音声言語情報処理 32−15
（2000．7．15）

2000年代における音声言語情報処理の研究課題

新田　恒雄　（豊橋技術科学大学　大学院工学研究科）

1．はじめに

2000年代には，情報処理の対象がbit-streamからcontents中心に移行する。contentsすなわちマルチメディア(MM)オブジェクトは，話し言葉，できれば音声言語主体で操作するのが自然である。また，次世代のMMサービスは，新しいUI環境と対話スタイルを必要としており，音声言語を中心としたマルチモーダル対話(MMI)に期待が寄せられている[1]。しかし，多様なメディアを統合制御しながら，効率良く人と対話するシステムの実現には多くの課題が横たわっている。本文では音声言語情報処理技術の現状を踏まえながら，その将来と研究課題について述べる。

2．音声言語情報処理技術の現状

音声認識・合成アルゴリズムの実現に専用ハードウエアを必要とした時代から，DSPファームウエアの時代を経て，'90年代後半にはソフトウエアのみで解決可能な時代が到来した。これに伴い研究－試作－商品化のサイクルが短縮され，新機能・高性能の音声関連ソフトが登場し始めている。

■音声認識：　英語版ディクテーションソフトに続き日本語版ソフトが登場した。成功の要因としては，音響モデル（HMM）および言語モデル(n-gram 文法)の双方に，強力な確率統計手法を支援する環境（音声・言語コーパスの整備，PC基本性能の向上ほか）が整ったことが大きい。一方，このことは現状のソフトウエアが，レール（音声・言語コーパス収集環境）から外れた発話に対して弱いことも意味している。

■音声合成：　波形重畳方式等の採用により，それまでの合成方式と比較して格段に音質のよい合成音が実現された。特定の話者に絞って多様な音韻環境を持つ波形素片を収集し，接続歪みの少ない素片を選択する方式が成功に繋がったが，同時に柔軟性（個人性・感情などの付与能力）を失った。

■音声インタフェース：　音声サービスに適した専用システム以外では，マウス操作とテキスト／画像表示によるGUIの時代が続いている。現在は，本格的なマルチモーダル対話実現のための基礎技術（マルチメディアのオブジェクト化，HIエンジンの高性能化，モダリティ統合化ほかの技術）を蓄積する時代にある。

3．音声言語情報処理技術の将来と研究課題

著者が企業で音声研究を始めた頃であるが，上司から「音声屋はexcuseが多過ぎる」とよく叱られたものである。この言葉は現在の音声研究者についても通用すると思われるが，眼を音声言語研究の地平線に向けて先へ先へと覗いてみよう。

(1) 言語音声が記述対象：　例えば現在のディクテーションは，音声をモノモーダル（言語）と仮定し，諸々の環境には全て眼をつぶった。

(2) 音声全体が記述対象：　音声の持つ多様なモダリティの中で「言語音声」を捉える。音声認識であれば，音声の他のモダリティ（個人性（男女・年齢・出身・性格…），感情，…）を特定することで，ロバストな音声認識に到達できるかもしれない。

(3) 音全体の記述：　音環境の中の「言語音声」を考える。聴覚情景分析のように環境音を同定したり，カクテルパーティ効果の機能を実現できると音声の応用は格段に広がるだろう。

(4) 言語情報に関わる実世界全体の記述：　実世界が発信する多様な言語情報（モダリティ）の中での「言語音声」を考える（注：ここでモダリティとは「受け手の意識するしないに関わらず，実世界（対話空間）が発信する知覚可能な情報」と定義する）。実世界（の発信する言語情報）からコトバ（の意味）を聴き取ることができれば，ロバスト性の向上に留まらず，コトバの意味を多

面的に扱えられるため，送り手の意図をより精確に解釈できるようになるだろう。

(5) 実世界全体の記述： 実世界が発信する多種多様なモダリティ（情報洪水！）の中の「音声言語」を考える。実世界から直接コトバ（の意味）を選択的に聴き取ることは，究極の音声言語情報処理技術であろうか？

もとより，技術は与えられた課題（バー）を乗り越えていれば良い（＝実用化できる）。ただ excuse が何処から来て，それがどのような影響を与えるかを音声研究者は知っておく必要がある。

音声認識： 人間は，離れた位置でかつ横を向いて話し掛けても，さらに高騒音の上に他の声が聞こえるような場所でも，目標とする音声を聞き分けることができる。これに比べると，音声認識に組込まれた現在の音響モデルは音環境に敏感すぎる。ロバストな音声処理方式と音響モデルを段階的に構築する必要がある。将来の音声対話応用では，音響モデル作成時と異なるマイクロホンを使用し，多少離れた位置から発話しても性能劣化が少ない音声認識ソフトを提供したい。一方，言語モデルでは，書き言葉から話し言葉への拡張（間投詞，助詞落ち，語尾の長音化，倒置，言い直し，言い淀み等への対応）が大きく進展すると期待される。また，制約の少ない自由発話に対しては，トピックが頻繁に変わるため，分野別言語モデルの整備と共に，トピックの追跡・適応といった新しい機能が必要になる。 一方，音の世界は音声だけでなく，より豊かな世界を含む。これらは聴覚情景分析の中で研究が始められているが，音声もいずれ音全体に対する意味記述の枠組みの中で，認識－理解されるようにならなければならない。

音声合成： 現在の朗読調合成音は音質は良いが柔軟性に著しく欠けた。音声対話には音質は多少悪くても，感情や個人性を表現できる音声合成方式が必要である。合成音と顔の表情が調和し，協調して動作することにより，擬人化エージェントの表現能力は格段に向上する。また，現状の音声合成ソフトウエアは，まず日本語解析に耐える文を入力する必要がある（テキスト－音声変換）。これに対して，自由対話では必ずしも文の形ではなく，意味形式を入力すれば応答が得られる柔軟な合成方式が望ましい（概念－音声変換）。

音声・マルチモーダルインタフェース： ここでは音声入出力と他のモダリティとの統合，およびMMオブジェクトとの直接対話の二点について展望と課題を述べる。

[A]マルチモーダル対話に関して

－ 音声出力と他のモダリティとの統合では，まず音声メディアに含まれる言語情報（音韻・韻律・プロミネンス等）と様々な非言語情報（話者の感情，態度等のモダリティ）との統合利用が始まる。また，映像メディアに含まれるモダリティとの統合では，擬人化エージェント（表情・動作）とのスムースな同期提示が数年内に実現してくるだろう。

－ 音声入力と他のモダリティとの統合では，上記出力における統合利用を追いかける形で，音声メディアに含まれる言語情報と非言語情報との統合利用が始まる。また，映像メディアに含まれるモダリティとの統合では，眼 (gaze) を中心とする顔画像からの意図抽出，ジェスチャとの統合，さらに非言語情報（動作，身体特徴，周囲空間情報等のモダリティ）との統合が進展すると期待したい。

[B]マルチメディアオブジェクトとの直接対話に関して

－ 画像世界のオブジェクト化が進展するのと並行して，これらマルチメディアオブジェクトと（音声言語で）直接対話可能な世界を構築することが大きな課題になる。コトバによる表現を如何に獲得するかという困難な課題に対しては：

step.1- 注釈(connotation)を含む大規模なMM対話コーパスの収集整備

step.2- マルチメディアオブジェクト世界のontology表現の整備

の二つが当面重要であると考えている。これらの研究は，今後，他の研究会と連携しながら段階的に，しかし精力的に進めていく必要がある。

参考文献： 新田：GUIからマルチモーダルUI(MMI)に向けて，情報処理学会誌, vol.36, No.11, pp.1039-1046 (1995-11).

Targets for Spoken Language Processing and Speech Interfaces : Synthesis

Nick Campbell

ATR Spoken Language Translation Laboratories,
Kyoto 619-02, Japan (nick@slt.atr.co.jp)

Abstract

This paper presents a personal view of the challenges facing speech technology during the first ten years of the 21st century. It suggests that whereas the basic component technologies are already well developed, there is still much work to be done before machines will be capable of the level of expression and understanding that is required for natural spoken interaction and for the comfortable transfer of digital information via voice. The paper focuses particularly on the challenges facing speech synthesis.

1. Spoken Language Processing

Both speech recognition and synthesis have reached a level where they are now appearing in a range of practical applications, and can be found on sale to the general public for low prices as software packages in computer stores. However, while people typically convey more than words when they speak, current speech recognisers and synthesisers are still only capable of processing the basic verbal content. They are not yet able to handle the full range of information carried by the human voice

Contrary to original predictions dating from the middle of the last century, general-purpose reading-machines are not yet found in common use. Instead, speech synthesis is being used in customer-care services, in car navigation systems, and in other such limited-domain applications. The technology for signal generation is mature, but the text-understanding problems involved in producing speech from writing remain difficult to solve. We are still not capable of the level of machine intelligence that is required to convert written text into meaningful speech-sized chunks, partly because the media differences require significant reformulation of the content before it is appropriate for an audio channel - even in formal presentations, the text of a natural spoken utterance is rarely the same as that of its written equivalent.

Speech recognition is similarly limited to word-level processing and takes little consideration of the way the component words are spoken, i.e., of the additional information that is signalled by the speech prosody. This may be adequate for basic limited transactions such as buying a ticket or booking a hotel room, but not for processing conversational-style speech, in which as much information is provided about the state of the discourse as about the content of the utterance. In natural spoken interaction, a human agent will adjust his/her speech in order to maintain a balance in the discourse and to satisfy social constraints of the dialogue, matching the needs of the listener as well as simply providing the required information. This requires an understanding of the way that an utterance is spoken, in order to efficiently process its verbal content.

We can suppose then, that if future speech processing is to be more *user-friendly*, it will also begin to take into account the non-verbal information which is a characteristic of human speech. For speech synthesis, this requires a higher level of voice quality and an expression of inferred meaning that current systems are still incapable of generating. It also requires a specification of the intent as well as of the content of each utterance.

2. Speech Synthesis

Before making predictions about the next decade of speech synthesis, it may be wise to look first at the trends which have emerged over the past quarter of a century. Although significant developments have also been made in text processing technology, we will focus here mainly on changes in the production of prosody and the voice.

2.1. Twenty-five years of progress

The most significant changes have reflected computing capacity, and have influenced the very basic concepts of speech synthesis. Dudley's Voder [1] was an extremely large small-memory machine, but when work began on MITalk [2], the electronic computer was already almost personal, albeit slow, limited in memory, and expensive. Now even notebook computers have a capacity and speed that far exceed their predecessors.

Figure 1 illustrates the effects that these changes have had on the approaches to generating computer speech, particularly with respect to the relationship between rules and data. MITalk (Fig 1.a) was almost entirely rule-based, relying on analysis of external databases for the generation of knowledge about speech. Phoneme formant targets and their transitions were derived from spectral analysis, and their prosodic modifications were determined by experiment from carefully prepared data. Apart from the sparcity of computer memory, neither the large speech corpora nor the tools to process them were available at this time, and the modelling of speech by rule was still deemed possible.

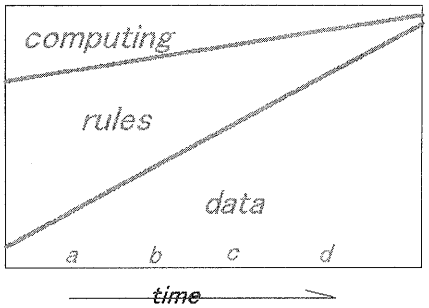

Figure 1: trends in speech synthesis systems

By the mid-eighties, several hours of speech material were available online, enabling better modelling and also better naturalness. However, this improved naturalness was not obtained as a result of the modelling; it was largely gained by the inclusion of segments of actual recorded speech (diphones, demi-syllables, non-uniform units, etc) to replace the rule-generated waveforms. Rule-based formant transitions could be replaced by their natural counterparts, reducing the modelling at the cost of increased memory consumption. ATR nu-talk [3] (Fig 1.b) illustrated this hybrid approach; relying on a database of 5000 words for its segmental content, and then subsequently modifying the prosody by rule. The increased complexity of the signal resulted in better quality speech, showing the models to be an over-simplification of the complex information in speech.

CHATR [4] (Fig 1.c) took further advantage of the availability of natural speech data and computer memory to also include the prosodic variations found in fluent speech. By thus eliminating signal processing, at the cost of an even-larger source corpus, almost human naturalness was achieved in the quality of the synthesised voice. As with the diphone approach, the increase in quality was gained by a knowledge of which variations can not be well modelled by rule or manipulation, in conjunction with a fortuitous increase in available computer storage memory.

It is interesting to note that whereas these developments made increasing demands on computer memory, they resulted in a significant reduction in computation. We can predict from these trends that future speech synthesis systems will continue to include more data and will perform even less processing in order to meet the demands of high-quality voice production.

2.2. The next ten years

The evolution from rules to data will affect all three stages of synthesis processing; including text and prosody as well as signal.

NATR [5] (Fig 1.d) will make further use of the extended source data to eliminate rule-based prosodic prediction by using direct-selection of speech units according to context-specific feature labels. Current systems use the large database as a source of knowledge from which to train prosodic models, but it now makes sense to eliminate this prediction stage and use database information directly, because using physical values as targets encourages both prediction and selection errors.

Furthermore, it is time to accept that prediction of meaning from text is difficult, even for humans, and to require annotation (e.g., XML markup) of texts to be synthesised so that their intention as well as their content can be expressed. It is not a coincidence that the prevalent html texts allow (and encourage) just such additional rendering information.

So the remaining challenge for speech synthesis is to design and construct large speech databases which contain representative samples of all necessary speech units in all likely prosodic environments such that direct concatenation is possible for all utterances to be synthesised. The programming challenges are for fast index-based retrieval; the scientific challenges are for the identification and automatic labelling of the salient features.

These design challenges are task-specific, and will differ according to whether the synthesiser is to speak in place of a person, providing a voice which listeners will trust and like, or whether it is to be domain-independent, in which case compromises must be made between voice quality and coverage of the almost infinite possibilities for new lexical items and sentence constructions.

3. Summary

The recent history of speech synthesis has shown a shift from rule-based systems to data-based systems, The synthesis systems of the early seventies were limited by memory and perhaps overly optimistic about the capability of computers to replicate all the meaningful variations of the human voice. The remaining challenge for speech synthesis is not to find further rules for predicting phonemic or prosodic variation from text, but to design a source database so that representative elements of all the perceptually-relevant prosodic and voice-quality variants can be guaranteed present.

4. References

[1] Voder: Dudley, AT&T, 1950.
[2] MITalk: Klatt & Stevens, 1986
[3] Nu-talk: Sagisaka et al, 1992
[4] CHATR: Collected Hacks from ATR (ITL), 1997
[5] NATR: Next-generation Advanced Text Rendering (concatenative synthesis, forthcoming)

NEDOシニア支援システムと音声研究の課題

鹿野清宏　（奈良先端科学技術大学院大学　情報科学研究科）

1．まえがき

現在行っているNEDOの「シニア支援システム」プロジェクトの現在の目標と、来年度以降の計画を述べる。最後に今後の音声研究の課題について述べる。

2．シニア支援システム

NEDOの支援で「在宅高齢者インターネットインタフェースソフトウエアの研究開発」を今年の3月から開始した。補正予算ベースであるので、当面は来年3月までのプロジェクトとなっている。その後は、新たにプロポーザルを提出して、4年程度のプロジェクトに延長する意向を持っている。

プロジェクトの統括は、イメージ情報科学研究所の釜江尚彦氏がつとめ、これに三洋電機、東洋情報システム、東芝、NTT西日本、松下電産、松下電工、旭化成などの企業からの派遣研究員と、京都大学の石田研究室、河原研究室、奈良先端大の鹿野研究室、木戸出研究室、松本研究室、名古屋大学の武田研究室、大阪市大の北村研究室が参画して行われている。

研究目的は、高齢者用インターネットインタフェースソフトの開発である。キーボード、マウスなどの第1世代のインタフェースから、第1．5世代？のインタフェースとして、音声認識・合成、ノンバーバル・マルチモーダル、知的総合利用支援などを生かしたインターネットインタフェース構築の基本ソフトウエアの確立を目指している。研究所は、学外に設置し、音声認識・合成・マルチモーダルの研究は、奈良先端大の学外の研究所で、インターネット統合利用支援ソフトウエアの研究は、京都大学の学外の研究所で行われている。

私が担務している音声認識・合成の今年度の目標は、

(1) 汎用的に利用可能な高齢者音声データの収集（300人×200文、60歳～90歳）、
(2) バリエーションの大きい高齢者音声に合わせた音声認識アルゴリズム、
(3) 高齢者のインターネット利用に合わせた語彙の設定と言語モデル、
(4) 高齢者に聞き取りやすい規則合成音、

などである。今年度では、マルチモーダル、インターネット検索部分との統合は行わず、各要素技術の確立に重点を置く計画である。音声認識の研究のベースとして、IPAの「日本語ディクテーション基本ソフトウエアの開発」のプログラムJuliusおよび開発ワークベンチを用いている。音声規則合成は、ATRのCHATRをベースにしている。主な研究項目の特徴をまとめておく。

(1) 高齢者音声データベース：　学習用（300人×200文）、評価用（100人×200文）
(2) 音韻モデル：　高齢者音声データベースを用いてPTM音韻モデルを作成する。また、発声環境適応アルゴリズム、高齢者発声者のバリエーションの対処として話者クラスタリングによる音韻モデルの構築の研究を行う。
(3) 言語モデル：　高齢者インターネットタスクを設定し、高齢者の話し言葉に対応できる言語モデルを構築する。
(4) 認識アルゴリズム：　不要語、ポーズなどを透過語として処理できるようにデコーダを改造する。また、合成音声の出力中でも入力できるようにバージインプログラムを開発する。
(5) 音声合成：　CHATRの音声データベースを明瞭度の観点から評価して、明瞭性の高いCHATR用音声データベースを構築する。

その他、音声ブラウザーの基礎検討、話し言葉の品詞タグコーパスの作成、マルチモーダル部のジェスチャーや口の動きなどとの統合の予

備検討を行う予定である。

来年度以降も、今年度の研究を基に、研究開発の継続をNEDOに提案する予定である。そこでは、今年度に研究する高齢者用音声認識・合成、ノンバーバル・マルチモーダル、インターネット統合利用支援ソフトウエアを基本として、それらの機能向上と統合化、さらにシステム化し、実環境で利用可能なシステムのプロトタイプに仕上げることを目標とする予定である。

音声認識・合成関連の研究としては、1000人以上の高齢者音声データベースの作成、話者・環境同時適応アルゴリズムの改良、マイクロホンアレーを用いたハンズフリー音声入力、ステレオ出力に対応できるバージインプログラム、知的音声ブラウザー、声質変換付き規則合成システムなどが考えられている。

3．これからの研究課題

（1）音声データベースの共有

NEDOのシニア支援システムでは、現状の音声認識・合成技術を改良して、実環境で動かそうとする短期的な戦略に基づいている。さらに、このプロジェクトでは、研究開発で共有できる高齢者音声データベースや、話し言葉のタグコーパスにも重点を置いている。現在の音声認識のパラダイムでは、音声データベースの蓄積は、確実に性能の向上につながるが、投資効率からすると問題がある。さまざまな機関やプロジェクトでは、商用化利用まで含めた共有できるデータベースの蓄積が望まれる。とくに、多言語音声データベースは、ますます世界的規模で必要となってきており、共有の必要性は極めて大きい、LDC、ELRA、GSKのイニシアチブおよび協力が期待される。

（2）ツール・プログラム・開発環境の共有

音声認識プログラムや開発環境ワークベンチは、ますます高度になり、多種多様になってきている。一研究機関では、とてもすべてを整備することは難しくなってきている。音声データベースと同様にこれらツール・プログラムを共有して、かつ、共同で開発できるパラダイムが望まれる。

（3）人間の情報処理機能の利用

２０年ぐらい前までは、聴覚機能を模擬したパラメータの利用や認識アルゴリズムの研究が散見された。現在では、聴覚や脳機能の解明の研究は盛んになっているが、音声認識・合成への応用を考えている研究者はあまり目立たない。これは、HMMをベースした学習アルゴリズムが、あまりにも強力で、簡単には乗り越えることができないことも原因である。

（4）マルチモーダル情報との統合

ジェスチャーや顔表情と音声との融合の試みは、散見されるが、なかなか強力なインタフェースとはなっていない。これは、音声認識率などの従来の評価尺度にこだわっているためかもしれない。米国のDARPA次期プロジェクトの候補である「Multimodal browser over internet」のように、パラダイムを転換して、ターゲットをかえて研究を進めることも必要であると思われる。時代に即した適切な研究のターゲットの提示が重要であると思われる。

（5）言語の壁を感じないインタフェース

インターネットでのWWW情報の発信が普及するにつれて、言語の壁が大きな障害になってきている。英語圏以外では、WWWは英語版も作成しなければいけないのであろうか？言語の壁を取り除く検索技術、翻訳技術、音声認識・合成技術がますます重要となってきている。とくに、特色のある言語を使っている日本にとっては、とくに大きな問題である。

4．むすび

NEDOのシニア支援システムプロジェクトを紹介した。これには、当面の音声認識・合成の研究開発課題が含まれている。また、少し長期的な（10年）展望にたって、思いつくままに研究課題についても羅列してみた。