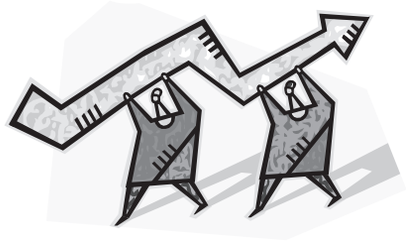
解 説

音声言語インタフェースのための情報処理学会試行標準



新田恒雄　豊橋技術科学大学　nitta@tutkie.tut.ac.jp

松浦　博　(株)東芝　hiroshi.matsuura@toshiba.co.jp

西本卓也　東京大学　nishi@hil.t.u-tokyo.ac.jp

西村雅史　日本アイ・ビー・エム(株)　nisimura@jp.ibm.com

音声言語インタフェースの標準化とは

IT 分野における標準化活動の役割が大きくなっている．背景には，情報端末とシステムの多くがネットワークに繋がる時代に入ったこと，同時に，IT サービスに必要な技術が高度化・複雑化したことがある．ネットワーク上に分散する異種混交（heterogeneous）なハード／ソフトコンポーネントを組み合わせて，より複雑なシステムを実現するには，コンポーネント間のインタフェースを標準化しておく必要がある．また，音声言語を利用するといった新しいサービスを普及させるには，ユーザとシステム間のインタフェースを標準化することも重要になる．

情報処理学会(以下，本会)には，国際標準化機構(ISO)および国際電気標準会議（IEC）のもとで，IT 全般に関する国際標準化活動を行う合同技術委員会に対応する情報規格調査会が設けられている．しかし，IT に関する標準化への我が国の貢献は，一部の分野を除いて大きくないのが現状である．そこで，日本発標準の促進，国際標準における日本の貢献の活発化を目指して，試行段階の標準化活動を支援する本会試行標準制度が 2001 年秋からスタートした[1]．成果は本会の Web で公開され，国内外の規格化作業に役立つことを目指している．

本稿で紹介する音声言語インタフェースは，学会試行標準専門委員会（委員長石崎俊慶應義塾大学教授）の WG 4 小委員会において検討している．委員会では，最初に標準化目的と対象テーマを関連企業の協力を得て討議した[2]．標準化への要望には，利用する立場によって次に示すさまざまなものがある．

(a) 音声処理エンジン(LSI，ボード，ソフト)開発者：多様なアプリケーションで利用できるよう，種々のデータ形式／ファイル形式／ API を標準化したい．性能の比較と向上のため評価方法も標準化したい．

(b) アプリケーション開発者：用途に合ったエンジンを自由に選択したり，エンジン変更時にもアプリケーションは変更のないようにしたい（そのためには (a) で挙げた API の標準化が重要）．音声の専門知識がなくともアプリケーションが開発できるようにしてほしい．エンジン性能からアプリケーション組み込み時の性能を予測できないか．

(c) エンドユーザ：異なるアプリケーション間で違和感なく，音声インタフェースを利用したい．そのためにも使用方法は共通化してほしい．また，音声や言語辞書の登録方法も標準化してほしい．

図 -1 に，音声言語インタフェースが抱える課題の例を，ユーザと開発者の視点に分けて示した．委員会では，以上に述べた要望・課題のうち，標準化を急ぐべきテーマとして，次の候補を取り上げて討議した．

(1) ユーザが新語（未知語）を登録する際に使用する「読み」表記の標準化

(2) 音声関連製品の取扱い説明書などで使用する「用語」の標準化

(3) アプリケーションに特化した標準化（例：操作コマンドや対象の呼称に対する「読み」の統一，評価方法に対するガイドラインなど）

(4)対話記述言語の標準化

(1) と (2) は，これまで各社が独自に対応してきたが，委員会で調査し 6 社の比較表を作成したところ，すべての企業に共通する表記・表現が少ないことが明らかになった．特に，(1) はエンドユーザに混乱を招くため，早急な対応が必要とされ，最初の試行標準化対象となった．(2)については，討議結果をもとに産官学における，この分野の有識者へアンケート調査した後，試行標準案をまとめることになった．

次に（3）は，音声認識応答システムを含むCTI（Computer Telephony Integration），ディクテーション，カーナビゲーション（以下カーナビと呼ぶ），PDA（Personal Digital Assistance）などの携帯端末を対象に，操作コマンドや対象の呼称に対する「読み」の統一と，評価方法のガイドラインが検討された．このうち「読み」に関する標準化は，多くのアプリケーションで共通化が困難とされ，当面，標準化を見送ることとなった．理由は，コマンドの場合，同じアプリケーションでも製品により機能やその意味が異なること，あるいはすでに長年使用されていることなどであった．また，対象の呼称では，検索したい目的地の読みや，楽曲の読みなどで問題が大きいことが共通に認識されたが，音声インタフェースを手掛ける企業とコンテンツを提供する企業が異なることが問題とされた．この課題は作業量が厖大となるが，音声対話技術の発展のためにも，第三者機関が本格的に取り組むことが望ましい．一方，ディクテーションは音声入力の基本的な応用であり，標準化の対象範囲を「ディクテーション中に入力する（キーボード上の）記号に対する読み」に絞って作業することになった．市販ソフトウェアを調査したところ，製品によって読みがかなりばらついていることが分かったこともある．

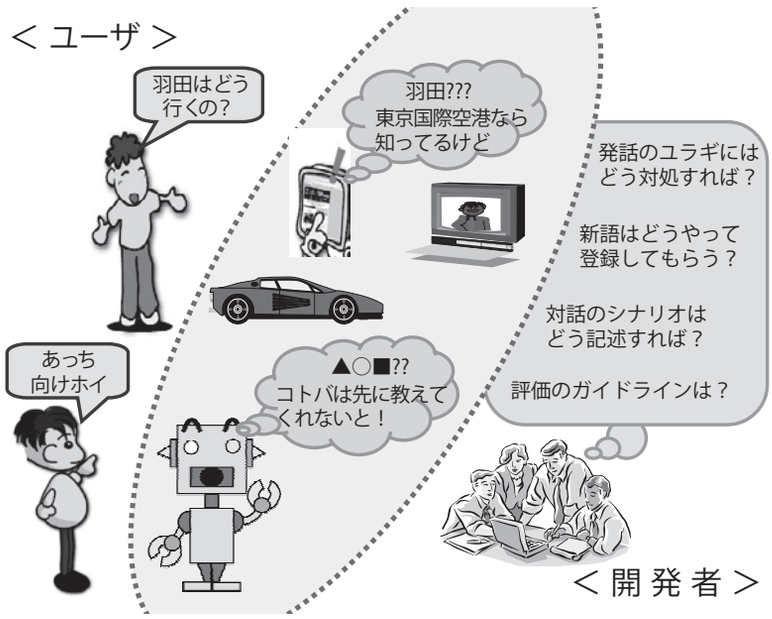


図-1　音声言語インタフェースが抱える課題

（3）のアプリケーションに特化した標準化，その中でも「読み」に関する標準化は，上に述べたように多くの困難な問題があり，基本的なもの以外は当面の作業をあきらめざるを得なかった．他方，評価のガイドラインでは，文を入力するため評価方法が容易なディクテーションや，古くから評価方法が検討されてきたCTIと比較し，カーナビの音声入力評価方法が問題とされた．カーナビでは使用状況が複雑多岐にわたる上，操作方法，コマンド名称も各社まちまちのため，統一的な性能評価が困難である．また，騒音下のハンズフリー大語彙連続音声認識といった非常に難しいタスクが対象となるため，環境の影響を大変受けやすく，正しい手順に従って評価することが特に重要である．標準化作業にあたっては，メーカごとの不公平が生じることがないよう，評価方法だけでなく評価対象や単語リスト選定などについても，多くの時間を割き議論が行われた．

以上の標準化作業は，評価のガイドランを除くと日本語の特殊性にかかわるものである．一方，国際標準における日本の貢献を活発化する活動に（4）がある．音声対話を記述する言語は，W3C（World Wide Web Consortium）[3] のVoice Browser WGが策定するVoice XMLが知られている．また音声を含むマルチモーダル対話（MMI：Multi-modal Interaction）もW3C MMI WGで策定作業が進められている．これに対して我が国では，Webサービス以外に携帯端末，カーナビ，情報家電など，音声言語インタフェースにかかわる先導的分野が少なくない．そこで委員会では，音声対話を含むMMI記述言語の試行標準作業を，平成18年から開始することとした．

表-1に委員会の現構成メンバを，**表-2**に標準化を終えWeb上に公開したテーマを示す．以下では，これらを順に紹介するとともに，次の標準化目標であるMMI対話記述言語の概要について述べる．

新語の読みを登録するには[1),4)]

音声は日常的に利用される便利なメディアではあるが，コンピュータとのインタフェースに利用する場合，いくつかの問題がある．その1つが未知語の問題である．人間同士では，知らない単語が出てくるとその意味（語意）を推測したり，質問することで自分の記憶に入れる．音声認識ソフトウェアの場合にも，システム開発者は必要な単語をあらかじめ登録するが，すべてをカバーすることはできない．そこで，エンドユーザが未知語を登録する機能，あるいは登録されているものとは異なる発音を登録できる機能を提供している．これにより，コンピュータはユーザの定義した未知語を含む発話を認識し，結果を表示し，音声で応答することが可能になる．

エンドユーザは，新語に対して表記（表示文字列：たとえば「最高値」）と，どのように発音したいか（読み表記：たとえば「さいたかね」）を登録する．登録方法がソ

主 査：	新田恒雄	豊橋技術科学大学
委 員：	赤堀一郎	（株）デンソー
	石川　泰	三菱電機（株）
	磯谷亮輔	日本電気（株）
	伊藤克亘	名古屋大学（現在，法政大学）
	大淵康成	（株）日立製作所
	河村聡典（←金澤博史←松浦　博）	（株）東芝
	國枝伸行	松下電器産業（株）
	外山聡一	パイオニア（株）
	西村雅史	日本アイ・ビー・エム（株）
	西本卓也	東京大学

表 -1　音声言語インタフェース委員会の構成

- ✓ 音声認識のための読み表記
  （担当　松浦　；　IPSJ-TS 0004）
- ✓ ディクテーションで用いる基本記号に対応する読み
  （担当　西本　；　IPSJ-TS0009）
- ✓ カーナビ用音声入力の性能評価のためのガイドライン
  （担当　西村　；　IPSJ-TS0011）

表 -2　試行標準化を終えたテーマ

フトウェアごとに異なると，エンドユーザに戸惑いと不便を与えることになる．実際に，音声認識の応用範囲が拡大するに従って，読み表記の不統一が混乱を招く恐れが出てきている．

今回まとめた学会試行標準は，エンドユーザが利用する簡易形式として，「音声認識のための読み表記」を最初に規定している．この規定では「経営：けいえい」，「ユーザ：ゆーざー」のように平仮名で読みを表記する．**表 -3** に例を示した．この中で「まるの・うち」とある中点は，長音化しないことを示す区切り記号である（この使用はオプション）．

一方，システム開発者が発音をより詳細に記述できる形式として，「音声認識のための詳細読み表記」を規定した．こちらは**表 -4** に示す例のように，片仮名で「経営：ケーエイ(あるいはケイエーなど)」，「秋田：ア(キ)タ」（括弧内は母音の脱落や無声化を表す），「三月：サンカ゜ツ」（カ゜は鼻濁音を表す）とより詳細な表記が可能になっている．詳細読み表記の標準化にあたっては，認識結果を音声合成によりトークバックする（音声で知らせる）場合を考慮し，「日本語テキスト音声合成用記号の規格（JEIDA-62-2000）」と連携できるよう，ほぼ同一の内容とした．

委員会では，音声認識読み記号の並び方に関する規則や制約についても討議を行ったが，ルールの複雑化と利用時の困難から各記号の並び方に関する規定は特に設けないこととした．また，「読み表記」に対する的確な用語は，英語の訳（今回は sounds-like symbols とした）も含めて今後の検討事項である．なお，（社）電子情報技術産業協会（JEITA）では，ここに紹介した試行標準を元に検討を加えた後，JEITA 規格 IT-4003 として，「日本語音声認識用読み記号」の規格をまとめている．学会試行標準の目的の 1 つである，迅速な活動と Web 公開による標準化作業支援がうまく機能した例といえよう．

東京	とうきょう（to-kyo-, toukyou）
武蔵新城	むさししんじょう（musashishinnjo-, musashishinnjou）
丸の内	まるの・うち（marunouchi）
丸の内	まるのうち（maruno-chi, marunouchi）
大阪阿倍野橋	おおさか・あべのばし（o-sakaabenobashi）
経営	けいえい（ke-e-, keiei）
経営	け・いえい（keie-, keiei）
ユーザ	ゆーざー（yu-za-）…… 外来語も平仮名で記述
秋田	あきた（ak (i) ta）…… 記号列から無声化を推測
鼻血	はなぢ（hanaji）
続く	つづく（tsuzuku）
私は	わたしわ（watashiwa）

（注）（　）内は，ある音声認識エンジンが読み表記を解釈し，内部表記に変換した一例を示している（規定ではない）．

表 -3　読み表記の例

東京	トーキョー（to-kyo-）
丸の内	マルノウチ（marunouchi）
大阪阿倍野橋	オーサカアベノバシ（o-sakaabenobashi）
経営	ケーエー（ke-e-）
経営	ケイエー（keie-）
私を	ワタシオ（watashio）
私が	ワタシガ（watashiga）
私が	ワタシカ゜（watashinga）
デッドボール	デッドボール（deqdobo-ru）
デッドボール	デットボール（deqtobo-ru）
秋田	ア（キ）タ（ak (i) ta）
ストップ	（ス）トッ（プ）（s (u) toqp (u)）

表 -4　詳細読み表記の例

読み上げ途中の記号はどう発音すれば…[1), 5)]

音声で文を入力する場合，キーボード上のさまざまな記号（？，／，＊，「，」…）に対応する読み（たとえば“？”ならクエッションマーク，クェッション，はてな，…）

基本記号	読　み	
,	こんま	
.	ぴりおど	どっと
:	ころん	
;	せみころん	
?	くえっしょんまーく	
!	びっくりまーく	
^	はっと	
～	から	
~	ちるだ	

基本記号	読　み	
＿	あんだーばー	
ー	まいなす	はいふん
／	すらっしゅ	
＼	ばっくすらっしゅ	
｜	たてぼう	
＋	ぷらす	
＝	いこーる	
＜	しょうなり	
＞	だいなり	

基本記号	読　み	
¥	えん	
$	どる	
%	ぱーせんと	
#	しゃーぷ	
&	あんぱさんど	あんど
*	あすたりすく	
@	あっとまーく	
(空白)	すぺーす	
'	くおーてーしょん	
"	だぶるくおーてーしょん	

基本記号	読　み	
(	かっこ	
)	かっことじる	
[	だいかっこ	
]	だいかっことじる	
「	かぎかっこ	
」	かぎかっことじる	
、	てん	
。	まる	
・	なかてん	

表 -5　ディクテーションに用いる基本記号の読み

を規定する必要がある．市販の日本語ディクテーションシステムを調査したところ，各社とも複数の読みを登録するなどの対応を行っていたが，共通の読みが存在していないことが少なくなかった．将来，ディクテーションシステムの利用が日常化すると，基本記号に対する読みの不統一がユーザに混乱を招く恐れがある．

読みのゆらぎを考慮すると，標準化規定にも幅を持たせる必要があるが，既在のセットから短時間に読みを選定することは大変難しい作業である．委員会は選定の基本方針を決め，これに沿って試行標準を策定した．具体的には，まず対象に関しては：

① 日本語キーボードから入力可能なもののうち，英数字，漢字，かたかな，ひらがななどの文字を除いたものとする，

②「改行」など，操作の名称は含まない，

③ 視覚的形状が同じ場合は 1 つにまとめて扱う，

（例）「－」と「-」（全角および半角のマイナス）は同じ，一方，「～」と「~」（全角「から」と半角チルダ）は視覚的形状が異なるなど．

また，読みの選定にあたっては：

④ 既存のシステムで広く使用され，覚えやすく読みやすい，

⑤ 関連する基本記号の読み同士に一貫性がある，

（例 1）意味的な対応を考慮し，「ぷらす」と「はいふん」でなく，「ぷらす」と「まいなす」を用いる．

（例 2）「XX まーく」は，「くえっしょんまーく」，「びっくりまーく」，「あっとまーく」などのように統一．

（例 3）「XX かっこ」と「XX かっことじる」は同じ形式を用いる．（「だいかっこ／だいかっことじる」「かぎかっこ／かぎかっことじる」）

⑥ 1 つの読みは 1 つの基本記号に対応する，

ことを考慮して決定した．

表 -5 に今回標準化した基本記号と対応する読みを示した．外来語を語源とする読みは，多くのバリエーションがあるため，特に次の方針，「語源となる単語に置き換えた後，広く用いられる単語を選択し，さらにその単語に対して広く用いられる読みを選択する」を設けて読みを決定した．詳しくは規定を参照されたい [1]．表には，既存アプリケーションに依存した例外から，複数の読みを許容しているものがある．まず「．（ぴりおど）」は URL の入力などで「どっと」と呼ばれること，また「－（まいなす）」は電話番号や URL の入力などで「はいふん」が使われるためである．「& （あんぱさんど）」は，一般になじみが薄いため特に「あんど」を許容した．なお，1 つの読み（たとえば「てん」）が入力された時，どの視覚的形状の基本記号（「，」（全角）もしくは「, 」（半角）など）を対応させるかは，アプリケーションに依存する．

試行標準に適合するディクテーションシステムは,「基本記号」に対して，少なくとも規定の「読み」を認識するよう設計する必要がある．しかしその他，独自の「読み」を認識しても構わない．

カーナビ音声入力性能を比較するには…[1), 5)]

カーナビの音声入力は，十字カーソルで目的地を選ぶように階層をたどることなく，すぐに入力が完了するといった利点を持つ．しかし，使用環境が多岐にわたるうえ，操作法やコマンド・目的地の呼称が各社まちまちなため，統一的性能評価が困難であった．共通のガイドラインが試行標準として提供されると，各社の認識装置が本来持つ能力を比較することが可能になる．このことは，カーナビメーカだけでなく音声認識装置開発各社，およびエンドユーザにとっても有益と考えられる．

委員会は，性能評価方法を示すガイドラインとして,「簡易評価方法」と「基本評価方法」の 2 つを規定した．「簡易評価方法」制定の目的は，異なる機種間の性能比較を可能にし，雑誌社など第三者による評価に裏付けを与えることにある．一方,「基本評価方法」の目的は，カーナビメーカがこの評価基準で認識率を測定し，結果をパンフレットに記載するなどの利用を想定している．なお，システムのユーザビリティ評価は本ガイドラインには含まない．

ガイドラインは，性能評価の際に守るべき一定の基準を規定している．実験条件では，実車に搭載したシステムを使用すること，車は停車しエンジンはアイドリング状態とすること等々を指定している．また実験手順では，**図 -2**（簡易評価手順の場合）に示すように，現在地の設定, POI（Point of Interests：関心地点）入力方法の事前確認，発話者の事前訓練，性能評価の方法および正解・不正解の判断基準を与えている．基本評価の手順は，POI リストに対して 100 地点以上を選定するとしていることと，話者数を簡易評価の男女各 1 名に対して男女各 10 名以上としている以外は，ほぼ同じである．ただし，POI リストについてはサイズが実験ごとに異なるのが普通であろう．そこで「基本評価方法」は，リストサイズを加重平均値として求め併記することを推奨している．

音声入力の評価を POI の認識率のみで行うことにした理由は，他の詳細住所入力や電話番号入力などでは，カーナビの機種によりポーズを入れる位置が異なるなどの事情があり，現状では認識性能の対等な比較が困難と判断したことによる．

「簡易評価方法」では，誤った使い方により，誤った評価が行われることを避けるため，評価時の負担を減らすさまざまな工夫をした．また，自車の現在地を“新宿－東京都庁”に固定し，多くのカーナビが入力可能と考えられる東京周辺の施設名称 152 カ所を評価用 POI リストとして提供している．POI リストには，駅，ゴルフ場，遊園地，公園，ホテル，ホール，等々が選定されている．

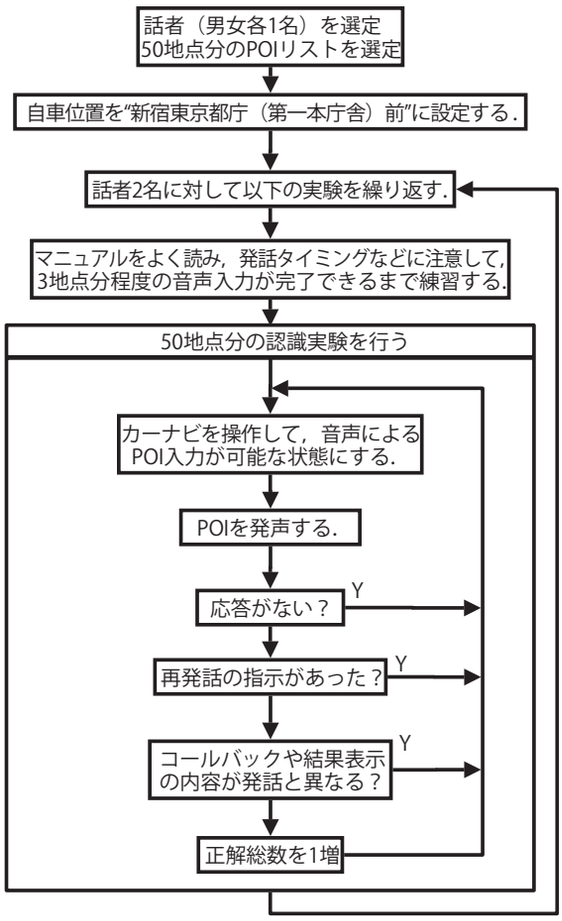


図 -2　カーナビ音声入力の簡易評価手順

音声（マルチモーダル）対話をどう記述するか

ディクテーションでは，汎用の統計的言語モデル（N-gram）を新聞記事などから作成し，入力文中の単語を予測することで精度良い認識を達成している．ただし，文中に未知語があると，認識結果はしばしば入力からは想像だにできないものとなる．一方，音声対話では個々のタスク（案内，予約，検索，購入，etc.）とドメイン（予約タスクの場合，航空チケット，新幹線，ホテル，etc.）に応じて，統計的言語モデルを作成する必要があるが，この作業コストに見合うサービスは少ない．そこで多くの場合，対話を直接記述している（この中で，単語とその並び（文法）を指定）．音声認識エンジンは，対話管理システムから対話の中で使用する語彙と文法をそのつど

```
<?xml version ="1.0"?>
 <vxml version="2.0">
  <form>                          ユーザから変数を得る
   <field name="dest">            変数を特定する
    <prompt>                      入力を促す発話
     行き先をどうぞ
    </prompt>
    <grammar src="dest.xml"/>     文法を規定
   </field>
   <filled>          変数の値が埋まった後の処理を記述する
    <submit next="http://www.dest.example/dest2.jsp"/>
   </filled>
  </form>
 </vxml>
```

図 -3　VoiceXML による音声対話記述の例

受け取り，単語候補を探索する．

音声対話を記述する言語は，これまで各社が独自のスクリプト言語などを定義し使用してきた．音声対話を閉じたシステムで利用する際には，こうした実現方法も可能であった．しかし，今日の Web サービスのように，ネットに開かれて提供されるサービスを音声対話で利用するには，共通の言語が必要になる．W3C では，主に CTI（Computer Telephony Integration）を対象に音声対話記述言語 VoiceXML を標準化している [3]．**図 -3** は，システムが行き先を尋ねた後，ユーザからの発話を代入する対話例を示している．

一方，音声入出力は比較的新しいインタフェースのため，これまでの画面を中心としたインタフェースと共存する方策が必要である．そこで，両者を融合したマルチモーダル対話（MMI：タッチ，音声など複数の入力インタフェースを持つ対話）が検討されている [6]．Web 上のサービスが MMI に対応するようになると，携帯端末やカーナビから画面タッチと音声インタフェースの双方でアクセスできるようになる．また，ディジタル TV の Web アクセスも，リモートキーと音声インタフェースの双方から利用できるようになるだろう．

MMI 記述言語は，現在，W3C の MMI WG で標準化の検討が行われている [6]．これまでのところ，複数の記述言語（VoiceXML, SMIL, SCXML, InkML など）を組み合わせた compound documents で MMI を記述する方向が討議されている．国内では，本会音声対話技術コンソーシアム（ISTC：代表新田恒雄（豊橋技術科学大学））[7] の中で，SIG/MMI WG が連携した討議を行っている．これまでに，国内関連企業の技術者が参加し，MMI を利用したユースケースと要求仕様などをまとめ，Web 上に公開した．**図 -4** は，この WG で検討中の MMI の階層構造を示している．委員会では，WG のメンバに参加していただき，合同で試行標準化作業を進める予定である．

MMI は，Web アプリケーションからロボットまで，近未来のヒューマンインタフェースの中核技術である．しかし，この実現には要素技術から統合技術まで多くの困難

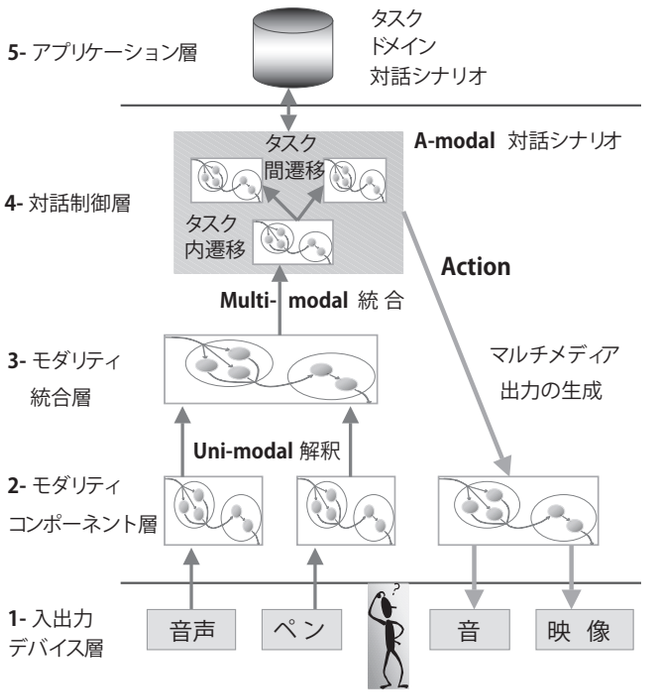


図 -4　マルチモーダル対話の階層

が横たわっている．そこで，試行標準も段階的に行うことを考えている．たとえば，図 -4 の 3 層と 4 層の間で，適切なインタフェースを定義すると，現状の情報端末から MMI を利用して Web アクセスできるようになる．一方，タスク，ドメインごとに参照する（Web からダウンロードする）文法記述中の意味タグも共通化が必要になろう（< 出発地 > 東京 | 横浜 | … < / 出発地 > といった記述から，地名が出発地（到着地）であることが区別できるなど）．さらに本格的な対話には，タスクにまたがる対話制御（入力された複数のタスクを仕分ける／前の発話を修正する／…）にかかわる記述方法も検討する必要がある．

最後に，音声言語処理インタフェース委員会は，今後とも，音声，言語，ヒューマンインタフェース研究者の方たちの幅広いご支援をいただきながら，「役立つ試行標準」を提供していきたいと願っている．

参考文献

1）情報処理学会試行標準 Web サイト
http://www.itscj.ipsj.or.jp/ipsj-ts/index.html

2）新田恒雄，石川　泰，伊藤克亘，畑岡信夫，松浦　博，磯谷亮輔，西村雅史，西本卓也：音声言語情報処理に関する情報処理学会の試行標準策定活動，情報処理学会研究報告 2002-SLP-40-10, pp.57-60（2002）．

3）World Wide Web Consortium Web サイト　http://www.w3.org/

4）松浦　博，西本卓也，金子　宏，磯谷亮輔，石川　泰，西村雅史，伊藤克亘，新田恒雄：音声認識読み記号および音声関連ソフトウエアに係わる用語の試行標準案，情報処理学会研究報告 2003-SLP-45-11, pp.65-70（2003）．

5）西本卓也，西村雅史，赤堀一郎，石川　泰，磯谷亮輔，伊藤克亘，大淵康成，金澤博史，國枝伸行，外山聡一，新田恒雄：音声認識応用に関する学会試行標準，情報処理学会研究報告 2005-SLP-55, pp.47-52（2005）．

6）新田恒雄：マルチモーダル対話の深化と記述言語の今後，情報処理学会研究報告 2004-SLP-50, pp.15-22 (2004).

7）情報処理学会音声対話技術コンソーシアム（ISTC）Web サイト
http://www.astem.or.jp/istc/

（平成 18 年 2 月 28 日受付）