ヒューマンインタフェース 97−10
音声言語情報処理 40−10
（2002. 2. 1）

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

新田 恒雄[1] 石川 泰[2] 伊藤 克亘[3] 畑岡 信夫[4]
松浦 博[5] 磯谷 亮輔[6] 西村 雅史[7] 西本 卓也[8]

1豊橋技術科学大学大学院工学研究科 2三菱電機情報技術総合研究所 3産業技術総合研究所
4日立中央研究所 5東芝研究開発センター 6NECマルチメディア研究所
7日本IBM東京基礎研究所 8京都工芸繊維大学工芸学部
〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘1-1 E-mail : nitta@tutkie.tut.ac.jp[1]

あらまし： 本報告では，平成14年より発足した情報処理学会試行標準専門委員会下のWG4小委員会：音声言語処理インタフェースの活動内容を紹介する。小委員会では，(A) 音声認識・合成，(B) 音声言語に関するデータベース表記（メタデータほか），および(C) Voice BrowserやMulti-Modal Browserにおける音声対話に関する試行標準を策定することを目的としている。本文では，(A)と(B)に焦点を当て，現状と課題を述べる。

キーワード： 音声言語処理，標準化，データベース，音声認識，音声合成，音声対話

Activity of IPSJ Trial Standard WG4 (Spoken Language Interface)

Tsuneo NITTA[1], Yasushi ISHIKAWA[2], Katunobu ITOU[3], Nobuo HATAOKA[4],
Hiroshi MATSU'URA[5], Ryosuke ISOTANI[6], Masahumi NISHIMURA[7], and Takuya NISHIMOTO[8]

1 Toyohashi University of Technology, 2 Mitsubishi Electric Corp., 3 AIST, 4 Hitachi Ltd.,
5 Toshiba Corp., 6 NEC Corp., 7 IBM Japan Ltd., 8 Kyoto Institute of Technology
1-1 Hibariga-oka, Tempaku, Toyohashi, 441-8580 JAPAN E-mail : nitta@tutkie.tut.ac.jp[1]

Abstract: This paper describes the activity of IPSJ Trial Standard WG4 that has the objectives of standardization for spoken language interface. In WG4, three areas are targeted for IPSJ trial standard, that are (A) speech recognition and synthesis, (B) description framework of spoken language database, (C) description language of spoken dialogue system. In this report, we discuss some standardization problems in Japan by focusing on the areas of (A) and (B).

Key words: Spoken Language Processing, Standardization, Database, Speech Recognition, Speech Synthesis, Spoken Dialogue

1．はじめに

平成14年より情報処理学会に学会試行標準専門委員会（委員長石崎俊慶応義塾大学教授）が正式に発足した。同時に，この委員会の下でWG4小委員会：音声言語処理インタフェースが発足している。学会標準の目的は次の二点にある。

(a) 国際標準を成立させるには長時間を要するので，準備段階のものを学会として標準化する。

(b) 国際標準の基礎となるデータで，国際標準として制定が難しかったり，国際標準になじまないものを学会標準とする。

学会試行標準は，情報処理学会のWebで公開され，国内外の規格化作業に役立てて頂くことを目指している。なお，内容は原則3年程度で見直すことになっている。音声言語処理インタフェース小委員会では，(A) 音声認識・合成，(B) 音声言語に関するデータベース表記（メタデータほか），および(C) Voice BrowserやMulti-Modal Browserにおける音声対話に関する試行標準を策定することを目的としてスタートした。以下では，このうち(A)と(B)に焦点を当て，現状と課題をまとめた。

2．音声認識・音声合成の標準化

音声認識・音声合成は，長い研究の歴史を経て，現在，実用化の時代を迎えている。しかし，研究開発をさらに進め，また利用を促進させるには，単に技術内容の進展だけでなく，種々の標準化・共通化を同時に図る必要がある。以下では，音声認識・合成に関する標準化の目的と動向を概説した後，検討すべき課題を列挙する。

2.1 標準化の目的

音声認識・音声合成に関する標準化の目的は，利用対象者から以下のように分類することができる。

(1) 音声認識・音声合成エンジンの開発者

音声認識・音声合成エンジンが多くのアプリケーションで利用されるには，種々のデータ形式，ファイル形式，APIの共通化が必要である。また，アプリケーションから要求される性能を把握して，複数エンジン間の比較が正当に行えることも必要になる。問題点を把握するため，共通の性能評価手法も重要である。

(2) アプリケーション開発者

音声認識や音声合成エンジンを使用するアプリケーション開発者にとっては，用途に合ったエンジンを自由に選択でき，エンジン変更時にもアプリケーションに変更の必要がないことが望ましい（そのためには，最低，データ形式，ファイル形式，APIが共通化されている必要がある）。エンジン性能からアプリケーション組込み時の性能を予測したいという要求もある。最終的には，後述するエンドユーザにとっての利便性を配慮し，音声認識・合成を組込んだアプリケーションを普及させて，利用分野を拡大することが目標である。

(3) エンドユーザ

エンドユーザが音声認識・合成のアプリケーションを利用する場合，利用方法が共通化され，異なるアプリケーション間で違和感なく円滑に利用できることが重要になる。そのためには，音声言語辞書などのリソースが共通になっている必要がある。アプリケーションの性能を正しく把握できる性能評価法の標準化も重要である。

2.2 音声認識・音声合成の標準化の動向

音声認識・音声合成の標準化は，これまでにも検討されてきた。国内では，（社）電子情報技術産業協会（以下JEITA）の前身である（社）電子工業振興協会（2000年11月に(社)日本電子機械工業会と統合）において，80年代初頭から継続的に検討されている。海外では，欧州の音声処理プロジェクトEagles (the Expert Advisory Group on Language Engineering Standard)が，学術的見地から検討を進めた例がある[1]。最近では，VoiceXML Forum [2]や，その検討結果を受けたW3C [3]の音声ブラウザ用マークアップ言語，さらにマルチモーダルへの拡張を検討するSALT（Speech Application Language Tags）[4]など，業界団体での取り組みが目立つ。一方APIについては，プラットフォームへの依存性などの理由から，特定の企業が規定するAPIへのデファクト化が進んでいる。しかし現状では，上述した

活動によって標準化が進展した部分があるものの，未着手の課題もまだ多い。音声処理技術は研究の歴史は長いが，製品としては萌芽期からやっと成長期に入った段階にある。このため業界主導の標準化では，利害関係から作業が遅れる，もしくは詳細規定がなかなか決まらないといった弊害が懸念される。また評価手法などについては，研究的な要素も大きく，技術の発展と普及には，学会と業界団体が協調して検討を進める必要がある。

2.3 標準化の課題

(1) よみ記述

音声認識，音声合成にとって最も基本的な課題は，音声の表記である。日本語では，かな漢字記述，かな表記，ローマ字表記，音素記号表記などが考えられている。音声認識，合成用辞書を記述する場合，1) 表記コード・表記法，2) 有効な範囲，3) 記述形式が，検討項目となる。例えばかな記述は，一般には分かりやすい表記である。しかし，音声認識辞書に登録する際，「佐藤」に対して「サトウ」「サトオ」「サトー」からどれを選ぶか，使用可能な外来音の範囲をどう決めるか，カタカナにしかない「ヴ」などのコードは？といった問題がある。また，無声化・鼻濁音・撥音などの表記には，ローマ字表記が適しているが，日本語のローマ字表記にもゆらぎがある。さらに IPA（国際音声記号）を扱おうとするとコンピュータでの扱いが困難で，これを ASCII コードで表現する SAMPA などを使用した場合，専門知識が必要になる，音声学的記述が必ずしも音声認識や合成で有効とは限らないなど問題も多い。JEITA では，音声合成用の記述として，かなレベル，ローマ字，異音レベルの表記を定めたが[5]，音声認識との共通化が可能か，またユーザが記述する辞書表記とエンジンレベルでの表記の問題，さらにはマルチリンガルを視野に入れた場合など今後の検討課題は多い。

(2) 合成用表記

音声合成では，読みの他，アクセント記述や音質制御のための記述が必要になる。主な課題には，1) アクセント記述，2) 文章中に埋め込む発話速度など合成音を制御するタグ，3) テキスト合成の言語処理部の出力結果に相当する中間言語の記述がある。制御タグについては，JEITA，各種 API，マークアップ言語策定の中でも検討されているが，モーラなど日本語特有の情報表現の扱いなどの課題がある。

(3) 記号の読み

PC 向けディクテーションソフトが，各社から販売されている。しかし，“(“などの記号や「改行」をどのように読み入力するか等，標準化に絡む問題は少なくない。音声合成においても，これらをどう発音するかが問題となる。

(4) 音声認識用言語モデル

音声認識用の CFG などの文法，N-gram など統計的言語モデルの記述やファイル形式を標準化する必要がある。

(5) 音声対話システムの利用方法

音声認識を利用するアプリケーションでは，メニューをすべてユーザに開示する GUI と異なり，「何を言っていいか分からない」ために利用できない場合が多い。アプリケーションが普及すればユーザが慣れるという見方もあるが，利用方法が分からないための語彙外発話は，結果的にシステムの「認識率が低い」と判断され，普及の大きな阻害要因ともなりかねない。特に利用が拡大し始めたカーナビや音声ポータルでは，深刻な問題となる。自己表現性が高く，特に誤認識時にユーザがシステムを容易に制御できる I/F を構築するには，利用方法を共通化しておくことが望ましい。システムの制御用語，対話の基本構成，プロンプトなどの共通化が課題である。

(6) 評価手法

音声認識・音声合成の評価法は，2.1 で触れたように，製品普及にとって，また研究開発者自身にとって（他の手法と比較して問題点を正確に把握しなければならない等）究めて重要である。単語音声認識であれば，評価単語やデータの共通化に的を絞ればよいが，ディクテーションや音声対話の評価[6]は困難な課題が多い。また符号化と異なり，原音声のない音声合成の評価も簡単ではない。単に難しさを強調した議論をするだけでなく，課題や問題点を技術的に詰めながら明確化し，作業を進めることが大切である。

3．音声言語データベースに関する標準化

1990 年代以降，記録媒体容量の増加と計算機性能の向上といったハード面での後押しと，確率モデルの普及とその膨大な学習データ処理の必要性というソフト面からの要請が相まって，わが国においても様々な音声言語データベース，あるいはコーパスが整備されつつある。このような状況のもとで，一つの研究機関，もしくはプロジェクトで整備したデータベースを広範囲に利用できるようにするには，そのデータベースがどういうものであるか（メタデータ）を第三者に分かるように記述する必要がある。そこで以下では，データベースをより広く流通させるために役立つ記述とはどのようなものかを検討する。

これまで広く使われてきた日本語音声言語データベースに，単語データベースや音素バランス文データベースがある。これらは主に，音素モデルの学習に使われてきたが，その際によく利用された標準として，電子協の音節一覧やATR の音素バランス文セット[7]がある。一方，音響的な条件については，標準的な記述方法がなかったため，同一マイクを利用するなどの解決策[8]が採られてきた。

音響的諸条件については，実際に処理してはじめて，目的に合致しないものであることがわかることもあり，利用可能かどうかを知るすべが必要であろう。収録条件が統一，もしくは記述されていないデータベースは，利用価値も低いのではないだろうか。収録環境に関する記述方法は定まったものがないため，流通している多くのデータは，雑音の少ない環境だけで収録されたものである。今後，音響的な面で，より幅広いコーパスが構築されるために必要な記述項目として，以下のようなものが考えられる。

1) 音響環境：部屋（諸次元，残響時間），騒音特性
2) 収録環境：マイクロホン（単一指向／無指向／近接マイク／携帯電話，モノラル／ステレオ／ラインマイク，F特，S／N），設置位置（音源との位置関係含む）
3) 音源情報：話者（男女，年齢層，出身地，現居住）
4) デジタル信号情報：サンプリング（周波数，ビット数），符号化方式（GSM / XXCELP 等）

これらには，既存のデータベースでも（異なる形式ではあるが）記述されている項目も含まれている。したがって，「役立つ試行標準」の議論には，これまでの記述方法で十分か，十分でなければどのような記述がよいかを含めて検討する必要がある。

これまでに広く使われてきたコーパスは，読み上げによるものがほとんどであった。今後は対話コーパスの必要性が高まるであろう。これに関しては，人工知能学会言語・音声理解と対話処理研究会の「談話・対話研究におけるコーパス利用研究グループ」がタグなどの標準化の活動を進めてきており[9]，今後，当該 WG との連携が重要である。

4．おわりに

音声言語処理インタフェース小委員会活動は，スタートしたばかりであるが，今後，音声，言語，ヒューマンインタフェース研究者の方達の幅広いご支援を頂きながら，順次，「役立つ試行標準」の提供を行っていきたい。

参考文献

[1] D.Gibbon, et.al. “Handbook of Standards and Resources for Spoken Language System”, Mouton de Gruyter (1997)

[2] http://www.voicexml.org

[3] http://www.w3c.org/

[4] http://www.saltforum.org

[5] 赤羽,板橋,”音声合成システムの性能評価方法”, 日本音響学会秋季講演論文集,pp.215-218（2000）

[6] 石川,”音声対話システムの評価法”,音響学会誌 54 巻 11 号,pp.807-811（1998）

[7] 磯,渡辺,桑原,”音声データベース用文セットの設計”, 日本音響学会春季講演論文集, pp. 89-90 (1988)

[8] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka,T. Kobayashi, K. Shikano, S. Itahashi,” JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research”, J. Acoust. Soc.Jpn,Vol. E20, No.3, pp.199-206 (1999)

[9] 土屋,堀内,石崎,前川,”音声対話コーパスの共有化へ向けて”, 人工知能学会論文誌, Vol.14, No.2, pp. 231-242 (1999).