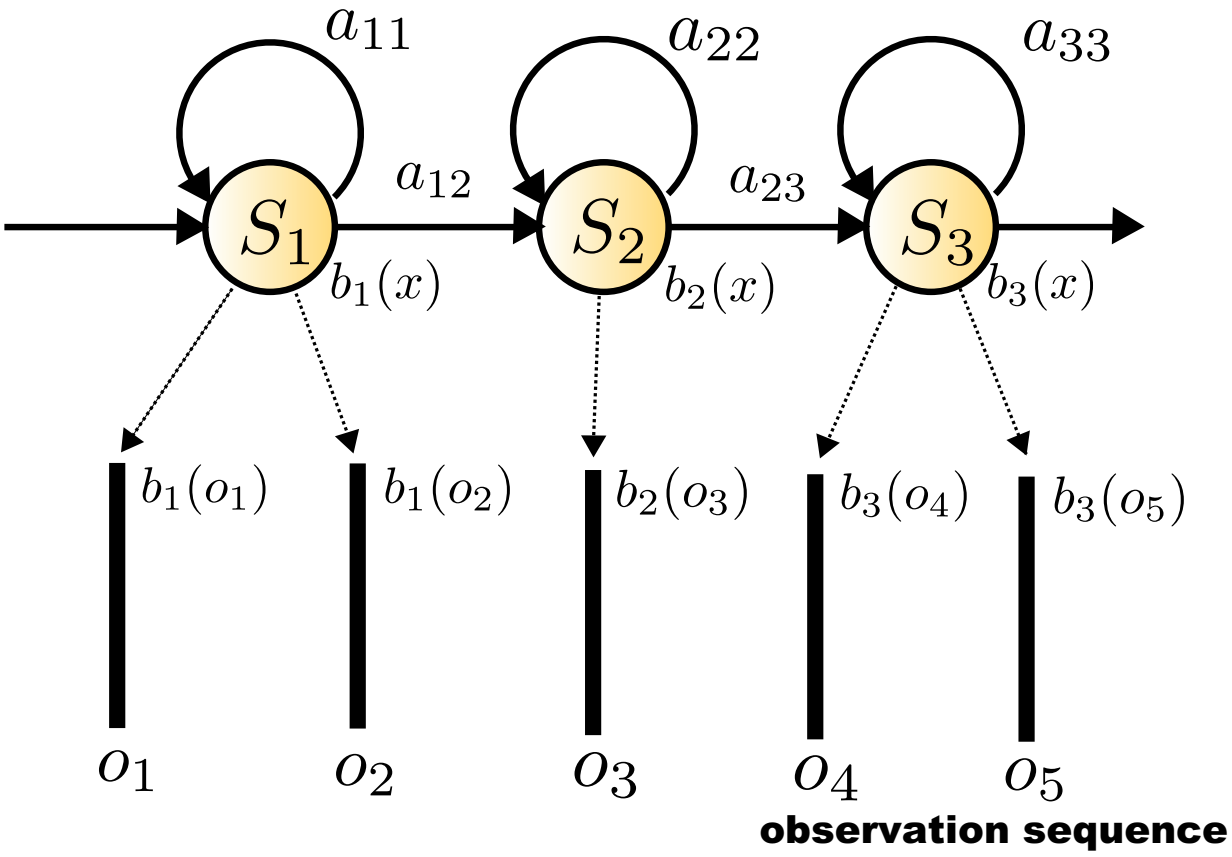

図 3.2: HMM

構が不可欠となり，複数の階層のモデルを組み合わせる場合も多い [14].

3.2.2　音声合成部

音声対話システムにおける音声合成は，対話管理部で生成された応答文から音声を合成する，テキスト音声合成（Text To Speech：TTS）が基本である．音声の合成方法には大雑把に分けて 2 種類ある．それが，調音音声合成方式と，信号処理的音声合成方式である．

前者は，人間の調音運動に着目し，そのメカニズムを近似するような方式である．基本的には声帯から声道にかけての調音器官の発声時の動きをモデル化·シミュレートする [15] ことによって近似を行うが，誉田ら [16] のように，実際にロボットに声帯·肺及び各調音器官を実装することにより，人間と同じように喉を震わせ口を動かして発話することの出来るロボットの研究などもある．調音音声合成は，身体障害者の発声訓練支援などの応用が検討されている．また，人間の音声生成過程はソースフィルタモデルでは近似できない非線形な要因を含んでいるため，このようなアプローチは人間の発声メカニズムの解明という点においても期待は大きい．

後者は，人間の発声メカニズムには触れず，実際の音声データを利用して，要求を満たす音声を実現しようという手法である．こちらの手法は，調音音声合成方式と対比して信号処理的音声合成方式という．信号処理的音声合成方式はさらに 2 種類に分けられる．

一つ目は波形接続方式といい，収録した人間の発声を切り貼りするように接続して出力する．電車の駅名アナウンスなどはこの方法を用いており，実際の人間の音声を接続部以外はそのまま使用しているため，合成音声の品質がよい反面，ドメインが限定されていないようなタスクにおいてはデータ量が大きくなる欠点がある．近年のストレージ増大傾向に後押しされて様々な研究がなされている [17, 18, 19].

二つ目はパラメータ編集方式という．こちらは，人間の発声を分析してパラメータに変換する．発話内容とパラメータの対応関係を音響モデルとしてモデル化することで，特徴量空間における柔軟な音声合成が可能である [20].

また，近年においては TTS のみならず，概念音声合成（Concept To Speech：CTS）[21] の研究も行われている．TTS がテキストを入力とするのに対し，CTS ではシステムの内部表現（概念）から直接音声を合成するため，文の生成過程で正確な言語情報が得られ，統語構造を韻律に反映させたり，談話情報で韻律の制御を行なうといったことが容易に行なえる [22]．統語構造や談話情報等の高次の言語情報，あるいは意図や感情等の非言語 · パラ言語情報は，音声の韻律と関連する点が多く，この観点からの研究が重要である．

3.2.3　対話管理部

対話管理部は，まさに音声対話システムの要であり，人間の言語機構における脳と同じ役目を担っている．図 3.1 からもわかるように，その機能は認識した音声からの発話者の意図の理解 · 対話管理 · 適切な応答文の生成の 3 つに大きく分けることができ，それぞれに焦点をあてた研究がこれまでにも多くなされてきた．そのような先行研究を例にとり，対話管理部での処理を次節で説明する．

3.3　様々な対話管理技術

3.3.1　音声認識結果の理解 · 解釈

前節で述べたように，音声対話システムにおける音声認識は大語彙連続が前提である．しかし，自然発話では人間が逸脱と認めることの出来る絶対的に不適切な表現も普遍的に出現するため，その全てを網羅した文法を記述することは現実的に不可能である．又，もし可能だったとしても，処理に無駄がありすぎて実時間処理を前提とするシステムには不向きである．したがって，一般的には限定されたタスクドメインを仮定し，語彙数や文法のサイズを小さくする．それでも不適格な発話は存在し，それに対する頑健な言語理解に関してもたくさんの研究がなされてきた [23].

例えば，文法解析の際に解析の途中で得られた句や節を保存しておき，全体の解析に失敗したときに途中保存した断片を用いて意味の推定を行う部分解析と呼ばれる手法 [24] であるとか，一度通常の文法で解析を行い，解析が失敗した場合に制約の緩い文法に切り替えて解析を行う制約緩和 [25] などがある．これらは主に音声認識の結果（テキスト）に対する自然言語処理であり，対して，音声の非言語 · パラ言語情報に着目し，発話者の意図の理解を言語情報の認識とは独立に行う研究もある [2, 3, 26]．但し，ここでいう発話者の

意図とは，発話に込められた感情や，肯定／否定という狭義での意図のことを指し，前に述べた自然言語処理による意味理解とは異なるということに注意しなければならない．後者の非言語 · パラ言語情報を用いた発話者の意図の理解に関しては 4 章で詳しく述べる．

3.3.2　対話管理

対話管理の主眼とも言うべきが対話戦略の組み立てであるが，その大きなカテゴリの一つとして音声認識のリスクマネジメントが挙げられる．すなわち，大語彙連続音声認識において不可避である音声認識誤りをどのように処理するかという問題である．この問題は，さらに 2 つの問題に分けることが出来る．一つ目は，どのようにして音声認識結果が誤りであるかを判断するかということである．二つ目は，音声認識結果が誤りであると推定できた場合にどのような処理を行うかということである．

まず一つ目に関してだが，コンピュータによる音声認識は，入力された音声に対して最も尤度の高い単語列を出力するという過程であるため，正しい認識結果と認識誤りを識別するためには何らかの尺度が必要である．一般に発話検証では，小語彙では音節モデル · 競合音素モデルと比較することが有効であるが，大語彙では他の候補（N-best）と比較することが有効であると知られている [27]．そこで，駒谷ら [28] は，音声認識結果の内容語が正しく認識されたかどうかを推定するために，信頼度という指標を提案している．内容語 w に対する信頼度 p_w は 3.1 式で表される．

$$
\begin{aligned}
p_w &= \sum_{i=1}^{N} p_i \cdot \delta_{w,i} \qquad (3.1)\\
p_i &= \frac{e^{\alpha \cdot score_i}}{\sum_{j=1}^{N} \alpha \cdot score_j}\\
\delta_{w,i} &= \begin{cases} 1 & (w \in \text{Sentence No.}i) \\ 0 & (w \notin \text{Sentence No.}i) \end{cases}\\
\alpha &: \text{Smoothing Coefficient.}\\
N &: \text{N-best.}\\
score &: \text{Likelihood of Recognition.}
\end{aligned}
$$

これは，重み付きの事後確率に相当する．この信頼度を尺度として適切なしきい値を設定することで，音声認識結果の第一候補をそのまま受理する従来手法と比較して，誤り率が減少したことを確認した．

このようにして，音声認識誤りを検出することが出来たとき，どのような戦略をもってユーザに意図を確認するかという問題が二つ目である．堂坂ら [29] は，この問題に対して，ユーザ発話の全文を確認するのではなく，システムが持つタスクに関する知識を用いて確認を行うドメインを限定するという方法を提案している．図 3.3 に堂坂らが提案するシステムを利用した場合の対話例を示す．図中，下線が引いてある部分が，このシステムにおけ

前提：システムは気象情報案内システム．現在国中のどこにも警報は発せられていない，または，警報が発せられている場所は少数であるという情報をシステムは保持している．

· 従来の対話戦略：
ユーザ：東京都に大雨警報は発表されていますか？
システム：質問は，東京都 に 大雨警報 が発せられているかどうか，でよろしいですか？

· 堂坂らの対話戦略：
ユーザ：東京都に大雨警報は発表されていますか？
システム：質問は，警報 について，でよろしいですか？

図 3.3: システム知識制限下における効率的な対話戦略

るドメインである．この例の場合，ユーザが関心のある場所が東京都であることや警報の種別が大雨であることは，確認する必要がない．たとえ確認を行ったとしてもシステムがユーザに提供できる情報はほとんど変わらないからである．逆に，システムが認識している “東京都” や “大雨” といったキーワードも，認識誤りをしている可能性があるので，それらの項目を確認することでユーザの訂正発話を招き，対話が不必要に長くなってしまう恐れがある．堂坂らは，ドメインに関わるキーワードに対して確認コスト（確認の際にかかるコスト：確認発話に含まれるキーワードの数で定義）と情報伝達コスト（対話の結果ユーザに情報を伝達する際にかかるコスト：確認コストと同様に発話に含まれるキーワードの延べ数で定義）という二種類のコストを導入し，2 つのコストの合計が一番小さくなるように応答発話を生成することで，音声認識誤りに対処できるとしている．

また，単純には，そもそも認識誤りが起きないように全ての対話をシステム主導にしてしまうという考え方がある．この場合，ユーザは最小限のキーワードと Yes/No のみを答えればよく，システムに実装すべき語彙サイズも小さくてすむため，認識誤りは起きにくい．しかし，これはかなり単純なタスクにしか適用できない方法であることは明らかである．一般の検索タスク等においては，ユーザにとって必要な情報はユーザごとに異なるため，常にシステム主導で対話を遂行するやり方は効率性に欠ける．したがって，ある程度自由度のあるタスクで，効率的かつ頑健な音声対話システムを構築するには，ユーザに自由な発話を許しながらも，必要なときにはユーザへの質問やユーザの誘導を行う混合主導対話 [30, 31, 32] が望ましい．

このように，対話戦略において効率は大変重要なファクタであり，音声認識誤りのリスクマネジメントと同様に，対話管理研究において大きなカテゴリの一つとなっている．フライト検索（ATIS）[33, 34] や，列車時刻案内（RAILTEL）[30, 35] などに代表される情報検索型の音声対話システムでは，関係データベースのフィールド名やその値の集合から明

示的に検索に必要なキーワード集合を定義できるので，それらのキーワードをユーザの発話からスポッティングすることで効率よくユーザの意図の確認ができる．しかし，近年研究が進んでいる Web ページなどの大規模知識ベースを検索するようなタスク [36, 37, 38] においては，キーワードだけで，タスクが達成できるようなキーワードの集合を明確に定義することは不可能であり，キーワードスポッティングではなく，音声認識結果全体を自然言語文として解釈する必要がある [39]．しかし自然発話の場合，音声認識誤りだけでなく，フィラーや話者性による多様な言い回しなどの冗長性が多いため，認識結果そのもの全てが情報検索に有用とは限らない．そのため，音声認識誤りに対しては適切に確認/棄却を行い，検索に有用でない部分に関してはその部分を取り除いたり，逐一確認を行わないことが望ましい．音声認識誤りを回避する方法に関しては，この話題の本筋ではなく，また，前に述べたこともあり割愛する．また，前に述べた堂坂らの手法 [29] は認識の段階でキーワードが確定されており，スポッティングが可能であることが前提であったため，この話題においては利用できないことを確認しておく．認識対象が自然発話であるような場合に，検索に必要なキーワードをどのように抽出し，他を棄却するかという問題に対して，翠ら [40] は検索対象の知識ベースのみから求められる統計量と，音声認識の N-best 候補に対する検索結果を用いて，音声認識結果の各文節が検索に有用かどうかを判定している．具体的には，音声認識結果の N-best 候補を用いて検索を行い，検索結果の違いを検索重要度と定義する．すなわち，検索重要度が大きい場合，検索結果の違いを与えている語句が検索のキーワードの一つであると考えられるため，その語句についてユーザに確認をとることで検索結果を絞り込むことが出来る．また，検索に決定的な影響を与えるような語句に関しては，前もって確認を行う．そのような語句の検出には認識の際と検索の際に用いる 2 種類の言語モデルから得られる単語パープレキシティを利用して得られる検索整合度をいう尺度を利用する．このような 2 重の確認を行うことで，ユーザへの確認回数が減らすことが出来る．

最後に，利用者に対する親和性（ユーザフレンドリネス）向上に関する研究を紹介する．これは，その効果を直接的に数値で評価することが難しい問題であるが，実際にユーザが使うことを考えたとき，よりユーザにとって使いやすい·親しみやすいシステムであるということは大変重要なことである．本節で前に述べた音声認識誤りに対するリスクマネジメントなども，ユーザにとって使いやすいシステムにするための取り組みのひとつであるが，ここでは，人間同士の対話のように，ユーザの発話に対して適切に相槌を打ったり，システムの発話中にユーザが割り込んで発話を行った際にも動的に対処する，など，人間同士のような自然な対話を実現するための手法について述べる．

人間とシステムが対話を行う際に，システムが人間同士の対話と同じように相槌や割り込みなどの応答をすることができたら，より円滑に対話が進行するものと考えられる．人間同士での対話において適切なタイミングで挿入される相槌や割り込みは，対話の潤滑剤の役目を果たしている．特に，指向性のない雑談においてその傾向は顕著であり，竹内らは人間同士の雑談を分析することでシステムが相槌や割り込みを行うべきタイミングについて検討を行っている [41]．発話句末 100[ms]（およそ 1 モーラ分）の基本周波数と波形パワー，及び発話句末の終端単語などが決定木学習の素性として用いられており，評価実験

において実際の対話と同程度の自然性を実現した．また，河原らは対話における自然な“間”を音声対話システムで実現するために，漫才や落語の音声を分析し，発話者の意図（肯定／否定，同意／不同意）と発話タイミングとの関連を調査している [42]．人間同士の対話における発話のタイミングを学習することがシステムと人間との対話において常に有効であるかどうかは自明ではないが，少なくともシステムを人間と同じように捉え対話する傾向のあるユーザに対しては，一定の効果があるようである．

ユーザに対する親しみやすさを向上させる方法は他にもある．人間は対話をする際，相手の情報を得てそれを応答に活かしている．相手の情報とは，相手の年齢や性別などの静的な情報から，発話時の感情などの動的な情報まで様々であるが，これらをシステムが認識して応答に活かすことは，より人間同士の対話に近づくこととなり，ユーザにとってより使いやすいシステムとなると考えられる．峯松らは，音響特徴を分析することで主観的高齢者を同定し，なおかつ，同手法に基づいて主観的年齢推定について検討を行っている [43]．峯松らは，まず高齢化の音声スペクトルへの影響を考慮し，話者同定技術の応用で，スペクトル情報のモデル化による主観的高齢者音声の同定を行った．これにより 9 割を超える同定が実現されているが，データを替えても同じように同定誤りが起こることからスペクトル情報のほかに話速や波形パワー · 基本周波数パターンなどの韻律的特徴も同定に利用し，その有効性を確認している．同様に，篠田らはユーザの発話から対話のトピックに関する知識レベルを推定し，動的に応答を変化させるシステムを構築している [44]．知識レベルに応じて提供する情報の量や質を変えるだけでなく，知識レベルの高いユーザに対してはユーザ主導の対話に切り替えるなど，対話の主導権を切り替えることで柔軟な対応を実現している．

3.3.3　適切な応答生成

対話管理技術において，応答生成に主眼をおいた研究は少ない．それは，音声対話システム自体がそうである場合が多いが，評価が利用者による主観評価に頼らざるを得ないことが理由の一つとして挙げられる．しかし，多くの音声対話システムに実装されているテキストベースの応答生成器には，改善すべき点が多い．例えば，情報検索タスクを目的とするようなシステムの場合，多くは応答文のテンプレートを用意しておき，そこに検索結果となる単語を挿入することで応答文を生成する．しかし，この手法の場合，応答文のテンプレートが想定されている単語群にしか使えないことが多く，対話ドメインや利用する単語を拡張した場合，テンプレートも増やさなくてはならず，汎用性に欠ける．そこで，八木らは従来のテンプレート式応答文生成法の拡張と韻律的特徴を合成時に導入することでより柔軟な応答生成の可能な音声対話システムを構築した [45]．まず，従来は文そのものに対しテンプレートが作られていたのに対し，これをフレーズごとに変更する．これにより，語順の変化や修飾語の挿入，文末表現の変化に対応が可能となる．また，文節や単語の連結によりアクセント規則が変わる単語が存在することから，それらの韻律制御規則を導入することにより，より自然な合成音声の生成が可能となった．後者の韻律制御規則の導入については，越智らが基本周波数パターン生成過程モデルに基づく焦点制御の研究を行っている [46]．人間同士の対話においても，強調したいフレーズに対して，単語のアク

セントとは別にフレーズ全体に強勢をおくような話し方をする場合がある．音声対話システムにおいても，システムがユーザに対して一番伝えたいフレーズを強調することで，他の部分の韻律や統語情報が多少不完全であってもユーザが最も知りたかった情報は伝えられる可能性が高まると考えられる．システムの発話の焦点を制御するためには，強調したいフレーズの音調を制御することが必要であり，それは基本周波数パターン生成過程モデルのパラメータを制御することでほぼ実現できる．

3.4　本研究の位置づけ

前節で，多様な対話管理技術の一部を述べた．近年音声の非言語 · パラ言語情報に着目し，対話管理に活かそうという研究が増えてきてはいるが，多くは音声認識から言語理解 · 解釈のフェーズまでで終わっているものが多い．そこで，本研究では主に発話者の意図を識別し，その結果を応答生成に活かす手法について検討する．

3.5　まとめ

本章では，一般的な音声対話システムの構成と各部の働きを簡単に述べた．また，音声対話システムの肝である対話管理技術について先行研究を例に説明した．最後に，本章を振り返り，本研究の位置づけについて明確にした．

第4章

発話に込められた意図・感情の識別

4.1　はじめに

前章で，音声対話システムの基礎理論と，音声対話システムの要である対話管理技術について先行研究を例に紹介した．先行研究の例からみえてくることは，音声認識誤りに対するリスクマネジメントや対話進行の効率化という観点では，音声認識結果のテキストに対する自然言語処理によってある程度の成果を挙げられるが，ユーザに対する親和性を実現するためには音声の非言語 · パラ言語情報に着目した処理が必要不可欠であるということである．ここで改めて音声をシステムの入出力に利用する利点の最たるものは，直観的にその操作方法が理解できること，であり，やはり，ユーザにとって使いやすいシステムを設計することが重要であることがわかる．本研究では，そんなヒューマンフレンドリーなシステムの構築のために，音声の非言語 · パラ言語情報を話者の意図の識別に利用することで，音声認識誤りに対して頑健な意図識別を目標とし，かつ意図識別の際に得られた情報を利用して，システムの応答をよりユーザの心的状態を汲んだものに動的に変更することを目指した．本章では，前章でも触れた対話戦略の中でも，特に音声の非言語 · パラ言語情報を利用して話者の意図を識別する技術に関して主に利用している特徴量ベクトル別に先行研究を例に説明する．なお，話者の意図の識別の研究に関連して話者の感情の識別の研究も意図識別の研究以上に盛んに行われている．利用する特徴量や識別のプロセスが非常に似ていることから，感情識別の研究に関しても触れる．

4.2　韻律的特徴量を利用した識別

本節では特に韻律的特徴を主に識別に利用している研究について述べる．韻律的特徴を意図や感情の識別に用いる場合，一般的に次のような手順で行う．

1. 識別に利用する物理量を各音声から抽出する．
2. 抽出された物理量から最大値や偏差などの各項目を静的に計測し，特徴量ベクトルを作成する．
3. 各音声から得られた特徴量ベクトルを用いて機械学習によってモデルを構築する．
4. 構築されたモデルを用いて識別を行う．

藤江らは，上記の流れに沿って発話者の意図の識別を行う音声対話システム:ROBISUKE[47]を開発した．ROBISUKE は，ユーザの意志決定支援型（ユーザがある事柄に関して何か決定しなければならないとき，その決定を支援する）の対話システムであり，具体的な対話タスクとしてはレストラン決定支援タスク（昼食をどこに食べに行けばよいのかについて相談に乗るタスク）を取り上げている．入力として画像と音声の両方を受け付けるマルチモーダルシステムであり，画像処理によって人間の首の動き（“かしげ”や “うなづき”）を認識する．本節では音声入力を受け付けてユーザの反応が肯定的態度か否定的態度かを識別する 2 値の発話態度識別について述べる．識別までのプロセスの概略は前に示した通りである．識別に関わる詳細な条件を表 4.1，表 4.2 にまとめた．特徴量ベクトルに利用した項目に関しては，肯定的発話と否定的発話で次のような違いが現れるとしている．

表 4.1: 意思決定支援型音声対話システム ROBISUKE の特徴抽出条件

音声サンプル	男性 20 人分：計 2000 発話．発話の内容は表 4.2 参照.
特徴量ベクトルに利用した項目	x_1：基本周波数パターンにおける第 1 モーラの母音部分の傾き
	x_2：発話全体における基本周波数のレンジ
	x_3：最終モーラの継続長
モデル	混合正規分布（Gaussian Mixture Model；GMM）
識別手法	ベイズ識別

表 4.2: ROBISUKE でモデルの学習 · 識別に利用された音声の収録内容

カテゴリ／店	ハンバーガー，ラーメン，弁当，カレー
	学食，マクドナルド，味源，夢民，ホカ弁，そばの実
言い回し	○○か，○○ね，○○，いいんじゃない，そうだね
	※ “○○”はシステムの提案に含まれるカテゴリ／店の復唱.
態度	肯定的／否定的

x_1： 肯定的な場合に正になり，否定的な場合に負になる傾向がある．

x_1： 肯定的な場合に否定的な時よりも大きくなる傾向がある．

x_1： 肯定的な場合に否定的な時よりも小さくなる傾向がある．

識別の際には，収録した 2000 人分のデータを 4 分割（1 セット 5 人分，500 発話）し，交差検定を行った．結果は混合数 16 の際に 82.9% と最もよい結果が得られ，人間の知覚実験を行ったところ，ほぼ同程度の識別が可能であることが示された．藤江らは，後にカテゴリに “中立”を加えて 3 値の識別も行い，なおかつ識別誤りに対するリスク回避の方法も提案している [48].

McGilloway ら [49] や，Dellaert ら [50] は，韻律的特徴を用いて感情の識別を行った．どちらも，喜び · 悲しみ · 怒り · 恐怖の 4 つの感情に，“平静”を加えた 5 つのカテゴリ区分で識別を行い，McGilloway らが 55%，Dellaert らが 79.5%を実現している．それぞれの実験環境は表 4.3，表 4.4 の通り． どちらの研究も，識別後に分析を行って各特徴の寄与を調べている．その結果，McGilloway らの実験では，識別への寄与が最も高かったのは Number of F0 points recoverd，つまり，基本周波数が抽出できた区間の時間長の和であった．音節数による正規化を行っているため，無声音が少ないほうが識別率が高いということになる．他にも，Inter-quartile range for intentnsities at minima：波形パワーの下方 4 分の 1 区間のレンジの最小値や，Median of silence durations：無音区間長の中間値などの寄与が大きいことがわかった．一方，Dellaert らの実験では，識別への寄与が高かったのは，基本周波数の最大値や中間値，及び，基本周波数が正に増加している区間の増加率の平均値などであった.

最後に，識別器のことについて触れる．McGilloway らの実験では，サポートベクターマシン（Sapport Vector Machine; SVM）[51]，生成的ベクトル量子化（Generative Vector

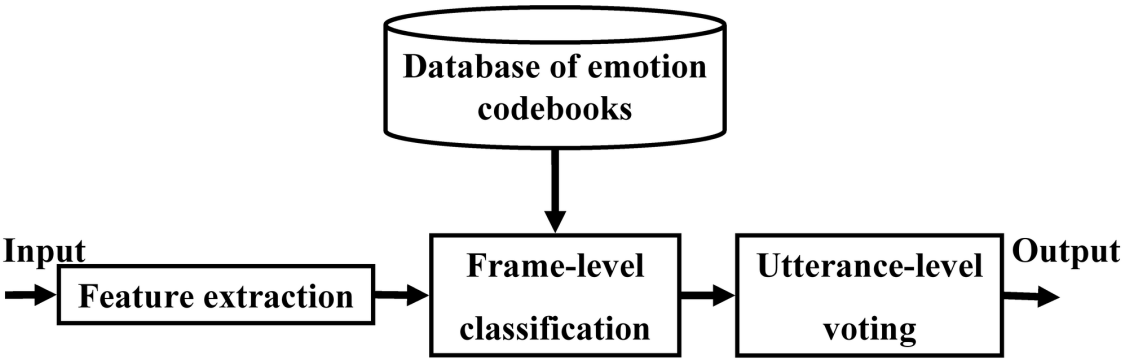

図 4.1: 佐藤らによる特徴抽出のフローチャート

Quantisation; GVQ）[52]，及び線形識別（Lenear Discriminants）の 3 種類の識別手法による実験を行い，最も精度のよかった線形識別を採用している．また，Dellaert らの実験でも同様に KNN 以外にも 2 種類の識別手法で試行し，最も識別率のよかった KNN を利用している．実験条件が同一でないためどの識別器を使うのが最もよいのかということは，上記の研究からはわからない．しかし，識別器の違いは，特徴量の違いに比べて識別精度に与える影響は小さく，あまり本質的な問題ではないという意見もある [3].

4.3　MFCC を利用した識別

MFCC は，音声認識の分野では早くから音素の動的特徴を捉えるために利用されていたが，感情認識の分野でその有効性が認められたのは，近年になってからである．それは，MFCC が主として音韻的特徴を抽出するために利用されるものであり，感情認識に有効な情報は音韻的特徴よりも韻律的特徴に含まれていると考えられてきたからであった．しかし，音韻と韻律は，少なくとも日本語において，互いに独立な特徴ではありえず，それは語感という形で話者の意識の外で音韻が韻律に影響を与えていることからも明らかである．したがって，一般的な音声対話システムのように，話者やテキストが限定されないような状況で音声認識を行う場合には，MFCC から得られる音韻情報は貴重である．また，韻律的特徴のみで識別を行う場合，相互独立な物理量が数少なく，特徴次元が小さくなってしまう．佐藤らは，このことが識別の精度に影響を与えていると考え，MFCC のみを用いた感情識別手法を提案している [3]．佐藤ら以前にも，MFCC を感情識別に応用している研究はあったが，発話単位で統計量を求めており，フレーム単位での分析を行っていないもの [53, 54]，逆に，フレーム単位の統計量は識別に利用しているが発話単位での分析が足りないもの [55] など，フレーム単位での分析と発話単位での分析を両方行っているものはなかった．図 4.1 に佐藤らの特徴量抽出のプロセスを示す．まず，識別対象である発話文を，フレームの集合と捉え，フレーム単位で感情識別を行う．次に，フレームでの識別結果による多数決によって発話単位での感情識別を行う．図 4.2- 4.3 に，識別手法のイメージを示す．発話単位の識別フェーズにおいて，識別誤りが起こらないように，しきい値 L を導入

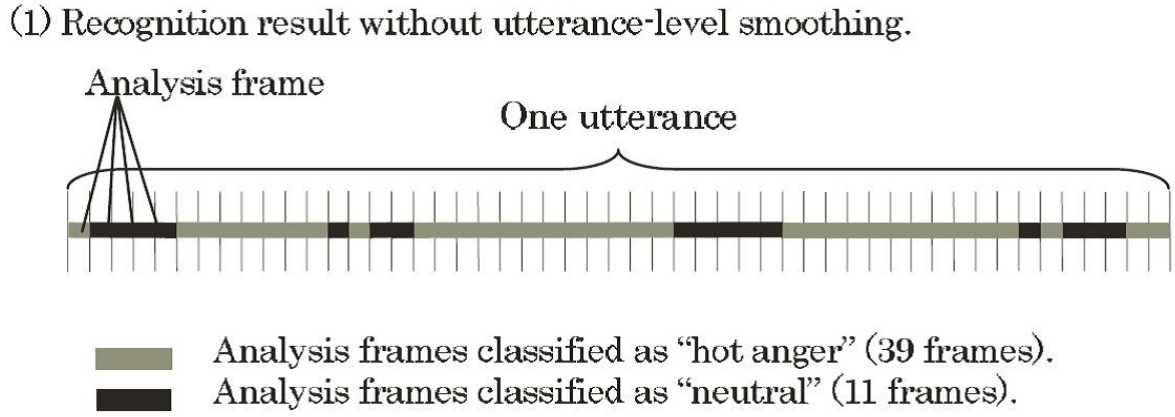

図 4.2: フレーム単位の識別結果の集合としての発話単位の識別結果

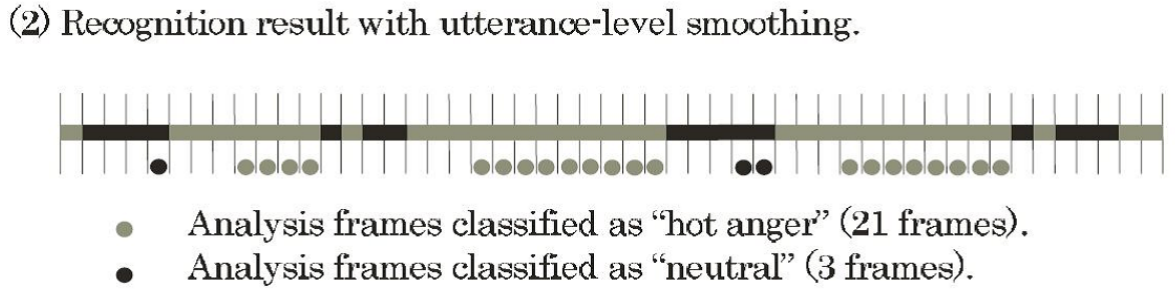

図 4.3: 発話単位の識別の際のスムージング

する．同じ識別結果を持つ連続したフレーム群を一つの集合として考え，長さ L 以下の集合に関しては識別結果を無視することで，識別結果を明確にしている（図では L=4）．この手法を用いて，“怒り (hot anger)”·“平静”·“悲しみ”·“喜び”の 4 感情の識別を行った結果，韻律特徴量のみによる手法と比べて約 16%の精度向上が実現された．

4.4　ST 特徴を用いた識別

MFCC と並んで，近年注目されているのが ST 特徴（時間周波数特徴）である．この手法は，人間の音声知覚過程をできるだけ忠実に再現するという，音声認識の根本的なコンセプトに沿った方法である．人間の音声知覚機構には中心周波数が連続的に変化するような帯域フィルタが存在する．これを聴覚フィルタという．聴覚フィルタの存在は Fletcher によって示唆され，Zwicker によって実験的に証明された [56]．聴覚フィルタは以下のような特性を持つ．

- 信号音に一番近い中心周波数をもつ帯域フィルタが信号音の周波数分析を行う
- 信号音のマスキングに影響する雑音成分はこの帯域フィルタ内の周波数成分に限られる

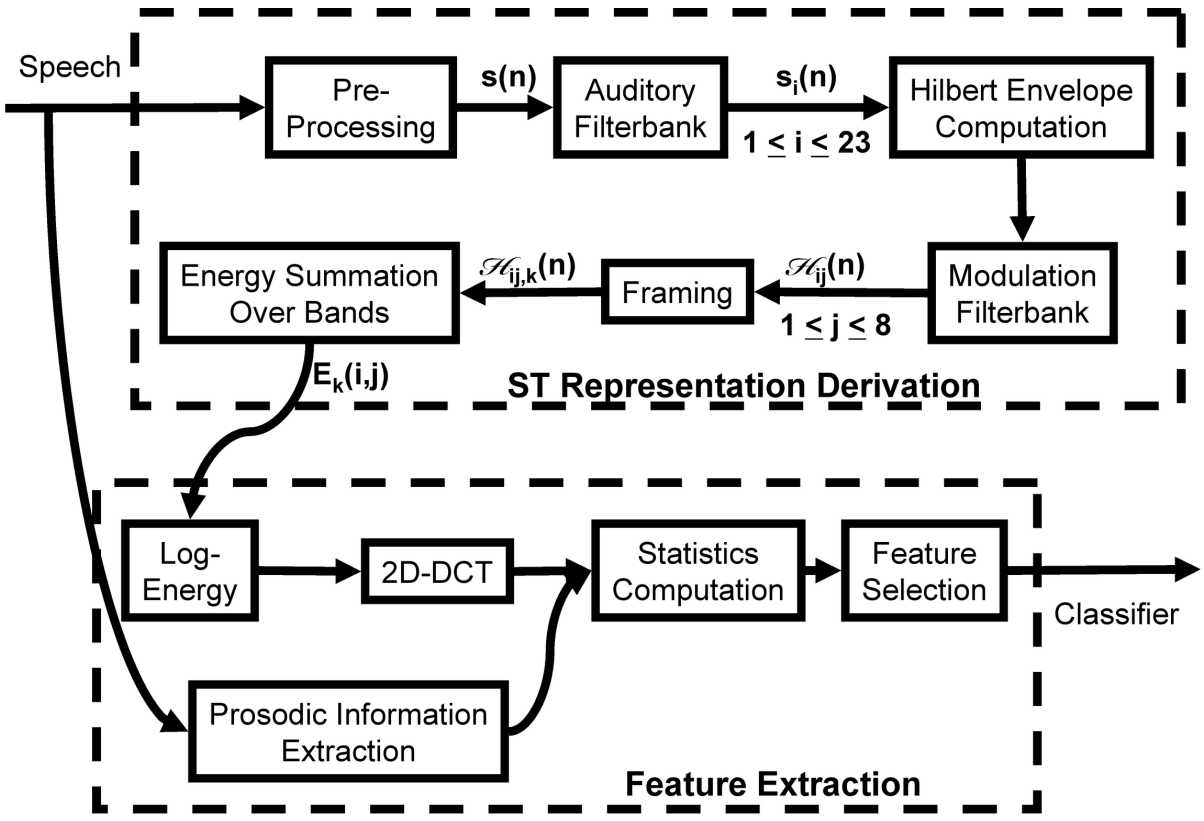

図 4.4: ST 特徴抽出のフローチャート

この帯域フィルタのバンド幅を臨界帯域という．つまり，この臨界帯域をバンド幅とするようなフィルタバンクは聴覚フィルタをよく近似するということになる．しかし，Zwicker の臨界帯域の測定法に誤差が生じる可能性が発見されたため，新しく臨界帯域に対応するものとして等価方形幅（Equivalent Rectanglular Bandwidth；ERB）が定義された．この ERB は，Zwicker らの測定した臨界帯域よりも纓牛内の基底膜の特性により近いものであり，これを基に聴覚フィルタのモデルである Gammatone Filter も提案された．ST 特徴は，Gammatone Filter のような聴覚フィルタをはじめ，Modulation Filterbank などのフィルタバンクを用いて感情発声に影響を及ぼす特徴を音声資料から抽出する手段である．

近年，哺乳類の大脳に時間周波数特徴を認識する受容野が存在し，人間の脳は数百ミリ秒単位の時間周波数特徴を知覚することができるという発見があった．Siqing ら [2] はこれを受け，長時間周波数分析を行うことで，感情識別の精度向上を図った．時間周波数特徴量の抽出方法を図 4.4 に示す．図中，破線四角内が特に ST 特徴の抽出プロセスとなっている．聴覚フィルタには，臨界帯域の 23 次元 Gammatone フィルタバンクを用いており，その出力波 23 次元に対して包絡線を計算し，それぞれに 8 次元の Modulation フィルタバンクをかけると，23×8 の特徴行列が得られる．これは，行方向と列方向がそれぞれ聴覚フィルタ成分と Modulation フィルタバンク成分に対応している．この行列に窓かけをするわけだが，この窓長を Siqing らは 250[ms] と比較的長く設定した．この 250[ms] という値は，生

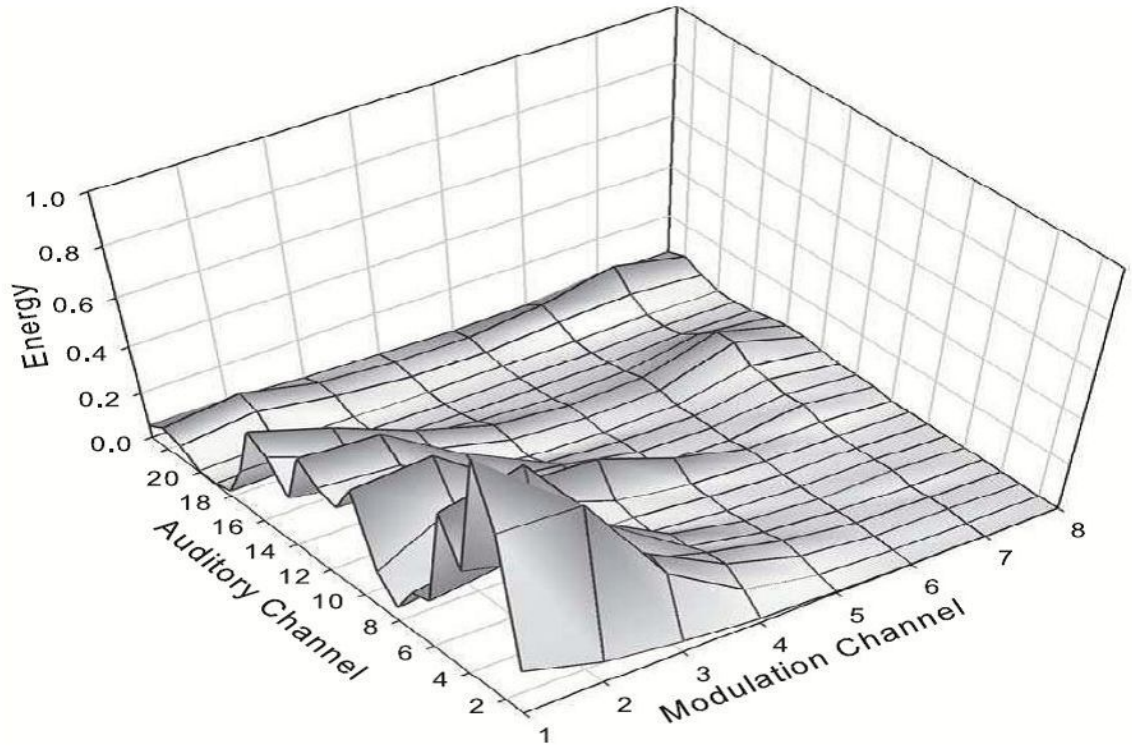

図 4.5: 変調エネルギー（感情ラベル：“平静”）

理学的研究によって事前に求められたもので，これによって従来の短時間周波数分析では得られなかった情報が得られることになる．最後に，そのパワーをとって，23 × 8 の変調エネルギー行列が得られる (図 4.5)．Siqing らはこの変調エネルギー行列に 2 次元離散コサイン変換 (2D-DCT) をかけて次元圧縮し，時間方向の分散が最も大きかった 25 個の DCT 係数を選択した．それぞれの係数ベクトルから最小値 · 平均値 · 標準偏差など 8 種類の統計量を計算し，最終的に 200 個 (25 × 8) の特徴量を時間周波数分析から得ている．これらの特徴量を韻律的特徴量に交えて感情識別を行ったところ，韻律的特徴量のみによる識別と比較して，44% もの誤答減少を実現した．なお，音声資料に用いたのは Berlin emotional speech database である．これは，男性 5 名，女性 5 名が 10 種類の文章を 7 つの感情で演じ分けたものである．話者は全て母語 (ドイツ語) 話者であり，7 つの感情とは，“怒り”·“退屈”·“落胆”·“恐れ”·“喜び”· “平静”·“悲しみ”である．なお，感情境界の学習・識別には SVC(Support Vector Classifier) を用いている．

Gammatone フィルタと同じく人間の聴覚特性をよく近似するといわれているフィルタに，ほかに Gabor フィルタなどが挙げられる [57]．Gabor フィルタは音声認識の分野ではすでに利用されており，対話システムのような high-end システムにおいても頑健な認識を実現できる [58]．感情識別の分野では，主に顔の表情に対して利用され，音声に対して利用されている例はまだ少ない．

4.5　まとめ

本章では，本研究で用いる特徴と同じ特徴量を用いて話者の意図や感情の識別を行っている先行研究を例に，意図 · 感情識別の実験の傾向と枠組みについて述べた．

表 4.3: McGilloway らの実験条件

音声サンプル	40 名（男性 20 名，女性 20 名）の意図的発声．計 197 発声
特徴量ベクトルに利用した項目	
Measures relating to tunes	
	tune duration，fit of tune to a quadratic function
	no of inflections in F0 contour per tune
Spectral	
	Energy below 250[Hz]
Intensity contour (excluding pauses)	
	Mean intensity，Median intensity，
	Inter-quartile range of intensity distribution
Intensity at local extrema in the intensity contour	
	Mean at maxima，Inter-quartile range for intensities at maxima
	Mean at minima
	Inter-quartile range for intensities at minima
Magnitude of rises or falls in the intensity contour	
	Inter-quartile range for magnitudes of rises
	Inter-quartile range for magnitudes of falls
Pitch of points in the F0 contour	
	Number of contributing observations，Mean，Inter-quartile range
Pitch at local extrema in the F0 contour	
	Inter-quartile range for pitch at maxima
	Inter-quartile range for pitch at minima
	Inter-quartile range for pitch at all local extrema
Magnitude of rises in the F0 contour	
	Median，Inter-quartile range
Durations of rises and falls in the intensity contour	
	Median duration for rises，Median duration for falls
Durations of level sections in the intensity contour (‘ plateaux ’)	
	Inter-quartile range for plateaux at intensity peaks
	Upper limit (90%) of range for plateaux at intensity
peaks	
	Median for plateaux at intensity minima
	Inter-quartile range for plateaux at intensity minima
Durations of features in the F0 contour	
	Median of silence durations，
	Inter-quartile range for durations of silences
	Median duration of falls，Median duration of plateaux at F0 maxima
	Inter-quartile range for duration of plateaux at F0 maxima
以上，計 32 次元.	
識別手法	線形識別

表 4.4: Dellaert らの実験条件

音声サンプル	複数話者による意図的発声．計 1000 発声以上.
特徴量ベクトルに利用した項目	
Statistics related rhythm：	
	speaking rate，average length between voiced regions,
	slope of maxima,
	number of maxima ／ numberof (minima+maxima)
	number of upslopes ／ number of slopes
Statistics on the smoothed pitch：	
	min，max，median，standard deviation
Statistics on the derivative of the smoothed pitch：	
	min，max，median，standard deviation
Statistics over the individual voiced parts：	
	mean min，mean max
Statistics over the individual slopes：	
	mean positive derivative,
	mean negative derivative,
以上，計 17 次元.	
識別手法	K-nearest neighbors; KNN

第5章

意図の識別実験

5.1　はじめに

本章では，2·4 章で述べた音響特徴を用いて話者の意図の識別実験を行い，識別精度とシステムに応用する上での扱いやすさの観点から話者の意図識別に有効な音響特徴に関して考察を行った．また，話者の意図のクラス分けを，肯定／否定の 2 値と，強肯定／弱肯定／曖昧／弱否定／強否定の 5 値で 2 種類識別実験を行い，その結果から，クラス拡張の可能性について検討した．

なお、本研究ではタスク指向の音声対話システムとして，ユーザに対して推薦を行うシステムを想定した．システムの推薦に対して，ユーザの応答が肯定的か否定的か，言語情報からは判断しにくいような場合に，ユーザの発話意図を識別することを目標としている．このようなシステムを想定したのは，他に多く見られる情報検索型の対話システムよりも，発話意図の識別の必要性が高いと考えられるためである [47]．特に日本人に多くみられる傾向ではあるが，あるトピックに対して提案を受けた際に，自分の意見を即座に決定して相手に伝えられるような状況にないことも多い．多くの場合には自分の意志を決定するまでの時間が存在し，その時間とその間に発した音声が，システムの誤認識を引き起こしてしまう．発話意図の識別はこのような問題に対応できると考えられる．

5.2　データ

音声対話システムが，近隣での食事に関する推薦を行ったと想定して以下のようなスクリプトを作成した．

User:この近くでご飯を食べるところを紹介して

System:○○なんていかがですか

User:○○ね

○○には，音素バランスやアクセント位置，語の長短などを考慮した 10 種類の名詞が入る（表 5.1 参照）．このスクリプトに従って，2 種類のデータを用意した．なお，音声の収録は両データとも防音室でおこなった．

データ 1

最後の，“○○ね” にあたる部分をそれぞれ 10 回（肯定的応答 5 回，否定的応答 5 回）ずつ発声したものを収録した．従ってラベルは肯定／否定の 2 値であり，発話者が発声時に込めた意図をそのまま採用した．収録には男性 5 名女性 5 名の計 10 名に参加してもらい，合計で 1000 発話をデータとして利用した．

データ 2

最後のユーザの応答を収録する点は**データ 1** と同じである．**データ 2** では，終助詞の種類を “ね” 以外にも，“か” や “な” も許すこととした．また名詞のみという音声も収録した．

表 5.1: 肯定否定判別の分析に使用する単語及びその特徴

単語	アクセント型	特殊拍	無声子音	モーラ長
ラーメンね *[ra:meNne]*	起伏型	長音	なし	4+1
うどんね *[udoNne]*	平板型	なし	なし	3+1
そばね *[sobane]*	起伏型	なし	[s]	2+1
カレー *[kare:ne]*	平板型	長音	[k]	3+1
スパゲッティね *[spageqtine]*	起伏型	促音	[s][p][t]	5+1
ハンバーガーね *[hanba:ga:ne]*	起伏型	長音長音	[h]	6+1
シュークリームね *[shu:kuri:mune]*	起伏型	長音長音	[sh][k]	6+1
お茶漬けね *[ochazukene]*	平板型	なし	[ch]	4+1
焼肉ね *[yakinikune]*	平板型	なし	[k]	4+1
そうね *[so:ne]* ※データ 1 のみ	起伏型	長音	[s]	2+1
和食ね *[washokune]* ※データ 2 のみ	平板型	なし	[sh][k]	3+1

但し，収録した話者は 1 名で，1 話者 ×4 終助詞 ×10 回 = 計 400 発話のデータとなっている．又，**データ 2** は，収録時には肯定的応答と否定的応答を 5 回ずつ発声したものを，発話者本人が収録後音声を聞いて改めてラベリングを行った．ラベルは強肯定／弱肯定／曖昧／弱否定／強否定の 5 値である. その結果，内訳は強肯定：98 発声，弱肯定：70 発声，曖昧：46 発声，弱否定：70 発声，強否定：116 発声となった．**データ 2** の収録に関する狙いについては後節で述べることとする．

5.3　識別に利用した音響特徴

以下の 2 値／5 値の識別実験で用いた音響特徴と，その抽出条件について述べる．

韻律的特徴量

観測した音声波形から，基本周波数と波形強度，音素持続時間を自動抽出し，表 5.2 に示すような項目を計算し，特徴量ベクトルを作成した．基本周波数パターンと波形強度の自動抽出は Praat[59] で行った．2·4 章で述べたとおり，基本周波数の正確な自動抽出は困難であり，本実験においても数点の抽出誤りが見つかった．しかし，システム化の観点からあえて修正はしていない．また，音素持続時間は，Julian（現在は Julius に統合）[60] を用いてセグメンテーションを行い，基本周波数などと同様に，各音素の持続時間の計測には Praat を用いた．

MFCC

Hidden Markov Model Toolkit：HTK[61] を用いて各音声から MFCC を計算した．抽出条件を表 5.3 に示す．収録時における収録機器との距離などについては発話者によって違

表 5.2: 韻律的特徴量の抽出項目

基本周波数 [logHz]，波形強度 [dB]	音素持続時間 [s]
最大値	短母音の持続時間 · · · (1)
最小値	句末音素の持続時間 · · · (2)
平均値	発声全体の持続時間 · · · (3)
最大値 - 平均値	(2) / (1)
平均値 - 最小値	(3) / (1)
最大値 - 最小値	
標準偏差	

表 5.3: MFCC の抽出条件

シフト長	抽出次元
25[ms]	12 次元の MFCC, 12 次元の ΔMFCC, 1 次元の ΔPower, 計 25 次元.

いがあるため，Power は時間変化量（Δ）のみ識別に利用する．

ST 特徴

Siqing らの感情識別 [2] と同じ手法で特徴量ベクトルを抽出した．抽出条件を表 5.4 に示す．各音声に対するするそれぞれのフィルタバンク処理に関しては MATLAB[62] を用いた．

5.4　2 値の識別

データ 1 を用いて，肯定的応答と否定的応答の自動識別実験を行った．実験条件を Table 5.5 に示す．表中の数字は 1 セットのもので，話者と単語の組み合わせを変えて 100 セットの試験を行った．話者 · 単語に対してオープンな試験になるようにした交差検定の変形版になっている．結果を，100 セットの平均値で示すと Fig.5.1 のようになった．図中の人間の識別率は，同じデータを用いて聴取実験を行った結果得られたものである．最も識別率がよかったのは，MFCC を用いたときで，混合数 8 のときに 92.6%を達成した．韻律的特徴を用いた場合に，平均値が他の 2 つの場合より小さくなっているのがわかる．韻律的特徴と ST 特徴での識別は，話者の違いによるばらつきが大きく，システムに応用する際には多様な話者が利用するようなタスクには不向きであることが示唆された．他方，単語に関しては MFCC や ST 特徴を用いると比較的偏差が小さくなっており，これらの 2 つの特

表 5.4: ST 特徴の抽出条件

窓長	シフト長	抽出次元
250[ms]	30[ms]	23 次元の Gammatone フィルタバンク× 8 次元の Modulation フィルタバンクによるフィルタリング．窓掛け後，DCT で次元圧縮して，分散が大きいものから 25 次元を選択.

表 5.5: 実験条件

識別器	HMM 25 状態
混合数	混合なし，2，4，8，16，32
特徴量	韻律的特徴，MFCC，ST 特徴
学習音声	810 発声（9 話者 ×9 単語 ×10 回）
評価音声	100 発声（1 話者 ×1 単語 ×10 回）

徴が単語に対して頑健に識別できる可能性が示された.

5.5　5 値の識別

データ 2 を用いてクラスを 5 値に拡張した識別実験を行った.

本実験を行うにあたり，データ 1 に対する人間による 5 段階ラベリングを試みた．ラベラーは女性 3 名男性 7 名の計 10 名．その結果，被験者間である程度の相関はみられたものの，ばらつきが大きく，また被験者の感想として再現性が低いことがあげられるなど，5 段階の識別が人間にとっても難しいことが伺える結果となった．そこで，本実験では 2 値識別実験において ST 特徴を用いた識別の結果，話者に対する偏差が大きかったことを考慮し，発話者及びラベラーを一人に限定したデータ 2 を利用することで，特定の個人専用の識別器の実現を目指すこととした．話者同定技術と併用することで家庭内などある程度利用者が限定される状況での使用が想定される.

実験条件を Table 5.6 に示す．表中の数字は 1 セットのものである．2 値識別の時と同様に，各単語に対してオープンな試験となるよう単語を変えて 10 セット行い，その平均値を識別率とした．Table 5.7-5.8 に MFCC と ST 特徴を用いた場合それぞれの confusion matrix を示す．2 値での識別に比べて識別率がだいぶ低くなっているが，confusion matrix を見ると誤りのほとんどが隣接するクラスに分類されていることがわかる.

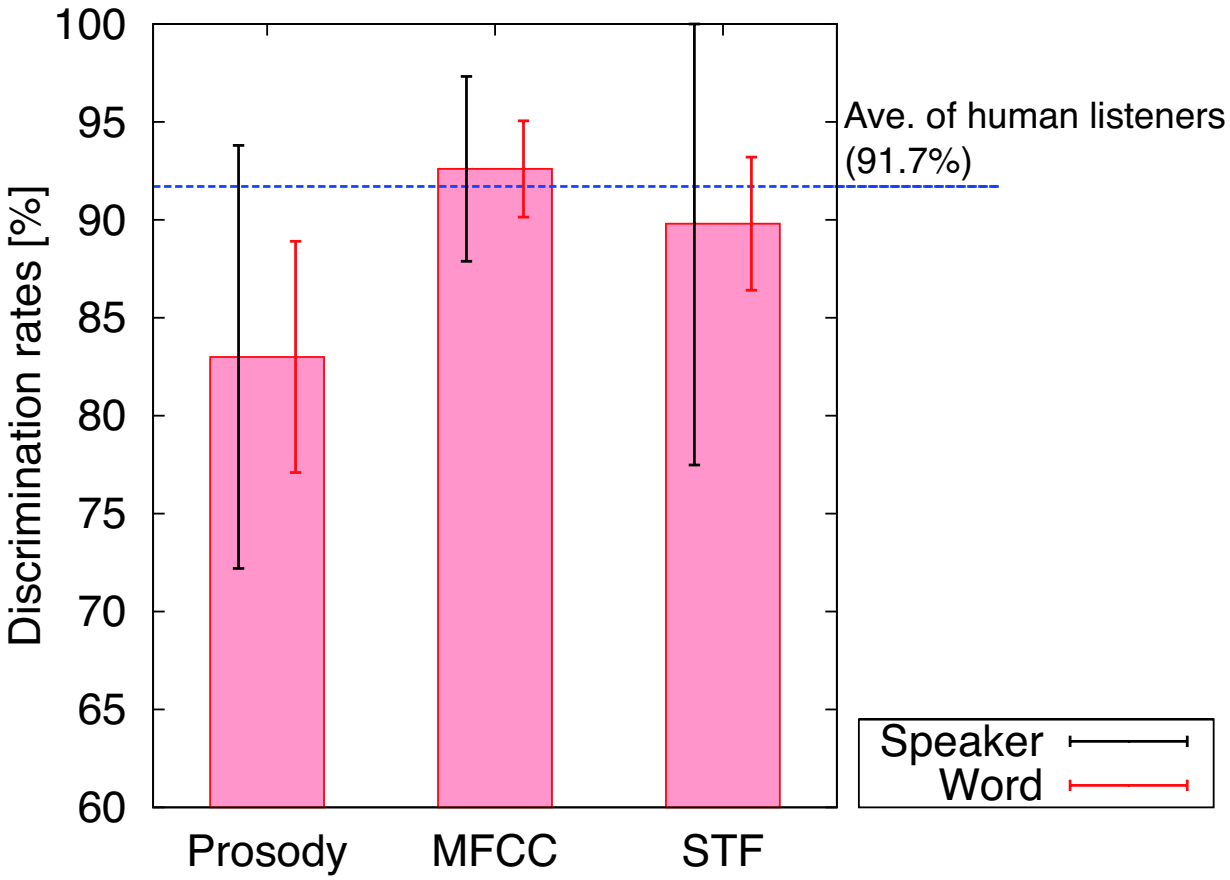

図 5.1: 識別率と偏差（混合数 8）

5.6　まとめ

5.6.1　識別に利用する特徴に関する検討

本章では韻律的特徴，MFCC，ST 特徴の 3 種の特徴量を用いて話者の意図の自動識別を行った．これら 3 種類の特徴のうち，韻律的特徴についてはそれぞれの要素が音の高低や大小などに直接影響を与えることから，他の 2 つの特徴に比べ直観的に理解しやすいという長所があった．しかし，音素持続時間を取得するためには音素アライメントのために言語情報を正確に認識する必要があり本研究の目的の一つである誤認識の回避を達成できない．また，基本周波数の自動抽出にはバイピッチや半ピッチなどのエラーが起こりやすい．以上の理由から，韻律的特徴は本研究で提案するようなシステムには不向きであると判断される．

5.6.2　クラス拡張に関する検討

本章では，2 値の識別の他に，5 値にクラスを拡張しての識別を試みた．その結果，やはり 5 値の識別は 2 値に比べて精度が落ちることがわかった．しかし，confusion matrix を見

表 5.6: 実験条件

識別器	HMM 25 状態
混合数	8
特徴量	MFCC，ST 特徴
学習音声	360 発声（1 話者 ×9 単語 ×4 終助詞 ×10 回）
評価音声	40 発声（1 話者 ×1 単語 ×4 終助詞 ×10 回）

表 5.7: confusion matrix : MFCC

	強否定	弱否定	曖昧	弱肯定	強肯定
強否定	**64.7%**	23.3%	6.9%	2.6%	2.6%
弱否定	50.0%	**30.0%**	17.1%	2.9%	0.0%
曖昧	6.5%	21.7%	**17.4%**	28.3%	26.1%
弱肯定	1.4%	0.0%	8.6%	**40.0%**	50.0%
強肯定	2.0%	0.0%	3.1%	32.7%	**62.2%**

表 5.8: confusion matrix : ST-features

	強否定	弱否定	曖昧	弱肯定	強肯定
強否定	**63.8%**	24.1%	5.2%	3.4%	3.4%
弱否定	48.6%	**34.3%**	17.1%	4.3%	1.4%
曖昧	8.7%	23.9%	**21.7%**	26.1%	19.6%
弱肯定	2.9%	0.0%	8.6%	**38.6%**	50.0%
強肯定	0.0%	0.0%	3.1%	33.7%	**63.3%**

ると識別を誤った発話のほとんどが隣のクラスに識別されているのがわかる．試しに 2-best で識別結果を出してみたところ，平均で 83%の識別率であった．よって，応答生成の段階で，隣のクラスであっても許容できるような応答を生成することで，5 値での識別も可能となる．

第6章

確信度を用いた柔軟な応答生成

6.1　はじめに

肯定／否定の 2 値識別実験において，MFCC や ST 特徴を用いて，人間の識別能と同程度の識別が可能であることが示された．次に，発話者の意図がどの程度定まっているのかを識別して対話管理に活かすべく，識別の際に得られた尤度の比をもって話者の意図の確信度と定義した．これは，話者の意図が定まっていない発話ほど尤度の比が小さくなるという推測に基づいたものである．5 章で述べた 5 値の識別結果を応答生成に反映させることも考えられるが，識別率の低さを考えると，実験結果をそのままシステムへ応用することは難しい．

本節では，意図の確信度を用いて識別誤りを回避し，動的に応答を変化させる可能性について検討する．

6.2　自動識別における誤識別回避

6.2.1　2 値に対する対応

識別実験で得られたそれぞれのクラスの尤度から，確信度（Confidence Index：CI）を 6.1 に従って計算した．

$$CI = \frac{L_1}{L_2} \times p(C_1) \tag{6.1}$$

$$\begin{aligned} L_1 &: \text{Likelihood of result 1.} \\ L_2 &: \text{Likelihood of result 2.} \\ p(C_1) &= \frac{\text{number of utterances in corpus} \in \text{Class 1}}{\text{number of all utterances in corpus}} \end{aligned} \tag{6.2}$$

これは，2 値の識別の場合，評価対象発話のクラス C_1 に対する事後確率を計算していることにほぼ等しい．実際にはそれぞれの尤度は対数で得られるので確信度 CI も対数で得られる．また，2 値の識別では，データ 1 を用いたため肯定的発話と否定的発話の数は等しくなっており，$p(C_1)$ は計算する必要がない．このようにして得られた確信度 CI（対数値）と識別結果が誤っていた発話の分布との関係を調査した．図 6.1-6.2 は識別を誤ったサンプルの数がどのように分布しているか，確信度 CI の対数を横軸にとってグラフ化したものである．確信度が低くなるほど識別を誤るサンプルの数が増えていることが見て取れる．ST 特徴を用いた識別の場合，識別を誤った発話のおよそ 90%が確信度 200 以下に存在することがわかった．これは，確信度で適切にしきい値を設定し，対話管理に利用することで識別を誤る発話の大半を検知し動的に対応することが可能であることを示している．

6.2.2　多値に対する対応

前章で，5 値における識別の結果が 2 値の時に比べてかなり悪かったことを述べた．本小節では，2 値の場合と同様に，識別を誤った発話の検出し，適切な応答を生成する方法

について検討する．

まず，2 値の場合と異なるのは，多値の識別の場合，必ずしも識別のクラス数以上の識別は求められていないということである．具体的に 5 値の場合で考えてみると，2 値の場合と同じように取り扱うと 5 値それぞれの中間のクラスが生成され，計 9 クラスの識別となる．人間の心的状況は離散的なクラスで詳細に分類できる類のものではなく，なおかつ本研究で想定しているようなユーザの意思決定支援タスクにおいては，識別を誤る確率の高いような発話では，ユーザは肯定と否定の間で迷っている可能性が高いと考えられる．従って，揺れ動く心的状況を 9 つもに分けられたクラスに一意に決定するのは返ってナンセンスである．多値の識別における誤識別検出法としては，肯定的発話を否定だと誤るような，識別を誤った際のコストが高い発話のみを検出するほうがユーザに対する親和度が高いと考え，以下のような 3 つの場合に該当するような発話を検出し，その場合の尤度の比などについて分析を行った．

1. 識別結果 1 位が，**強肯（否）定** で，かつ識別結果 2 位が，**強 · 弱否（肯）定**.
2. 識別結果 1 位が，**弱肯（否）定** で，かつ識別結果 2 位が，**強否（肯）定**.
3. 識別結果 1 位が，**曖昧** で，かつ識別結果 2 位が，**強肯定 · 強否定**.

上記のいずれかの条件に当てはまる発話は，MFCC で識別した場合で 13 発話．ST 特徴で識別を行った場合に 23 発話であった．条件別に分析を行った結果を以下に示す．

１：識別結果 1 位が 強肯（否）定 で，かつ識別結果 2 位が 強 · 弱否（肯）定 の場合．

この条件に該当する発話は，MFCC による識別の場合で 3 発話（全体の 0.75%），ST 特徴による識別の場合で 10 発話（2.5%）であった．MFCC による識別の場合，サンプル数が小さくて傾向の分析が難しいが，3 発話中 2 発話が識別結果第 2 位は正解ラベルであった．又，尤度比による有意な差は確認できなかった．尚，本分析においては 2 値の識別の際に用いてた確信度という言葉を用いない．クラスが 2 つ以上ある場合，6.1 式は事後確率と等価とは言えないためである．一方，ST 特徴を用いた識別の場合では，条件 1 に該当すると分類された 10 発話では，まず，1-best で識別を誤っていない発話が 3 発話あり，しかも，尤度比もバラバラであった．また，1-best でも隣接するクラスを正解とした場合に正解となる発話が 3 発話あり，同じく尤度比に関連は見出せなかった．2-best にした際に正解となる発話は 2 発話であった．

２：識別結果 1 位が 弱肯（否）定 で，かつ識別結果 2 位が 強否（肯）定 の場合．

この条件に該当する発話は，MFCC による識別では検出されなかった．また，ST 特徴による識別の場合も 1 発話（0.25%）しか検出されず，有意味な分析は出来なかった．

３：識別結果 1 位が 曖昧 で，かつ識別結果 2 位が 強肯定 · 強否定 の場合．

この条件に該当する発話は，MFCC による識別の場合で 10 発話（2.5%），ST 特徴による識別では 12 発話検出された．この条件に該当する発話では，識別を誤っていない発話が

表 6.1: 想定される対話例

システム	：	周辺案内システムです．何かご用ですか？
ユーザ	：	おなかが減りました．
システム	：	近くにおいしいラーメン屋があります．ご紹介しましょうか？
ユーザ	：	ラーメンね．

表 6.2: ユーザの入力の意図／確信度と，それに対するシステムの出力

確信度	意図	システムの応答
高	肯定	該当のお店を紹介する．
	否定	別のお店を紹介する．
低	／	該当のお店の追加情報と他の選択肢について示唆する．

MFCC で 5 発話，ST 特徴で 3 発話検出されてしまっていたが，隣接するクラスを正解とみなした場合，識別結果第 1 位と識別結果第 2 位の中間のクラスを出力することで，1 発話を残して全て正解となった．

考察

上記の分類結果から，識別結果 1 位が曖昧である場合には，識別結果 2 位との中間クラスを採用することで，対応できることが示唆された．一方，識別結果が強肯（否）定である場合には，何らかの損失関数を定義することで対応できそうな問題ではあるが，データを概観した限り識別結果 1 位と 2 位の尤度の比が有効な変数となりそうにない．この損失関数については，異なる角度からの検討が必要であると考えられる．

6.3　デモ

確信度の導入により 2 値識別から 3 値識別への拡張を行ったシステムのデモを構築した（図 6.3）．システムの根幹には Galatea Toolkit[63] を用いた．タスクはお昼ごはんを食べる場所をテーマとしたユーザの意思決定支援タスクとした．ユーザの入力が行われるまでの想定スクリプトは表 6.1 の通り．システムはユーザの入力の意図と確信度に対して，表 6.3 のような応答を出力する．なお，今回は意図と確信度を対話システムに渡す部分は人間の手で渡す WOZ システムとなっている．

6.4　まとめ

本章では，確信度の概念を導入することによって，2 値の識別の際に識別を誤る可能性の高い発話を検出し，そのような発話に対しては柔軟に対応することができる可能性につ

いて示した．また，多値の識別の場合でも，2-best の識別結果の順序比較等を行うことによって，一部柔軟な対応が可能であることを示した。

6.4.1　確信度の導入による柔軟な応答生成に関する考察

確信度の概念を導入することで，話者の意図識別を行うと共に，識別の際のクラスの数以上に多様な対応が可能であることを示した．これは，単純にクラスの数を増やすことでも，実現できるが，5 値の識別の結果からもわかるように，話者を限定した場合でも，識別率は低く抑えられており，システムに利用する場合には，前節で述べたような対応が必要であると考えられる．確信度の定義については，本研究では，2 値の場合に尤度の比を確信度と定義した．しかし，2 値より多くのクラスに対応する場合にはこの定義が有効であるかは自明でなく，今後検討すべき点である．

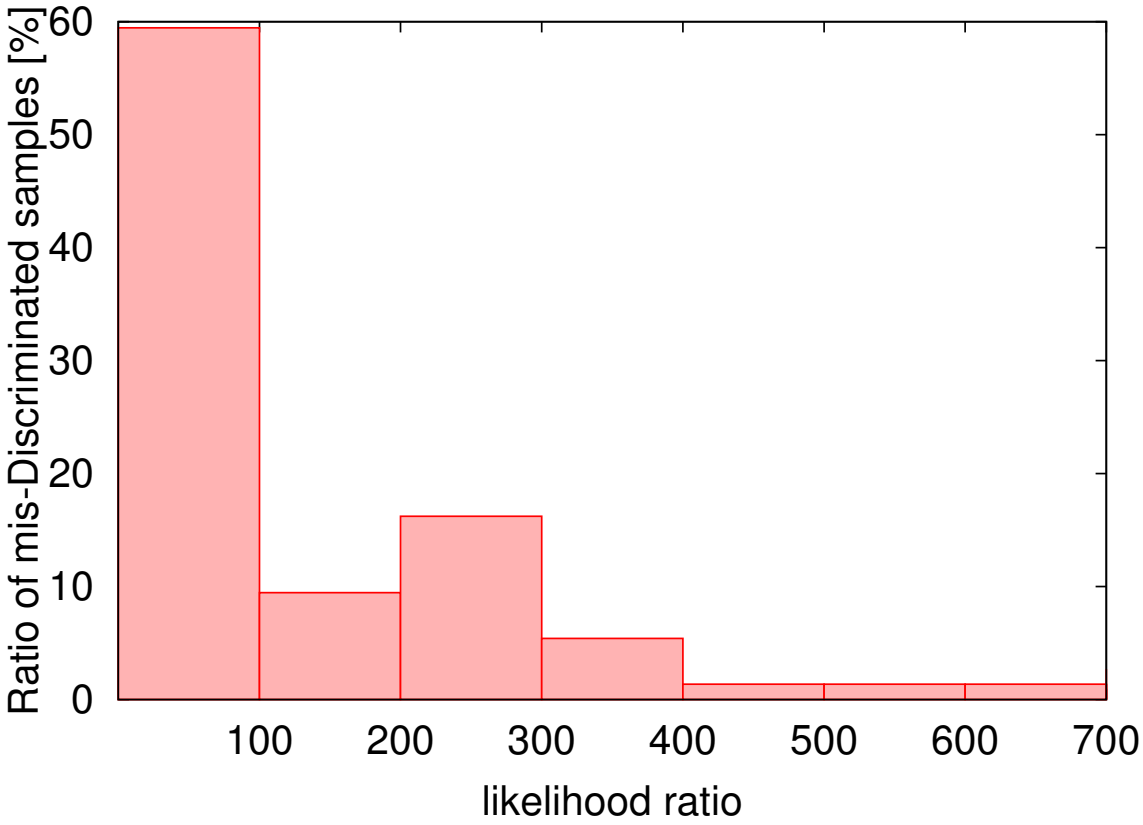

図 6.1: 識別を誤ったサンプルと確信度の関係:MFCC

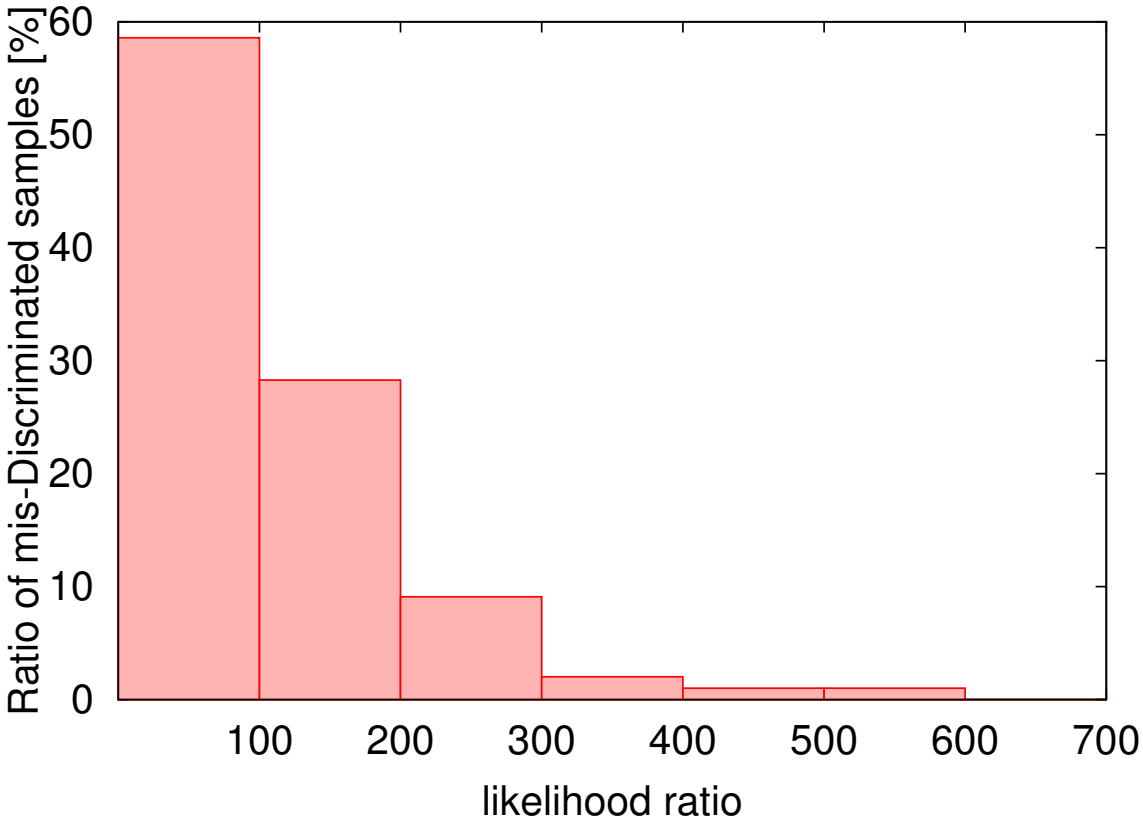

図 6.2: 識別を誤ったサンプルと確信度の関係:ST 特徴

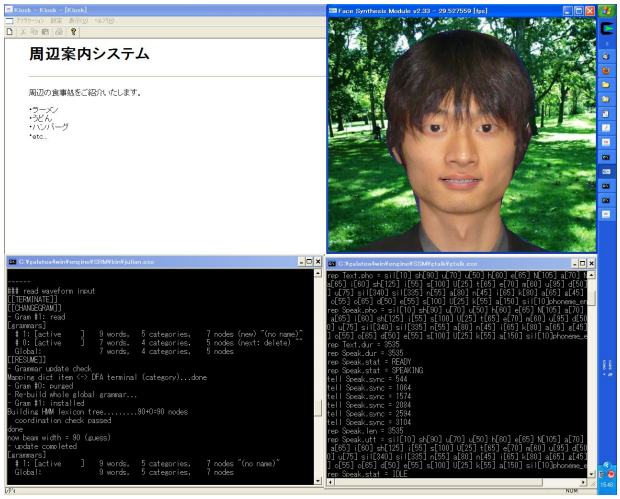

図 6.3: トップ画面

図 6.4: 肯定的入力に対する応答画面

図 6.5: 否定的入力に対する応答画面

図 6.6: 確信度が低い入力に対する応答画面

第7章

結論

7.1　まとめ

本論文では，高まる音声対話システムへの需要を受け，音声の言語情報だけでなく，発話者の意図や感情などの非言語 · パラ言語情報を入力として受理し，それらの情報から話者の意図を正確に識別し，その結果を応答に反映させるために，話者の意図の識別に必要な音響特徴についての検討 · 及び意図識別結果を反映させた柔軟な応答生成手法の提案を行った．

実験の結果，まず，音響特徴については，MFCC 及び ST 特徴が識別精度及び実用の面での取り扱いにおいて有効であることを示した．韻律的特徴については，自動抽出時の誤差及び音素持続時間の自動抽出の際に音声認識結果である言語情報が必要であることなどから，本研究で試行した特徴量ベクトルの作成法については，不備があったといわざるを得ない．次に，話者の意図の識別結果を応答生成に活かす方法としては，2 値の識別においては確信度の概念を導入することで誤識別を減らし，実際のクラス分け以上の柔軟な応答が可能であることを示した．また，多値の場合においても，2-best の識別結果の分析を行うことで識別誤りを起こす確率の高い発話を 2 値の場合とは別に定義し，それらに対して柔軟な対応をとることで，精度の改善が見込めることを示した．

7.2　今後の課題

まず，音響特徴について，韻律的特徴の時間方向に対する観察は不十分であったと言わざるを得ない．多くの先行研究において，基本周波数パターンや波形パワーなどの時間微分などが特徴量として採用されているにも関わらず，本研究ではそれらの特徴を観察できていない．また，他の多数の研究がそうであるように，韻律的特徴と他の分節的特徴を同時に識別に利用する取り組みについても本研究では行っていない．MFCC や ST 特徴により，人間と同程度の識別は可能となったが，識別の精度を追求するのであれば，考慮すべき手法の一つであったと言える．

次に，応答生成の部分において，多値の識別の誤識別回避のための損失関数について，明確な解を得るに至らなかった．多値の場合においても 2 値の場合と同様に確信度に相当する値を得ることが鍵であると考えられる．

また，本研究をデモシステムに実装することは，手法の評価の面でも今後まず取り組むべき課題のひとつであると言える．

謝辞

本研究ならびに本論文の執筆にあたり，多大なる御指導，御鞭撻を賜りました指導教員の広瀬啓吉教授ならびに峯松信明准教授に深く感謝いたします．また，日頃の研究室活動を支えてくださった高橋登技官，秘書の楠本由香里さん，池上恵さんに深く感謝いたします．本研究を進めていく上で，博士課程の斉藤大輔氏には数多くの鋭いご指摘やご意見を頂きました．また，研究室運営をはじめとして，研究室において先生方と同じか，それ以上に，先輩が後輩を丁寧に指導することで後輩が成長し，研究室全体の質が向上することも，その振る舞いで教えていただきました．自分が当研究室でその教えを実践できていたかどうかにははなはだ自信がありませんが，本当に深く感謝しております．研究のアプローチについて，常に冷静に，多角的な視点からご指摘を頂いた，特任助教の喬宇博士にも深く感謝いたします．研究及び研究室生活全般において，先輩として時に厳しく時に優しくご指導いただいた國越晶氏にも，深く感謝いたします．また，同期としてその研究室における時間を一番長く近い感覚で共有した青木史朗氏，岡雅之氏，鈴木雅之氏，高澤真章氏らに深く感謝いたします．同期ということで大変に刺激を受け，また，自分が苦しいときも前に進むことが出来たのも，共に頑張る仲間がいると身近に感じることが出来たからでした．特に，岡雅之氏，鈴木雅之氏には，同期であるにもかかわらず，大変に多くのことをご教授いただきました．深く感謝しています．また，高澤真章氏には，学部の実験班時代から就職活動に至るまで共にすごした時間が多く，その友人に対する気遣いや物事に対して積極的に取り組む姿勢など大変多くのことを学ばせていただきました．深く感謝しています．研究室生活におきましては，当時博士課程の平野宏子氏·朝川智氏·孫慶華氏，博士課程の高田靖也氏·小杉康宏氏·越智景子氏·羅徳安氏·馬学淋氏，当時修士課程の稲垣貴彦氏·鎌田圭氏·篠田知宏氏·下村直也氏·デゲル エルハン氏·ナジャンド アリ氏·ナリニョオ ホアン氏·三輪周作氏·レボルダオ アントニオ氏·印南圭介氏·馬敏懿氏·松浦良氏·細田聖人氏に感謝いたします．諸先輩方のおかげで，楽しく研究室生活をスタートさせ，かつ充実した研究室生活を送ることが出来ました．また，名前をすべて挙げることはかないませんが，広瀬峯松研究室で共に時間をすごした全ての諸先輩方·後輩諸君らに心より感謝いたします．また大学生活において一番長く時間を共有したであろうサークルの友人たちにこの場を借りて心よりの感謝を申し上げます．人間として一番多くを学ぶ機会をいただきました．最後に 25 年間ここまで自分のわがままを聞き入れ，自由に育てて頂いた家族と，近くから遠くから見守ってくださった親戚への謝意を表したいとおもいます．どうもありがとうございました．

2010 年 2 月 9 日
高橋琢己

参考文献

[1] R.Tato *et al.*：“Emotional Space Improves Emotion Recognition，” Proc. INTERSPEECH2002，pp.2029-2032，2002.

[2] Siqing Wu *et al.*：“Long-Term Spectro-Temporal Iformation for Improved Automatic Speech Emotion Classification，” Proc. INTERSPEECH2008，pp.638-641，2008.

[3] N.Sato *et al.*：“Emotion Recognition using Mel-Frequency Cepstral Coefficients，” Journal of Natural Language Processing Vol.14 No.4，83-96，2007.

[4] 藤崎博也：“音声の韻律的特徴における言語的 · パラ言語的 · 非言語的情報の表出，” 信学技報,，1994.

[5] 森山剛，斎藤英雄，小沢慎治：“音声における感情表言語と感情表現語と感情表現パラメータの対応付け，” 信学論，Vol.J82-D-II No.4 pp.703-711，1999.

[6] 鈴木久喜：“ピッチ抽出の今昔，” 日本音響学会誌 Vol.56 No.2，pp.121-128，2000.

[7] 広瀬啓吉：“21 世紀に向けての音声合成の技術展望，” 情報処理学会誌 Vol.41 No.3 pp.277-281，2000.

[8] Laver，J.：“Phonatory settings. In The phonetic description of voice quality，” Cambridge University Press Ch. 3，pp.93-135，1980.

[9] 石井カルロス寿憲 他.：“韻律および声質を表現した音響特徴と音声対話におけるパラ言語情報の知覚との関連，” 情報処理学会論文誌 Vol.47，No.6，1782-1793，2006.

[10] 大淵康成：“音声認識 · 理解のための特徴抽出，” 日本音響学会秋季講演論文集，pp.313-316，2009.

[11] S.Davis and P.Mermelstein：“Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences，” IEEE Trans. ASSP Vol.28 No.4，pp.357-366，1980.

[12] 室井貴司，滝口哲也，有木康雄：“スペクトル平面における勾配ヒストグラムに基づく音声特徴量の検討，” 情報処理学会研究報告. SLP, 音声言語情報処理 2008(123) pp.161-165，2008.

[13] 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄：“音声認識システム，” Ohm 社，2001.

[14] S.J.Young:“A review of large-vocablary continuous-speech recognition，” IEEE Signal Proc. Vol.13 No.5，pp.45-57，1996.

[15] 緒方公一，増矢拓郎：“Web アプリケーションとしての声道音響管モデルに基づく母音合成システムの開発，” 日本音響学会誌 Vol.62 No.3，pp.199-207，2006.

[16] 誉田雅彰，西川員史，高西淳夫，廣谷定男，持田岳美：“人間型発話ロボット -喉を震わせ口を動かして発話するロボット-，” 日本音響学会誌 No.61 Vol.2，pp.91-96，2005.

[17] 錘 V 信行，河井恒：“素片接続型音声合成における最良優先探索に基づく素片選択，” 電子情報通信学会技術研究報告 SP2005-161，pp67-72，2006.

[18] ニック · キャンベル，アラン · ブラック：“CHATR: 自然音声波形接続型任意音声合成システム，” 信号処理学会技術報告 Vol.96 No.39，pp.45-52，1996.

[19] W.Hamza，R.Bakis，Z.W.Shuang and H.Zen：“On building a concatenative speech synthesis system from the Blizzard Challenge Speech Databases，” Proc. EUROSPEECH，pp.97-101，2005.

[20] T.Nose，J.Yamagishi，T.Masuko and T.Kobayashi：“A style control technique for HMMbased expressive speech synthesis，” IEICE Trans. Inf. & Syst. Vol.E90-D No.9，pp.1406-1413，2007.

[21] S.J.Young and F.Fallside：“Speech Syntthesis from concept : A method for speech output from information systems，” J. Acoust. Soc. Am. Vol.66 No3，pp.685-695，1979.

[22] 広瀬啓吉：“音声合成技術，” 情報処理学会誌 Vol.38 No.11，pp.984-991，1997.

[23] 河原達也，松本裕治：“音声言語処理における頑健性，” 情報処理学会誌 Vol.36 No.11，pp.1027-1032，1995.

[24] Hobbs，J.R.，*et al.*：“Robust Processing of Real-World Natural-Language Texts，” Applied Natural Language Conferences，pp.186-192，1992.

[25] Douglas，S. and Dale，R.：“Towards Robust PATR，” COLING，pp.468-474，1992.

[26] 藤江真也，他.：“音声対話ロボット ROBISUKE による相談型対話の実現，” 情報処理学会研究報告 2004-SLP-53 No.10，pp.55-56，2004.

[27] 河原達也：“音声でスライド画面を操作する，” bit 4 月号 共立出版，2000.

[28] 駒谷和範，河原達也：“音声認識結果の信頼度を用いた効率的な確認·誘導を行う対話管理，” 情報処理学会論文誌 Vol.43 No.10，pp.3078-3086，2002.

[29] 堂坂浩二，安田宜仁，宮崎昇，中野幹生，相川清明：“システム知識制限下における効率的対話制御，” 音声言語情報処理 Vol.33 No.9，pp.49-54，2000.

[30] Bennacef，S.，Devillers，L.，Rosset，S. and Lamel，L.：“Dialog in the RAILTEL Telephone-Based System，” ICSLP Proc.，1996.

[31] Goddeau，D.，Meng，H.，Polifroni，J.，Seneff，S. and Busayapongchai，S.：“A Form-Based Dialogue Manager for Spoken Language Applications，” ICSLP Proc. ，1996.

[32] Sturm，J.，Os，E. and Boves，L.：“Issues in Spoken Dialogue Systems: Experiences with the Dutch ARISE System，” ESCA IDS’99 Proc.，1999.

[33] E. Levin，S.Narayanan，R.Pieraccini，K.Biatov，E.Bocchieri，G.DiFabbrizio，W.Eckert，S.Lee，A.Pokrovsky，M.Rahim，P.Ruscitti and M.Walker：“The AT&T-DARPA communicator mixed-initiative spoken dialog system，” ICSLP Proc. Vol.2，pp.122-125，2000.

[34] A.Potamianos，E.Ammicht and H.Kuo：“Dialogue management in the Bell Labs communicator system，” ICSLP Proc. Vol.2，pp.603-606，2000.

[35] J.F.Allen，B.W.Miller，E.K.Ringger and T.Sikorski：“A robust system for natural spoken dialogue，” 34th Annual Meeting of the Association for Computtational Linguistics（ACL-96） Proc.，pp.62-70，1996.

[36] S.Harabagiu，D.Moldovan and J.Picone：“Open-domain voice-activated question answering，” COLING Proc.，pp.502-508，2002.

[37] 藤井敦：“音声による言語バリアフリーな他言語情報アクセス，” 情報学研報 SLP-44-33，2002.

[38] C.Hori，T.Hori，H.Isozaki，E.Maeda，S.Katagiri and S.Furui：“Deriving disambiguous queries in a spoken interactive ODQA system，” ICASSP Proc. Vol.1，pp.624-627，2003.

[39] 駒谷和範，河原達也，清田陽司，黒橋禎夫，P.Fung：“柔軟な言語モデルとマッチングを用いた音声によるレストラン検索システム，” 情報学研報 SLP-39-30，2001.

[40] 翠輝久，駒谷和範，清田陽司，河原達也：“音声対話によるソフトウェアサポートタスクのための効率的な確認戦略，” 電子情報通信学会論文誌 Vol.J88-D-II No.3，pp.499-508，2005.

[41] 竹内真士，北岡教秀，中川聖一：“韻律·表層的言語情報を発話タイミング制御に用いた雑談対話システム，” 情報処理学会 2004-SLP-50 (14)，pp.87-92，2004.

[42] 河原達也，川嶋宏彰，平山高嗣，松山隆司：“対話を通じてユーザの意図·興味を探り情報検索·提示する情報コンシェルジェ” 情報処理学会 Vol.49 No.8，pp.1-6，2008.

[43] 峯松信明，広瀬啓吉，関口真理子：“話者認識技術を利用した主観的高齢話者の同定とそれに基づく主観的年代の推定，” 情報処理学会論文誌 Vol.43 No.7，2186-2196.

[44] 篠田知宏，広瀬啓吉，峯松信明，小杉康宏：“ユーザの特徴・知識を考慮した番組情報検索音声対話システムの構築，” 日本音響学会秋季講演論文集，pp.71-72，2007.

[45] Yuji Yagi，Seiya Takada，Keikichi Hirose and Nobuaki Minematsu：“Concept-to-Speech Conversion for Reply Speech Generation in a Spoken Dialogue System for Road Guidance and Its Prosodic Control，” 4th Joint Meeting of ASA(Acoustical Society of America)/ASJ(Acoustical Society of Japan), 2006-11.

[46] 越智景子，広瀬啓吉，峯松信明：“基本周波数パターン生成過程モデルに基づくコーパスベース韻律制御における焦点制御，” 日本音響学会講演論文集，pp.369-370，2008.

[47] 藤江真也，江尻康，菊池英明，小林哲則：“肯定的/否定的発話態度の認識とその音声対話システムへの応用，” 電子情報学会論文誌 Vol.J88-D-II No.3，489-498，2005.

[48] 八木大三，藤江真也，菊池英明，小林哲則：“韻律情報を用いた発話態度認識とその対話システムへの応用，” 日本音響学会研究発表会講演論文集 2005(1)，pp.65-66，2005.

[49] McGilloway，S.，Cowie，R.，Douglas-Cowie，E.，Gielen，S.，Westerdijk，M.，Stroeve，S.，：“Approaching Automatic Recognition of Emotion from Voice: A Rough Benchmark，” ISCA Workshop on Speech and Emotion，Belfast，2000.

[50] Dellaert，F.，Polzin，T.，Waibel，A.：“Recognizing Emotion in Speech，” Fourth International Conference on Spoken Language Processing 3，pp.1970-1973，1996.

[51] M.A. Hearst：“Support Vector Machines，” IEEE Intelligent Systems，pp.18-28，1998.

[52] M.Westerdijk and W.Wiegerinck：“Generative Vector Quantisation，” ICANN，pp.934-939，1999.

[53] Oh-wook Kwon，Kwokleung Chan，Jiucang Hao，Te-Won Lee：“Emotion Recognition by Speech Signals，” Proc. EUROSPEECH 2003，pp.125-128，2003.

[54] Björn Schuller，Ronald Müller，Manfred Lang and Gerhard Rigoll：“Speaker independent emotion recognition by early fusion of acoustic linguistic features within ensembles，” Proc. Interspeech 2005，pp.805-809，2005.

[55] Tin Lay Nwe，Say Wei Foo，Liyanage C. De Silva：“Speech emotion recognition using hidden Marcov models，” Speech Communication Vol.41 Issue 4，pp.603-623，2003.

[56] 赤木正人：“聴覚フィルタとそのモデル，” 電子情報通信学会誌 Vol.77 No.9，pp.948-956，1994.

[57] Bernd T.Meyer and Birger Kollmeier：“Optimization and Evaluation of Gabor feature sets for ASR，” ISCA，pp.906-909 2008.

[58] M.Kleinschmidt and D.Gelbart：“Improving word accuracy with Gabor feature extraction，” ICSLP Proc.，2002.

[59] Praat：http://www.fon.hum.uva.nl/praat/

[60] Julius：http://julius.sourceforge.jp/

[61] HTK：http://htk.eng.cam.ac.uk/

[62] MATLAB：http://www.mathworks.co.jp/

[63] Galatea Toolkit：http://hil.t.u-tokyo.ac.jp/ galatea/index-jp.html

発表文献

[1] 高橋琢己，広瀬啓吉，峯松信明：“音声対話システムの高度化のための韻律を用いた話者の意図判別，”日本音響学会講演論文集，pp.491-492，2009.

[2] 高橋琢己，広瀬啓吉，峯松信明：“意図の確信度を用いた柔軟な応答生成，”日本音響学会講演論文集，2010（予定）.