Vol.2010-SLP-84 No.23
2010/12/21

視聴覚情報を用いた 意図推定の為の決定木に基づく意図境界推定

伊藤大介[†1] 酒向慎司[†1] 北村正[†1]

計算機が人の意図を理解する事で，人と機械のコミュニケーションはより円滑なものになると期待できる．我々は，意図理解を複数のクラスに分類した意図の識別問題として扱い，自由対話からの発話意図推定手法について検討してきた．連続発話から意図を推定する為には，その意図の変化のタイミングである意図境界を推定する必要がある．本稿では，意図識別に一般的に用いられる視聴覚特徴を用いて，意図の継続と変化の特徴を決定木で学習し，未知の連続発話データに対して意図境界推定を試みる．

Decision Tree based Intention Boundary Detection for Estimation of Intention using Audiovisual Information

DAISUKE Ito,[†1] SHINJI Sako[†1]
and TADASHI Kitamuar[†1]

It can be expected that the computer understanding the user intention makes human-machine dialogue smoother. As identification problem of the intention that classified intention in plural symbols, we examined the utterance intention estimate technique from spontaneous speech. It is necessary to detection the intention border (timing of the change of the intention) to estimate intention from consecutive utterance. We detect intention boundary by using decision tree for two identifications of continuity and change with audiovisual feature which is used in the intention identification.

†1 名古屋工業大学
Nagoya Institute of Technology

1. はじめに

人間同士は意思疎通を図る為に言葉の交換を行う．しかし，本質的には同じ意味でも省略や婉曲などの多様な表現があるように，対話においては表面上の言葉だけでなく，その発話の中にある意図の伝達が重要である．また,「目は口ほどに物を言う」と言われるように，人間同士の対話では，意図を解釈する際に言語情報だけでは不十分である．言葉以外の要素としては，声の抑揚や顔の表情など，様々な視聴覚情報が意図の伝達に役立てられており，そのような情報のうち，意図の伝達に寄与する割合には諸説があるが，文献 [1] によると，話し手の意思のうち，言語情報によって伝わるものは全体の約 35 %であり，残りの 65 %は話し方，動作，表情などの非言語情報によって伝わるとされている．

音声入力型の対話システムは，機械が人の発話に対して応答を返すシステムであるが，その多くは，入力された音声を音声認識により書き起こしてから応答内容を決定している．対話システムにおいて意図を的確に捉える事で，応答生成の際に書き起こし情報だけでなく意図も考慮した柔軟な対話制御ができ，また，音声認識は話し言葉の認識が十分でないが，意図により認識誤りを修正するなどの幅広い応用が期待できる．このような背景を元に，我々は意図を対話制御に活用する事を考え，離散的なクラスとして表現し，それらを発話から得られる視聴覚特徴を用いて推定する手法を検討してきた [2].

意図の違いは予め与えられた発話区間における文頭や文末などに表れると考えられている [3, 4]．この発話区間は音声区間検出によって決められるものが一般的であるが，発話が持続している中での意図の移り変わる場合などがある為，音声区間と発話意図の継続区間は，必ず一致するわけではなく異なる部分もある．そこで，本研究では連続発話からの意図変化タイミングを検出する為，発話意図境界推定を行う．まず，対話データから音響特徴として，基本周波数 (F0) パターン・短時間平均パワーを，動作特徴として加速度・角速度センサから得られる頭部動作特徴を抽出する．次に意図の継続と変化を表すそれらの特徴量の傾向を決定木に基づいて学習する．未知の連続発話データに対して学習された決定木を用いて発話意図境界推定実験を行う事で，意図変化タイミングの検出への視聴覚特徴の有効性を確認する．

2. 発話文における意図推定

以下では，まず人間同士の意図理解機能について検討し，その上で計算機における意図理解について述べる．

Vol.2010-SLP-84 No.23
2010/12/21

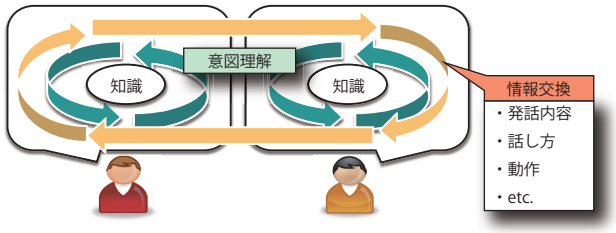


図 1 意図理解のプロセス

2.1 人同士における意図理解

人間同士のコミュニケーションにおける意図理解のプロセスとして，情報を交換する機能，得られた情報を解釈する機能の二つから構成される図 1 のようなモデルが提案されている [5]．前者は，言語や動作などの人間から発信される情報の時間変化から，意図に関係するものを抽出する特徴量抽出の問題として，後者は入力された複数の情報を総合的に解釈する識別問題として置き換えられる．

2.2 発話意図の分類

我々は意図を対話制御に活用する事を考え，離散的なクラスとしてそれらを表現する．対話は，Initiate-Response-(Follow-up) の三つにより構成されると言われている．Initiate は，新しい相手に応答を求める働き掛けの機能を持ち，Response は働き掛けに対する応答の機能を持つ．Follow-up は，現在の Initiate-Responseno のやりとりを終了させる機能や次のやりとりを促す機能がある．これらの構成要素は，役割に応じてさらに複数の発話意図に分類する事ができる．本研究では荒木らが提案している表 1 に示すような発話単位タグ [6] を発話意図とする．

本研究では，これらの意図を，人の発話から得られる音響的な特徴と加速度・角速度センサから得られた頭部動作特徴を用いて推定する事を検討してきた [2]．

3. 発話意図境界推定

ここでは，まず意図推定に関する関連研究から意図境界推定の重要性を述べ，その後本研究で行う発話意図境界推定手法について述べる．

3.1 意図推定に関する従来研究

意図や感情を識別する特徴量としては，声の高低の時間変化を表す基本周波数パターンや，声の強弱を表す短時間平均パワーがよく用いられる．文献 [7] では，発話単位基準を 400ms

表 1 発話意図の機能と詳細な分類

機能	発話意図
Initiate（働き掛け）	情報伝達，確認，提案，示唆，依頼， 未知情報要求，真偽情報要求， 約束・申し出，希望
Response（応答）	肯定・受諾，否定・拒否， 未知情報応答，保留
Follow-up（了解）	了解，相槌，復唱

以上のポーズによって決め，得られた一発話から声の高さ，声の長さ，声の強さ，声質に関する特徴量を抽出し喜びや悲しみなどの 8 種の感情の識別を行っている．文献 [4] では，発話の文頭と文末の F0 パターンから，肯定的発話か否定的発話の違いを表す特徴ベクトルを抽出し，ベイズ識別により意図の二識別を行っている．さらに，この文献では，意図の伝達を支える情報として頭部動作に着目し，画像処理によって得られた頭部動作特徴からうなずきや首のかしげを認識し，対話の進行に役立てている．

これらの従来研究では，音声認識における孤立単語認識のように一つの区切られた発話区間に対して意図識別が行われている．連続発話から逐次的に発話意図を推定する場合，どの意図かを求めるのと同時に意図の変化タイミングである発話意図境界も求める必要があるが，従来研究ではその議論はなされておらず，予め発話意図境界で区切られているか，タスクが限定的で発話意図境界が一意に決まる状態で意図識別を行っている．

そこで，本研究では，連続発話からの意図変化タイミングの検出手法を検討する．意図境界は，意図の種類と同時に求める事も可能であると考えられるが，本研究では，意図識別とは独立させ，意図変化タイミングを検出した後，その発話意図区間から意図識別を行う事を考える．以下では，自由対話における連続発話からの意図変化タイミングを検出する為の決定木に基づく発話意図境界推定手法について述べる．

3.2 決定木に基づく意図変化タイミングの検出

人が発話境界とは別に意図の境界を感知できるのであれば，意図識別に用いられる視聴覚特徴は，発話境界とは別の区間において特徴の変化が生じると考えられる．本研究では，そのような特徴量の値の変化を元に意図変化タイミングの検出を行う．まず，対話コーパスから一般的に意図識別に用いられる音響特徴や動作特徴をあるフレーム単位で抽出し，決定木により意図変化タイミングにおける特徴量の傾向を学習する．この決定木を用いる事で，未知データの各フレームに対して抽出した特徴量から意図が継続しているか変化しているかを識別できると考えられる．決定木の作成には C4.5 [8] を用いた．これはエントロピー

Vol.2010-SLP-84 No.23
2010/12/21

表 2 音響特徴

時間構造	1	無音区間の継続フレーム
基本周波数パターン	2	フレーム内の平均
	3	フレーム内の平均の変化量
	4	フレーム内の傾き
	5-7	境界前 2 フレームの平均（同様に 5,10 フレーム）
	8-10	境界後 2 フレームの平均（同様に 5,10 フレーム）
短時間平均パワー	11	フレーム内の平均
	12	フレーム内の平均の変化量
	13	フレーム内の傾き
	14-16	境界前 2 フレームの平均（同様に 5,10 フレーム）
	17-19	境界後 2 フレームの平均（同様に 5,10 フレーム）
	20-22	境界前 2 フレームの傾き（同様に 5,10 フレーム）
	23-25	境界後 2 フレームの傾き（同様に 5,10 フレーム）

基準で初期決定木を作成し，その後枝刈りを行う事で過学習を抑える事ができる．ここで，意図変化タイミングにおける視聴覚特徴は，意図の種類によって傾向が異なると考えられ，表 1 に示した全ての種類の傾向を一つの決定木で学習するのは困難だと考えられる．そこで，本研究では，発話意図の機能別の特徴の傾向を考慮する為に，機能に応じて分類した働き掛け，応答，了解のそれぞれで個別に決定木を作成した．

3.3 発話意図境界推定に用いる特徴量

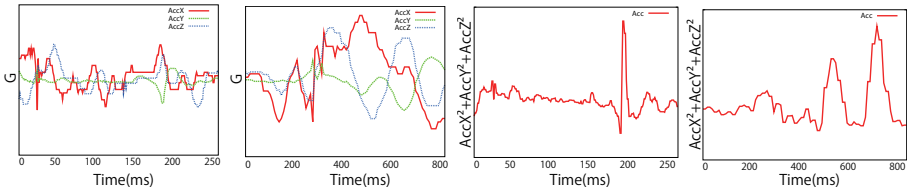
3.3.1 音 響 特 徴

意図の変化は，声の高低や強弱に瞬時に表れると考えられる為，音響特徴として基本周波数パターンと短時間平均パワーを用いた．基礎検討により，表 2 に示す 25 種類を意図境界推定の特徴量として用いた．ここで，無音区間の継続フレームは F0 から判断し，無音区間が続いているフレーム数をカウントしたものを特徴量とした．

3.3.2 頭部動作特徴

人は，相手に働きかける際や応答を返す際に，頷きや首をかしげたりする事がある．これらは視覚的な意図の表現として考える事ができ，本研究ではこのような頭部動作特徴も意図境界推定に役立てる．3D モーションセンサから得られた加速度・角速度値を，音響特徴と同様にフレーム分析を行い頭部動作特徴とする．

加速度・角速度値では，各軸から現在の姿勢やしぐさを推定する事ができる．しかし，意図境界の前後で頭部が動くと仮定すると，意図変化タイミングを検出する為には，姿勢やしぐさの情報に加えて，軸に依存しない頭部動作の強弱を見る事も重要であると考えられる．そこで，本研究では，X,Y,Z の各軸の各加速度を二乗和したものも特徴量として用いた．図



(a) 聴取時の加速度値　(b) 発話時の加速度値　(c) 聴取時の加速度の二乗和　(d) 発話時の加速度の二乗和

図 2 否定・拒否「ふとるよ」

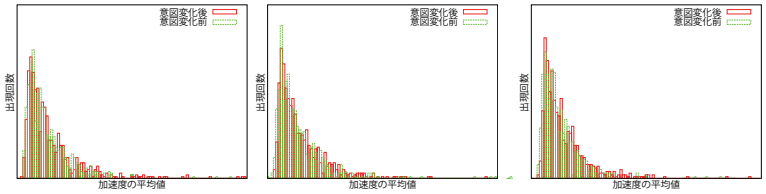
表 3 頭部動作特徴

加速度の二乗和	26	フレーム内の平均
	27-28	境界前 5 フレームの平均（同様に 10 フレーム）
	29-30	境界後 5 フレームの平均（同様に 10 フレーム）
	31	フレーム内の傾き
	32-33	境界前 5 フレームの傾き（同様に 10 フレーム）
	34-35	境界後 5 フレームの傾き（同様に 10 フレーム）
加速度	36-38	フレーム内の X 軸の値の平均（同様に Y 軸，Z 軸）
角速度	39-41	フレーム内の X 軸の値の平均（同様に Y 軸，Z 軸）

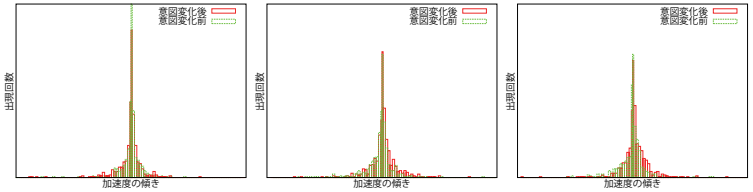
2 に，聴取時と発話時の各加速度値と，それらを二乗和したものの例を示す．二乗和をとる事により，聴取時に比べ，発話時に頭部動作が起きている事が明確にわかる．

また，実際に決定木学習に使用する特徴量を決める為に基礎検討を行った．図 3 に全 330 個の意図変化タイミングの前後のある一定区間における加速度の二乗和のフレーム内平均と傾きの分布を示す．この結果から，若干ではあるが，意図変化前と後で特徴量の傾向が異なっている事がわかる．加速度の二乗和の平均値では，どのフレーム長においても意図変化前より変化後の方が大きい値に分布している．加速度の傾きではフレーム長が短い時は分布の違いはほぼ見られないが，フレーム長を長く取る事で，意図変化後において大きい値に分布する割合が大きくなっている．つまり，頭部は意図変化前よりも変化後に動作が起こりやすいが，急に動き出すわけではなく，意図変化後から少しずつ動きを速めている事がわかる．これらを踏まえて本研究では，表 3 に示す 16 種類を音響特徴に加えて意図境界推定に用いた．

Vol.2010-SLP-84 No.23
2010/12/21



(a) 平均値（フレーム長:0.2sec)(b) 平均値（フレーム長:0.5sec)(c) 平均値（フレーム長:1sec）



(d) 傾き（フレーム長:0.2sec）(e) 傾き（フレーム長:0.5sec） (f) 傾き（フレーム長:1sec）

図 3 加速度の二乗和の平均値と傾きのヒストグラム

3.4 データ収録

4. 意図境界推定実験

連続発話データから意図境界推定実験を行う．まず，実験を行う為に収録した対話データについて述べ，その後発話境界推定の結果を推定精度と実際に学習された決定木の観点から考察する．

自然な意図を含んだ発話を得る為，台本の無い自由対話による対話収録を行った．本研究では，レストランにおいて二名のうちどちらかが相手に晩御飯を御馳走するという設定の下，二者間の自由対話を収録した．その際，メニューを見ながら対話を行う事で，自然な対話を模倣した．収録の際には，ヘッドセットマイクに 3D モーションセンサ（NEC Tokin 製 MDP-A3U9S）を接着し，音声と同時に加速度・角速度値を取得した．収録したデータの規模を 4 に，実際に行われた対話例を表 5 に示す．また，

4.1 実 験 条 件

本研究では，以下の 3 通りの特徴量で決定木を学習し識別性能を比較する..

- 音響特徴：25 種類
- 頭部動作特徴：17 種類（頭部動作特徴及び無音区間の継続フレーム）

表 4 収録データ

収録データ	3 分× 8 セット
収録人数	4 人
総発話意図	330 発話

表 5 対話例

話者	発話内容	意図
A	じゃあサラダ食べよ	提案
B	どのサラダ？	未知情報要求
A	どれがおいしそう？	未知情報要求
B	シーザーサラダ	未知情報応答
A	うん	相槌
B	かーシーフードサラダ	未知情報応答
A	シーザーそんなに好きだっけ	真偽情報要求
A	俺シーザーサラダ上位好きだよ	肯定・受諾
B	じゃあシーザーで	了解

- 視聴覚特徴：41 種類（音響特徴及び頭部動作特徴）

学習前に，まず各フレームに対して主観により継続か変化かのラベルを付与する．意図変化タイミンのフレームにおいては，意図の機能を表す働き掛け，応答，了解の 3 種に分類し，それ以外のフレームを継続とする．決定木の学習の際には，分類された個々の意図変化タイミング以外の変化タイミングのフレームを学習データから除く事で，意図の機能の種類を考慮した 3 種の決定木を作成する．ここで，主観によるラベル付けでは明確な意図変化タイミングを決める事はできない為，意図変化タイミングの前後 4 フレームを意図変化タイミングとする事で正解点をぼかしている．

評価は，以下のステップで行う．

(1) 評価データの各フレームに対して，意図の構造別に作成した 3 種の決定木を用いて継続か変化の二識別を行う．

(2) 4 フレームを 1 セグメントとして，セグメント内で多数決を取る (同数の場合は変化).

(3) セグメント単位で，3 種の決定木のうち一つでも変化と識別された場合変化，そうでない場合を継続として扱う．

(4) セグメント単位で正解ラベルと照合する．
  - 正しく意図変化を検出できた場合：正解
  - 継続を変化と誤検出した場合：挿入誤り

実験は話者毎に行っている．その他の実験条件を表 6 に示す．

4.2 発話意図境界推定結果

2 名の話者に対して発話意図境界推定を行った．各話者における評価データに対する正解数と挿入誤り数を**図 4** と**図 5** に示す．

全体的な特徴として，フレーム周期が短いと正解数が極端に少なくなり，挿入誤りが極端

Vol.2010-SLP-84 No.23
2010/12/21

表 6 実験条件

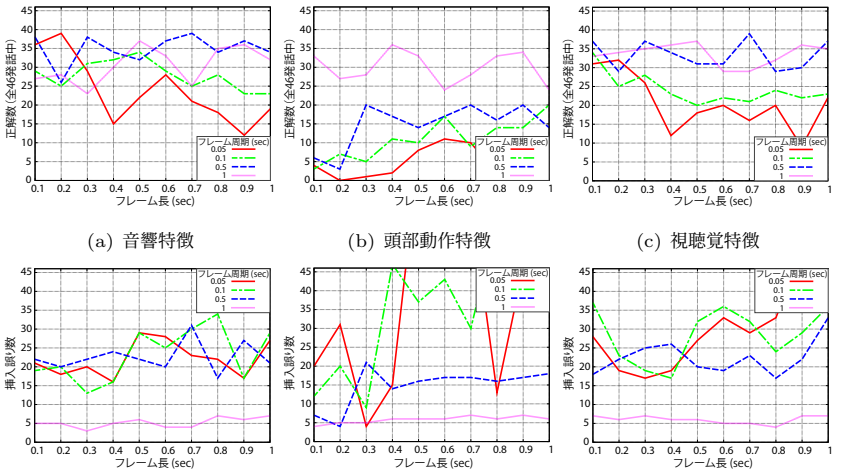
学習データ	3 分間の発話データ
評価データ	学習に用いていない 3 分間の発話のデータ
フレーム周期	0.05，0.1，0.5，1
フレーム長	0.01，0.1，0.2，0.3，0.4，0.5，0.6，0.7，0.8，0.9，1

に増える傾向がある．これは，変化の学習データ数が常に一定であるのに対し，フレーム周期が短くなるにつれ，継続の学習データ数が増えてしまう為，両者のバランスが取れず，決定木の枝刈りが上手く行えていない為であると考えられる．また，フレーム長の変化は推定精度にあまり影響を及ぼしていない．これは，特徴量として，複数のフレーム内の平均や傾きも見ている為，フレーム長が短い場合には，決定木に充分に長いフレーム数から得られる特徴量が反映されている為であると考えられる．また，どの話者においても，3 種類の特徴量の中では頭部動作特徴を用いたものが一番悪い結果となっている．これは意図境界推定に音響特徴の方がより有効であるという結果を示しているが，それ以外にも主観によるラベル付けの際に，音声のみの情報から境界を決定している事が原因として挙げられる．ラベル付けをビデオなどによる動作特徴も考慮して行う事で，頭部動作を用いた際の推定精度も向上すると考えられる．

次に，話者毎に見ていくと，話者 1 では，音響特徴に頭部動作特徴を加える事で，最高の正解数こそ変わらないものの，全体的に推定精度が低下した．この話者は音響特徴のみで高い推定精度を出している為，頭部動作特徴を加えた事で決定木が必要以上に複雑になってしまい，推定精度が低くなってしまったと考えられる．それに対し，話者 2 では，音響特徴に頭部動作特徴を加える事で，推定精度が向上している．話者 2 においても音響特徴のみで高い推定精度を出している為，この推定精度の向上は頭部動作特徴の有効性を示すものであると考えられる．これらの結果をまとめると，発話境界推定に用いる特徴量として，F0 やパワーなどの音響特徴は誰に対しても有効であるが，頭部動作特徴は個人に依存しやすく，有効に働く場合と悪影響を及ぼす場合があるとわかる．

4.3 意図別に推定精度の比較

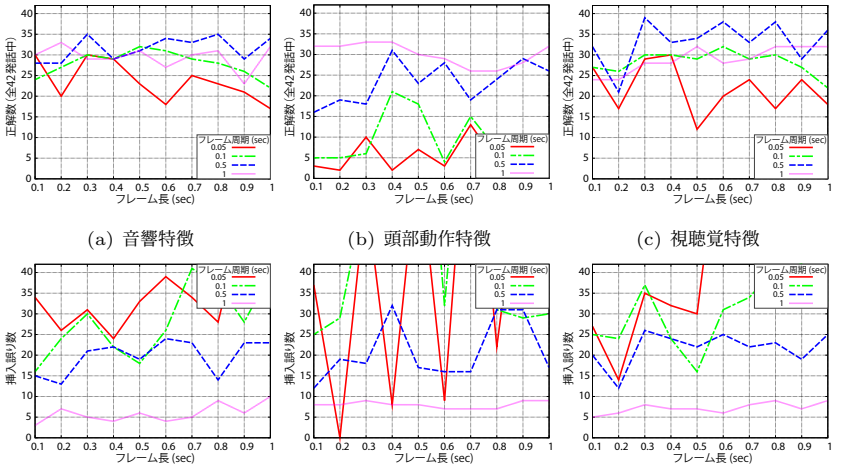
また，意図別の推定精度の比較を行った．16 個の意図に対して特徴量毎に正解数を算出し，比較を行った．その結果，3 通りの特徴量間で結果に大きな違いは見られず，どの意図も高い正解率を示した．これは，働き掛け，応答，了解という意図の構造の種類に従って 3 種の決定木を作成した為，全ての意図の特徴を考慮して決定木学習が行えたと考えられる．また，今回は実験の規模が小さい為，今後は大規模なデータによる実験によって，意図毎の



(a) 音響特徴　(b) 頭部動作特徴　(c) 視聴覚特徴

(d) 音響特徴　(e) 頭部動作　(f) 視聴覚特徴

図 4 正解数及び挿入誤り（話者 1）



(a) 音響特徴　(b) 頭部動作特徴　(c) 視聴覚特徴

(d) 音響特徴　(e) 頭部動作特徴　(f) 視聴覚特徴

図 5 正解数及び挿入誤り（話者 2）

傾向を調べる必要がある．

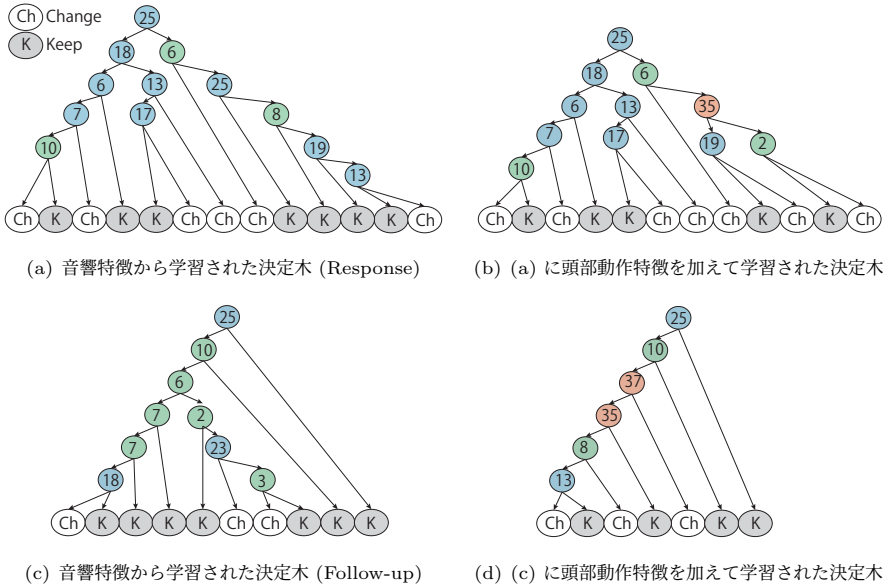
4.4 作成された決定木の考察

実際に学習された決定木について考察する．ここでは、音響特徴に頭部動作特徴を加える事が有効に働いた場合を見る為，全ての特徴である視聴覚特徴を用いた際に最も正解率が高くなった条件下で学習された決定木を図 6 に示す．ここで，各節に記述してある数字は表 2 及び表 3 で示した特徴量番号である．ここでは構造別に学習された 3 種の決定木のうち応答と了解である 2 種の決定木を示しているが、どちらの場合においても頭部動作特徴を加える事で，決定木がより簡潔に表現できている事がわかる．これにより，音響特徴のみで学習された決定木には，まだ無駄な素性が含まれ，それらの素性が頭部動作特徴に置き換わる事で，推定精度が向上したと考えられる．また，各決定木に反映された特徴量を見てみると，どの決定木においても，当該フレームのみから抽出された特徴量よりも，前後の複数フレームから抽出された特徴量が多く選択されている．これは意図変化タイミングの前後で音響特徴や頭部動作特徴は急激に変わるものではなく，意図境界推定にはできるだけ長い区間の特徴を見るべきである事がわかる．

5. む す び

本稿では，連続発話からの意図変化タイミングの検出を行う為に発話境界推定を行った．

特徴量として，F0 やパワーなどの音響特徴とセンサから得られる頭部動作特徴を用いて，意図が ‘継続か変化しているかを識別する為の決定木を学習した．推定実験により，音響特徴のみで，高い正解率を示し，また個人によっては頭部動作特徴を加える事でさらに高い正解率を示した．しかし，いくつかの条件下においては挿入誤りが極端に増えてしまう問題があり，特徴量を抽出する為の分析区間や，挿入誤りを抑制するような新たな特徴量の検討が必要である．また，本手法により得られた発話意図境界から実際に意図推定を行う事などが課題として挙げられる．



(a) 音響特徴から学習された決定木 (Response)　(b) (a) に頭部動作特徴を加えて学習された決定木

(c) 音響特徴から学習された決定木 (Follow-up)　(d) (c) に頭部動作特徴を加えて学習された決定木

図 6 学習された決定木の例

参 考 文 献

1) Marjorie Fink Vargas, 石丸 正 訳, ：非言語コミュニケーション，新潮選書，(1987).
2) 伊藤 大介, 酒向 慎司, 北村 正：発話文における発話意図識別に有効な音響特徴の検討, 第 9 回情報科学技術フォーラム, (E-015), pp.235–236 (2010).
3) 松本 宗也, 傅松 明, 白井克彦：音声対話システムにおける発話意図推定, 社団法人情報処理学会, pp.2-141–142, (2008).
4) 藤江 真也, 江尻 康, 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, 信学論, (D-Ⅱ), Vol.J880D-Ⅱ, No.3, pp.489–498, (2005).
5) 佐藤 知正，西田 桂史，市川 純理，畑村 洋太郎, 溝口 博：ロボットによる人間の意図の能動的理解機能，日本ロボット学会誌，Vol. 13, No. 4, pp.545–552, (1995).
6) 荒木 雅弘, 伊藤 敏彦, 熊谷 智子, 石崎 雅人：発話単位タグ標準化案の作成, 人工知能学会誌, Vo.14, No.2, pp.251–260, (1999).
7) 有本 泰子，河野 宏美，大野 澄雄，飯田 仁：感情音声のコーパス構築と音響的特徴の分析，電子情報通信学会 2006 年総合大会講演論文集，pp.S42–S43，(2006).
8) Quinlan, J. R.: C4.5:Programs For Machine Learning, Morgan Kaufmann Publishers (1993).