韻律情報の利用に基づく対話システム
Spoken Dialogue System using Prosodic Information

小林 哲則　藤江 真也　松坂 要佐
Tetsunori Kobayashi　Shinya Fujie　Yosuke Matsusaka

早稲田大学
Waseda University

Abstract: A spoken dialogue system which uses non-linguistic information to realize natural and effective conversation is proposed. The system is implemented on conversation robot ROBISUKE. Firstly, we proposed a blackboard-based module communication architecture with publish/subscribe model for the system. Then, we applied two functions to the system, para-language understanding function and back-channel feedback function. For para-language understanding, prosody based attitude recognition and head gesture recognition are proposed. These sorts of information make a conversation smooth and effective. For back-channel feedback, the prosodic features and linguistic features are used to determine the appropriate timing and the appropriate content respectively. Finite State Transducer based speech recognizer is introduced to extract the feedback content earlier than the time to generate it. Experimental results show the effectiveness of these methods.

Keywords: Conversation robot, system architecture, para-language, back-channel feedback

1. はじめに

ユーザの発話に主として韻律の形で含まれる言語情報以外の情報を考慮し，円滑でリズムのある対話を実現する音声対話システムについて検討する．

従来の音声対話システムの多くは，音声認識器によって文字に変換されたユーザ発話を，言語処理によって解釈し，その結果に基づいてシステムの発話を決め，それを合成するという流れで対話を進行する．しかしながら，実際の人間同士の対話では，言語情報以外に伝達されるニュアンスを理解したり，相手の発話に対してその発話中に自分の状態をフィードバックするなどして対話を円滑に進めている．本研究では，これらの機能を音声対話システムに実装し，円滑でリズムのある対話を実現することを目的とする．

2. 黒板モデルを用いたロボット内通信アーキテクチャ

本節では，後述する音声対話システムを動作させる為のプラットフォームとして用いる，音声対話ロボット ROBISUKE(図 1) 内部における通信アーキテクチャについて述べる．音声対話システムをロボット上に実装するためには，モータ制御，音声認識，画像認識，言語処理，対話制御等，様々な粒度のタスクを実行するモジュールを並列に動作させる必要があるとともに，それらの効率的なメッセージのやりとりを実現する必要がある．

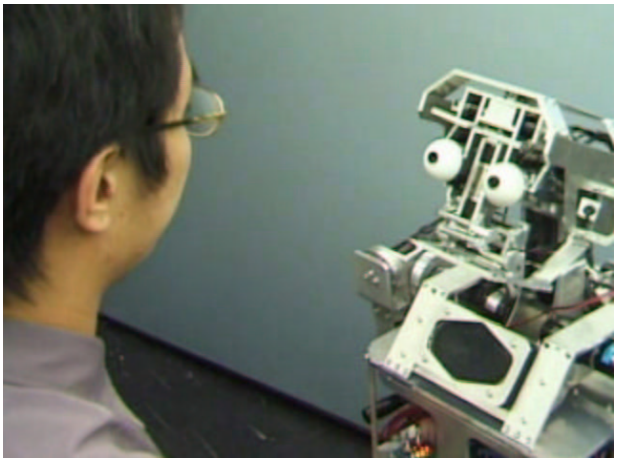


図 1: 音声対話ロボット ROBISUKE

各々のモジュールで生成された情報をシステム内のどのモジュールからも参照できるようにすることで，モジュール同士のフラットな関係を実現する．このような情報共有を実現する枠組みとして集中・開示型の情報公開モデルである黒板モデルを採用した．黒板モデルでは，各々のモジュールが共有の情報の開示場所に情報を「タグ」に書き込むことで，どのモジュールからもそれらの情報を参照することが可能となる．黒板モデルはデータ開示モデルとしては優秀なものであるが，データの変化を検知する用途には非効率である．そこで，メッセージ通知に適した枠組みである publish/subscribe モ

デルをあわせて採用した．このモデルでは，興味のあるデータを購読 (subscribe) リストに入れておけば，そのデータに対する変化の通知 (publish) を受けることができる．情報公開サーバは，黒板モデルに基づく情報開示サービスを提供するサーバと，publish/subscribe モデルに基づくメッセージ通知サービスを提供するサーバとの組合わせにより構成される (図 2)．

各々のモジュールは，情報開示サービスを利用して，自分にとって必要な情報を選定し，注目すべき情報はメッセージ通知サービスを利用して購読，新たに生成した情報を情報開示サービスを通して公開する．同じデータに対するアクセス手段として，情報取得とメッセージ配信が等価的に選択できるため，データを状態としてアクセスできると同時にイベントとしても取得できる．また，情報開示サービスには，センサーなどの外部情報だけでなく，システムの内部情報を区別なく開示できる．この枠組みによって，ロボット開発の要件となる自律性と応答性を両立できる環境でシステムを開発することが可能となる．

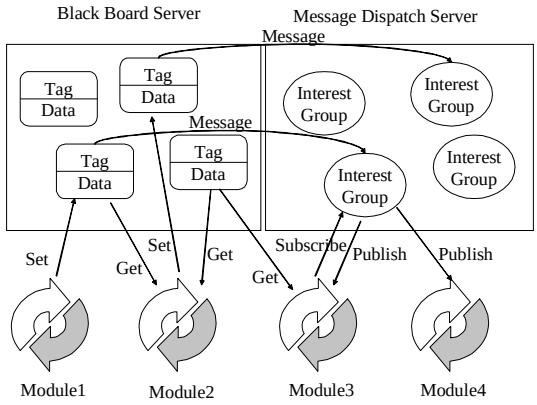


図 2: 情報公開・メッセージ通知モデル

3. パラ言語の理解と対話システムへの応用

本節では，音声対話システムへのパラ言語の理解能力の付与について述べる．

パラ言語として，ユーザの発話時の音声と画像から肯定的／否定的な態度を認識する．音声からは，基本周波数 (F0) や話速などの韻律情報から態度を認識する．画像からは，オプティカルフローを用いて，うなづき，かしげ，首振りを認識する．これら別々に得られた結果を統合し，システムの提案に対するユーザの反応を理解する．その理解に基づき，効率的な対話制御が可能なことを示す．

3.1. 韻律情報を用いた肯定/否定態度の認識

本節では，ユーザの発話態度を韻律情報から認識する手法について述べる．

まず，肯定的な態度，否定的な態度のデータを数多く収集するため，システムが提案をする際の発話を音声合成装置を用いて合成し，それに対するユーザの応答を肯定的，否定的な態度で発話するという形で収録を行った．収録した発話の種類を表 1 に示す．この表の言い回しの項目で,「○○」はシステムの提案に含まれるカテゴリや店の復唱を表す．収録は研究室の男子学生 20 人分 (計 2000 発話) について行った．

例えば,「ラーメン」「○○か」「否定的」というデータを収録する場合，まず合成音で「ラーメンなんてどうかな」という音声が流れる．その後ユーザが「ラーメンか」という内容を否定的な気持ちを込めて発話するという形で行った．

表 1: 収録音声の種類

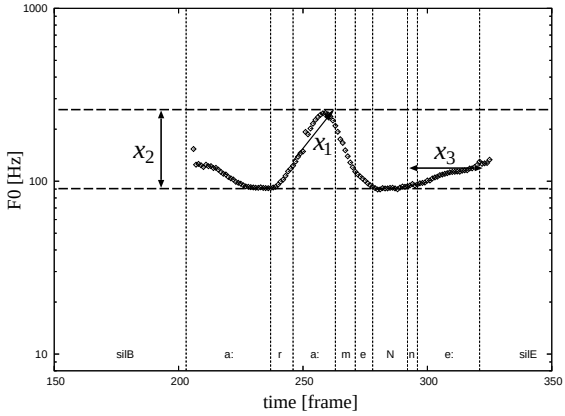
カテゴリ/店	ハンバーガー，ラーメン，弁当， カレー，学食，マクドナルド， 味源，夢民，ホカ弁，そばの実
言い回し	○○か，○○ね，○○ いいんじゃない，そうだね
態度	肯定的，否定的

収録した発話に関して，基本周波数 (F0) の抽出と音素アライメントを実行し，認識に用いる特徴量の検討を行った．図 3 に,「ラーメンね」という発話を肯定的，否定的に発話した場合の F0 と音素アライメントを示す．検討の結果，図中に示したように，以下の 3 次元の特徴量 $\boldsymbol{x} = (x_1, x_2, x_3)$ を識別に用いることとした．

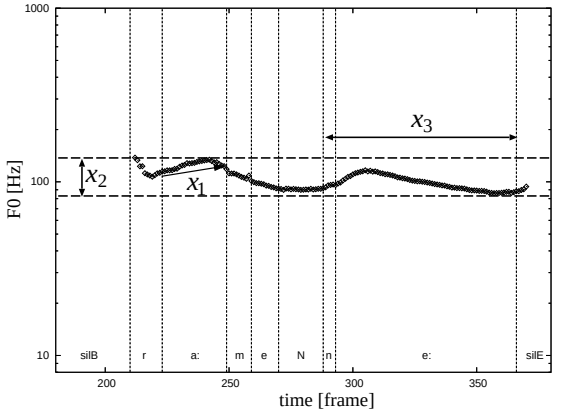
x_1：第 1 モーラの母音部分の F0 の傾き
x_2：発話全体の F0 レンジ
x_3：最終モーラの継続長

ここで，第 1 モーラの F0 の傾きは最小自乗法で求めた．抽出した特徴量を用いて，肯定的，否定的の 2 カテゴリについて混合正規分布 (Gaussian Mixture Model; GMM) を学習した．識別はこのモデルを用いたベイズ識別によって行う．

収録した 20 人分計 2000 発話のデータを，4 セット (1 セットあたり 5 人分 500 発話) に分け，交差検定を行った．結果を図 4 に示す．混合数 16 の時に 82.9%と最もよい認識率が得られた．



(a) 肯定的



(b) 否定的

図 3: 特徴量抽出の例

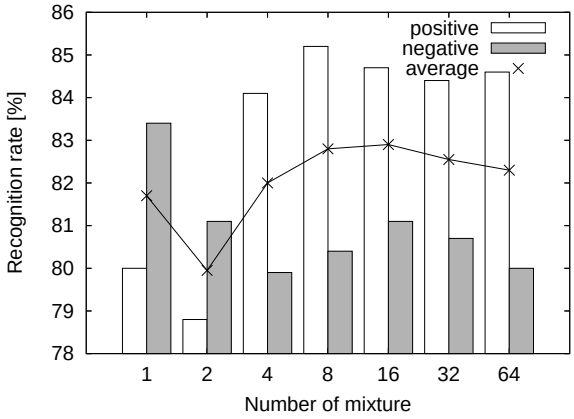


図 4: 韻律による発話態度認識の実験結果

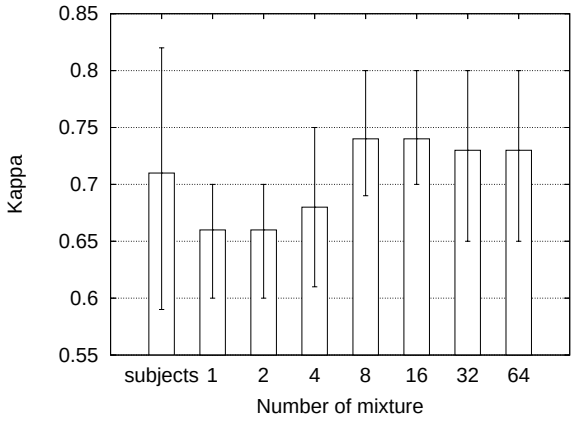


図 5: Cohen の κ の混合数による違い

また，人間の持つ識別能力との比較のため，収録した発話を人が識別した結果と，提案手法で識別した結果との一致率を見る．収録したデータのうち，態度が肯定的，否定的のものからそれぞれ 20 発話，計 40 発話をランダムに選択した．この 40 発話を 5 人の被験者 (A〜E) に順不同に聞かせ，それぞれ否定的か肯定的かを決めさせた．一致率の計算には，対話データのタグ付けの評価等に用いられる Cohen の κ を用いた．識別器の述べた識別結果を含む 6 組の結果について，それぞれの組合わせで κ を求めた．その結果を人毎に集計し，最小値，最大値，平均値を計算した．識別実験と同様に，モデルの混合数毎に結果を算出した．結果を図 5 に示す．この結果から，識別実験と同様，混合数 16 の時に一致率が最もよくなることが分かる．混合数 16 の時の κ の計算結果を表 2 に示す．表中の「認識器」は，認識器と各被験者間の κ の値を表し，A〜E は各被験者と他被験者間の κ を表す．表より，認識器と各被験者の一致率は，被験者間同士の一致率と同等であることが分かり，本手法による認識器が人間と同等の聞き分け能力を持つことが分かる．

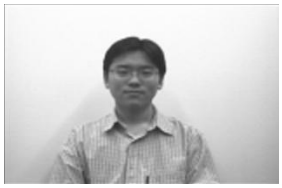
3.2. 頭部ジェスチャの認識

本節では，ユーザの発話態度を表す頭部ジェスチャとして，うなづき，かしげ，首振りの 3 ジェスチャを動画像から認識する手法について述べる．

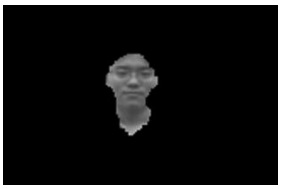
本研究で扱うデータは，自由対話のデータ (15 セット計 60 分) と，ユーザに特定のジェスチャを指示して行わせた様子を収録したデータ (25 セット計 144 分) の，計 40 セット 204 分のデータである．自由対話だけでは，極端に頻度の少ないカテゴリがあったため，これを補うため後者のデータセットを用意した．後者のものに関しても，ユーザにジェスチャを指示している最中のユーザの自然な動作も含まれるため，一部自然なジェスチャが含まれる．このデータのジェスチャ部

表 2: Cohen の κ の計算結果

	最小	最大	平均
認識器 (混合数 16)	**0.70**	**0.80**	**0.74**
A	0.52	0.85	0.72
B	0.52	0.79	0.62
C	0.65	0.80	0.72
D	0.65	0.85	0.76
E	0.61	0.79	0.72



(a) 入力画像



(b) 色モデルによる抽出

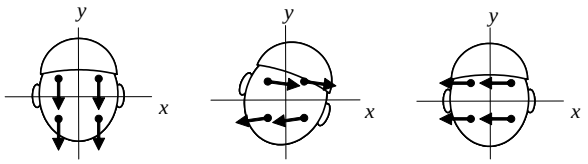
(c) 首領域の除去

図 6: 頭部の抽出

分に，人手でタグ付けを行なった．その結果，データから計 2148 個のサンプルが得られた．

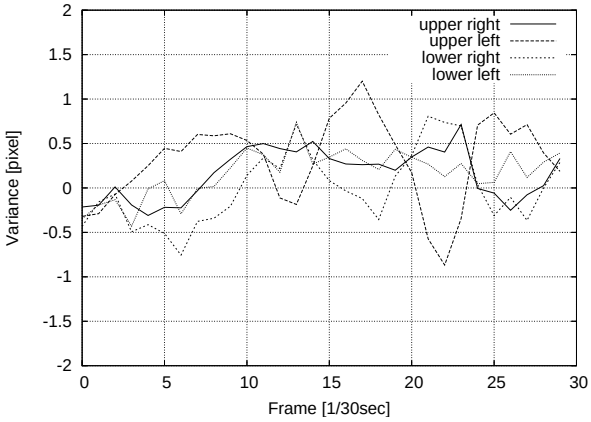
頭部領域の抽出は，入力画像 (図 6-(a)) から肌色情報を用いて，頭部領域を抽出する (図 6-(b))．さらに，頭部の縦横の比を用いて首領域の除去を行う (図 6-(c))．抽出した頭部領域の全画素に対しグラディエント法によって求めたオプティカルフローを，頭部の動きを表す特徴量として用いる．

抽出された頭部領域を上下左右に 4 分割し，各領域毎のオプティカルフローの平均ベクトルを求め，計 8

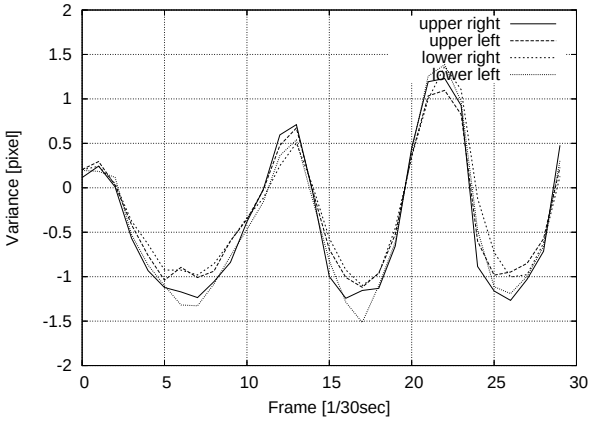


(a) nod (b) tilt (c) shake

図 7: 各ジェスチャの特徴的なフロー



(a)



(b)

図 8: うなづきのフローの時間による変化． (a) x 軸方向, (b) y 軸方向

次元の特徴量とする．このようにすることで，図 7 のように，首振りであれば 4 つの領域でフローが左右に同期するなどの，各ジェスチャの特徴を捕らえられることが期待できる．特徴量の時間変化の一例として，うなづきの特徴量を図 8 に示す．この図を見ると，x 軸方向の変化は不規則だが，y 軸方向の変化は 4 つのパラメタが同期して変化することが分かる． この様な時間的変化をモデル化するため，認識に用いる確率モデルとして，音声認識でよく使われる left to right 型の HMM(Hidden Markov Model) を用いる (図 9)．ただし，うなづきと首振りは繰り返しを含むため，モデルの構造にループを持たせる (図 10)．

用意するモデルとしては，今回設定した三つの認識対象のジェスチャ(うなずき，かしげ，首振り) 以外に，静止モデルとガーベージモデルの二つのモデルを設定

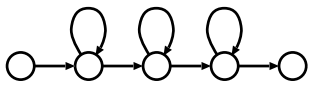


図 9: かしげ，静止状態，ガーベージ用の HMM

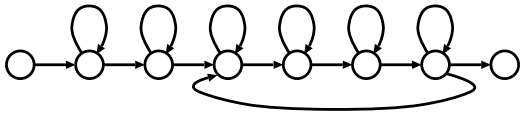


図 10: うなづき，首振り用の HMM

表 3: 頭部ジェスチャ認識の実験結果

	認識結果			
	うなづき	かしげ	首振り	脱落誤り
うなづき	**1213**	2	1	305
かしげ	12	**341**	5	197
首振り	3	0	**312**	14
挿入誤り	190	138	17	

した．静止モデルは頭部が動かないものを表すモデル，ガーベージモデルは設定した 3 つのジェスチャ以外の動きを表すモデルである．ガーベージモデルの導入によって，認識対象以外の頭部の動きに悪影響を受けることなく，認識対象の頭部ジェスチャのみを検出できる．

先に述べた 40 セットのデータを，学習データ 35 セット，認識データ 5 セットに分けて交差検定を行った．認識結果をコンフュージョンマトリックスで表したものを表 3 に示す．脱落誤りは正解ジェスチャが静止もしくはガーベージとして認識された回数，挿入誤りは静止もしくはガーベージが正解ジェスチャのどれかとして認識された回数を示す．この実験に用いたモデルのパラメータは，HMM の状態数が 11，正解ジェスチャモデルの混合数 4，静止モデルの混合数 1，ガーベージモデルの混合数 16 である．これらは様々なパラメータで実験を行った結果，最良の認識率を与えたものである．

認識対象とするカテゴリ以外の動作 (ガーベージとなるべきもの) が非常に多いので，結果として挿入が多くなっている (挿入誤りは，ガーベージから認識対象への置換と考えることもできる) が，割合としては小さく，アルゴリズムが有効に機能していることが分かる．

3.3. パラ言語情報の統合

前節までに述べた韻律情報による態度の認識と頭部ジェスチャの認識のそれぞれの結果を，パラ言語情報として応用するためには，両者の情報を統合することを考えなければならない．特に，両者に関して相反する結果が得られた時 (一方が肯定的な結果で，もう一方が否定的な結果の時)，どのように判断するかが問題である．

本研究では，韻律情報から得られる態度と，頭部ジェスチャの共起によって受ける印象から，表 4 に示す統合を行った．例えば，頭部ジェスチャが「うなづき」で韻律情報による態度が「肯定」の場合は「強い肯定」と解釈するが，韻律情報による態度が「否定」の場合はユーザは判断に迷って考えていると解釈する．

表 4: 認識結果の統合
(* は対話中ほとんど出現しない組み合わせを表す)

		頭部ジェスチャ		
		うなづき	かしげ	首振り
韻律	肯定	強い肯定	弱い肯定	思案*
	否定	思案*	否定	強い否定

3.4. 対話例

対話例を図 11 に示す．本研究で対象とする意思決定支援対話においては，システムの提案に対するユーザの応答が，発話の言語情報のみを見ていただけでは判断しかねる場合が多い．対話例においても，システムのカレーや弁当の提案に対するユーザの反応は,「カレーか」「ハンバーガーね」などと，言語情報を見ただけではどちらとも判断出来ない．

提案システムにおいては肯定的/否定的態度認識を用いることにより，提案に対して否定的だった場合は代替案を提案し，肯定的な場合はより具体的な提案を行うなど効率的な対話進行が可能になっている．

対話例では，ジャンルの提案に対して肯定的だった場合に，より具体的にそのジャンルに該当する店を提案している．

ここで，パラ言語の理解能力を持たない対話システムで上記の対話を実現することを考えてみる．システムの提案に対するユーザの反応と，その後のやり取りを図 12 に示す．

図中の R3’，U4 は，ユーザの最初の反応 U3’ の言語情報からでは提案に対する評価が得られないために起こるやり取りである．U3’ が，パラ言語情報により明

U1: お昼ご飯なんだけど,
　どこかいいところ無いかな?
R1: カレーなんてどう?
U2: カレーか. **(強い否定)**
R2: それじゃあ，ハンバーガーなんてどうかな.
U3: あーハンバーガーね. **(強い肯定)**
R3: ハンバーガーなら,
　近くにマクドナルドがあるよ.

図 11: 対話例.
U:はユーザの発話，R:は ROBISUKE の発話

R2: ハンバーガーなんてどうかな.
U3': ハンバーガーか.
R3': いいですか.
U4: いや，違うものがいいな.

図 12: 対話例: パラ言語理解機能がない場合

らかに肯定的/否定的な場合であっても，その理解能力を持たないシステムは必ずこのやり取りを挟まなければならない．このやり取りを避けるために，ユーザが最初の反応として「違うものがいい」と発話することも考えられる．しかし，提案の後にそのような発話を最初にするためには，反射的に起こる U3' のような発話を意識的に抑える必要があるため，ユーザは逆に煩わしさを感じ，対話は自然なものとはなりえない．

以上のようなことから，ここで作成したパラ言語の理解機能は，対話にリズムを与え，その自然性を向上させるのに大きく貢献するものと考えられる．

4. 韻律情報と言語情報を用いた相槌生成

本節では，ユーザの発話中に，適切なタイミングと内容で相槌を打つ音声対話システムについて述べる．

相槌には，相手の発話を促す効果と，相手の発話内容に対する理解度のフィードバックという機能がある．そのどちらも，適切なタイミング，適切な内容で生成することができなければ，返って不自然な対話になってしまう．本節では，タイミングと内容を韻律情報と言語情報からそれぞれ決定し，ユーザの発話に対して適切な相槌を打つ音声対話システムについて述べる．

4.1. 韻律情報を用いた相槌生成タイミングの決定

F0 とパワーをもとに計算した特徴量を用いて，相槌・復唱すべきタイミングを検出する手法について述べる．

まず，実際に相槌・復唱生成時の相手側の発話を収集するため，人同士の対面対話を収録した．5 分程度のトピックフリーの対話を 9 対話，映像と音声両方に関して収録した．動画アノーテションツール Anvil[1] を用いて，収録した動画像全てについて各話者の相槌・復唱にタグ付けを行った．タグ付けされた相槌・復唱の直前の対話相手の音声を切り出し，F0 抽出，ログパワーの計算を行い，観察した.「あー，一日目でー」という発話の抽出例を図 13 に示す．笑い声等の F0 抽出が困難になるケースを取り除いた結果を観察すると，相槌・復唱時点からおおよそ 100msec から 500msec 前の区間に特徴が現れていることが分かった．

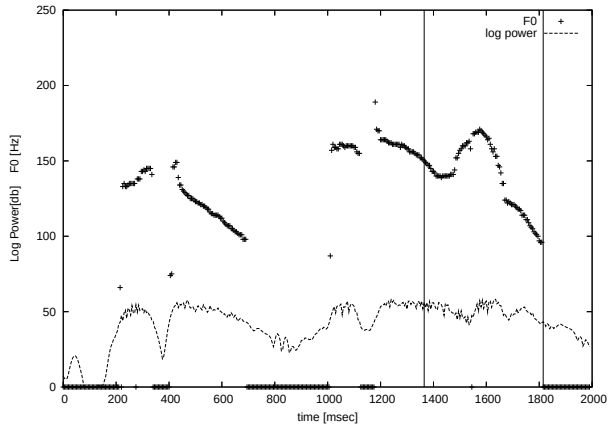


図 13: 発話の F0，ログパワー抽出例

特徴量は，過去 450msec の区間を 150msec ずつの 3 区間に分割し，各区間の F0，ログパワーの抽出結果を，最小自乗法を用いて二次曲線

$$y'(t) = at^2 + bt + c \tag{1}$$

に近似した場合の，係数 a，b と，平均誤差

$$\frac{1}{N}\sum_i |y_i - y'(t_i)| \tag{2}$$

の三種類の値を用いた． ここで y_i は抽出された F0 またはログパワー，N はその区間で抽出された値の個数を表す．本研究では，音声のサンプリング周波数 16kHz，フレームシフト 80sample で分析を行ったため，1 区間 150msec あたり N は最大で 30 となる．各区間毎に，F0，ログパワーに関する特徴がそれぞれ 3 種類ずつ計算されるため，合計で 18 次元の特徴量となる．図 13 に示した区間における F0 とログパワーの二次曲線近似の結果を図 14 に示す．

対話収録により得られた 119 サンプルの発話についてこの特徴量を計算し，正規分布を学習した．この正

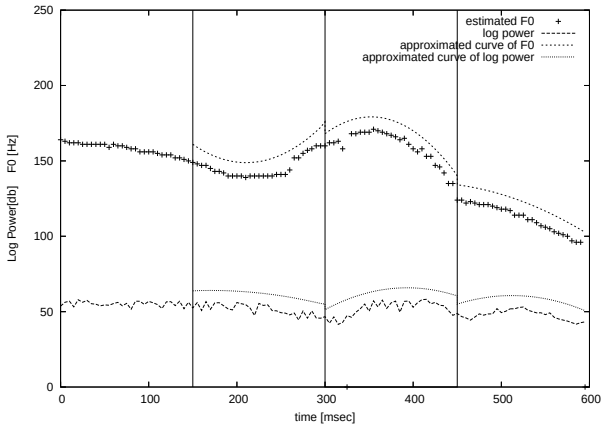


図 14: 抽出された F0 とログパワーの二次曲線近似の例

規分布が特徴量を出力する尤度がある閾値以上になった時に，相槌・復唱をするべきタイミングと判断する．

4.2. FST デコーダを用いた相槌内容の早期確定

有限状態トランスデューサ (Finite State Transducer; FST) とは，有限状態オートマトンに出力を付与したものである．複数のトランスデューサを合成 (compose) することにより，望みの入出力対のトランスデューサを構成することが可能である．また，最小化 (minimization)，決定化 (determinization) 等の最適化を行うことにより，出力の早期確定機能を実現することが可能である．大語彙連続音声認識器 (デコーダ) は，音響モデル，単語辞書，言語モデル等の複数のネットワークを用いた複雑な探索を行う．これらのネットワークを FST で表現することにより，デコーダを単純化することが可能なことから，近年，音声認識への FST の応用が注目を集めている．

音声認識の最終出力は一般的に単語列であるが，本研究では FST の早期確定機能を活かして入力文に対する相槌・復唱内容の早期確定機能を実現する．図 15 に示すように，音素を入力とした辞書ネットワークの出力である単語列をもとに，その単語列に相応しい相槌・復唱の内容を出力とするネットワークを構成する．

4.3. 対話ロボット上での相槌・復唱機能の実現

従来から開発している音声対話システムに，本研究で提案した FST デコーダと韻律処理を加えた．音声対話システムは，ROBISUKE 上に実装されている．システム構成を図 16 に示す．

ユーザが発話した音声は，音声認識に必要な MFCC 特徴量を計算するモジュールと同時に，本研究で提案した韻律処理を行う韻律処理モジュールに入力される．

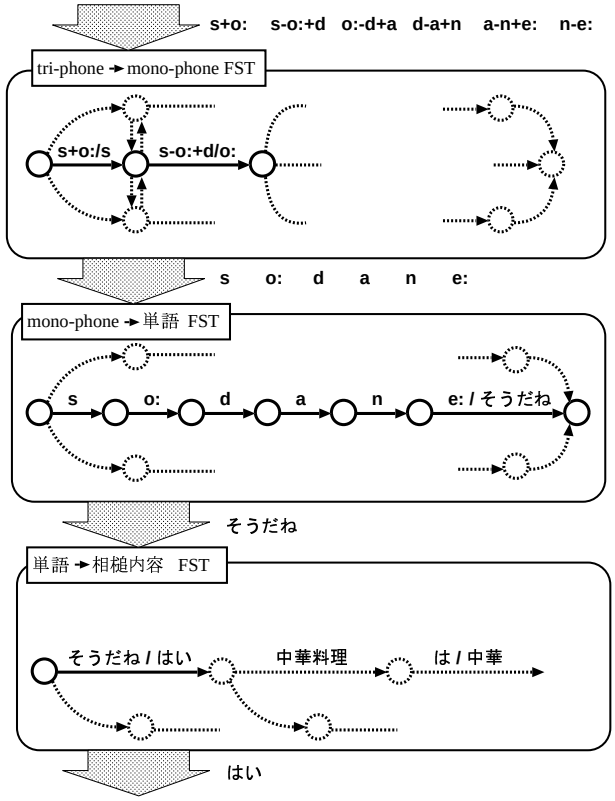


図 15: FST を用いた相槌・復唱内容の早期確定のネットワーク構成

韻律処理モジュールでは，サンプリング周波数 16kHz の音声を，フレームサイズ 1024sample，フレームシフト 80sample(5msec) で処理を行う．F0 抽出には，後藤ら [2] が提案した瞬時周波数と combfilter を用いた手法を用いている．韻律処理モジュールではさらに，発話開始から 450msec 経ち，F0，パワーの値が十分揃った状態になると毎フレーム特徴量の計算を行い，相槌・復唱タイミングの尤度を計算する．計算した尤度が閾値を超えた場合，即座に相槌・復唱発生モジュールに信号を送る．

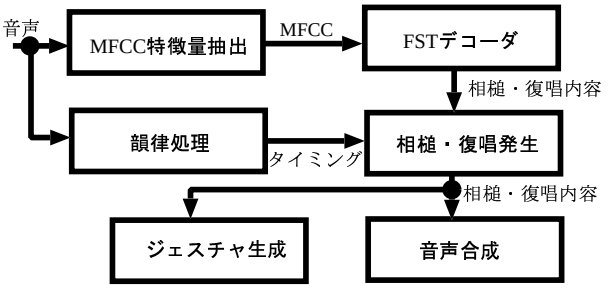


図 16: システム構成

FST デコーダは，MFCC 特徴量を受け取りデコーディングを行う．デコーディングは仮説を展開する形で進行するため，FST に早期確定機能があったとしても，発話終了まで最終的な出力を得ることができない．そこで本システムでは，一定数の連続フレームで最もスコアの高い出力が変わらなかった場合に，それを一時的に早期確定した出力とすることとした．その後にその出力と違う出力の尤度が高いことがわかった場合は，出力のキャンセルを表す特別な記号 `delete` を出力し，新たに確定したものを出力する．

相槌・復唱発生モジュールは，FST デコーダから相槌・復唱内容を受け取ると，まずその内容をバッファに保存する．そしてその後，韻律処理モジュールからタイミングが出力されると，バッファに保存された内容を音声合成モジュール，ジェスチャ生成モジュールへと送信する．

音声合成モジュールは，受け取った相槌・復唱内容を発声する．ジェスチャ生成モジュールは，相槌，復唱のどちらの場合も，首を 1 回だけ短時間のうちに縦に振るという動作をロボットに行わせる．

4.4. 評価実験

構築したシステムを評価するために，主観評価実験を行った．韻律情報を用いることと，FST デコーダを用いることとの効果を確認することがこの実験の目的である．

韻律を用いることによる効果は、FST デコーダのみを用いた場合に比べて，より適切なタイミングで相槌を打てることである．一方，FST デコーダを用いることにより期待できる効果は、ひとつにはユーザの発話内容に従って相槌の内容を変更することが可能になったことによる効果が挙げられる．また，ユーザが相槌をするべき内容を発話するまでは出力をしないことから，韻律のみを用いた場合に比べて，余計な発話を抑制する効果が得られることが期待できる．

組み合わせ 1

FST デコーダだけを用い，内容が出力された時に相槌・復唱を出力するシステム (A) と，韻律情報と FST の結果を利用して相槌・復唱を生成するシステム (B) の比較．A では韻律的な情報を無視して，相槌・復唱を行うが，B では韻律的なタイミングを考慮している．この比較で，韻律情報を用いた処理によるタイミング制御の効果が，対話の印象を向上させるかを評価する．

表 5: 組合わせ 1 の結果

A の方が良い	A の方がやや良い	同じ	B の方がやや良い	B の方が良い
3	4	0	15	10

表 6: 組合わせ 2 の結果

C の方が良い	C の方がやや良い	同じ	D の方がやや良い	D の方が良い
2	6	3	8	1

組み合わせ 2

韻律処理モジュールだけを用いて相槌を生成するシステム (C) と，韻律情報と FST の結果を利用して相槌を生成するシステム (D) の比較．C ではユーザの発話内容を無視して，韻律的に相槌を打つべき箇所の全てで，相槌を行う．この比較で，FST デコーダにより意味的に打つべきでない箇所の相槌を抑える効果が，対話の印象を向上させるかを評価する．

評価の結果を表 6，表 5 に示す．

組み合わせ 1 の結果から，韻律情報を利用し，適したタイミングで相槌を打つシステムの方が好まれることが分かる．

一方，組み合わせ 2 の結果が両極端に分かれていることから，言語情報を用いることによる効果は明らかになっていない．

5. むすび

パラ言語の理解と，適切な相槌の生成により，円滑でリズムのある対話を行うことができる音声対話システムを提案，実装した．これらから，ユーザの発話から得られる言語情報以外の情報を考慮することの重要性が確認された．また，肯定的/否定的な態度の認識を音声から行う場合や，相槌生成のタイミング決定において，韻律の果たす役割は大きいことが分かった．

参考文献

[1] M. Mipp, “Anvil - A Generic Annotation Tool for Multimodal Dialogue,” In Proceedings of of the 7th European Conference on Speech Communication and Technology, Aalborg, pp. 1367-1370 (2001).

[2] 後藤 真孝, 伊藤 克亘, 速水 悟, “自然発話中の有声休止箇所のリアルタイム検出システム,” 電子情報通信学会論文誌 D-II, Vol.J83-D-II, No.11, pp.2330–2340 (2000).