論文 / 著書情報
Article / Book Information

論題(和文)	音声認識・合成技術の将来 - NAS Colloquium on Human-Machine Communication by Voice の話題から -
Title	
著者(和文)	古井 貞熙
Author	SADAOKI FURUI
出典(和文)	日本学術振興会 「文字言語・音声言語の知能的処理 第152委員会」第31回研究会資料31-1, , , pp. 1-8
Journal/Book name	, , , pp. 1-8
発行日 / Issue date	1993, 3

音声認識・合成技術の将来

— NAS Colloquium on Human-Machine Communication by Voice の話題から—

古井　貞熙
NTTヒューマンインタフェース研究所
〒180 武蔵野市緑町3-9-11

あらまし　1993年２月８日と９日の２日間、米国California州Irvineで開かれた、NAS Colloquium on Human-Machine Communication by Voiceでの発表と討論を紹介するとともに、それをもとに、音声認識・合成技術の将来についての展望と、今後の課題について述べる。

1. 会議の概要

1.1 会議の名称：NAS Colloquium on Human-Machine Communication by Voice
1.2 主催：米国のNational Academy of Sciences
1.3 場所：The Arnold and Mabel Beckman Center (Irvine)
1.4 期間：1993年２月８日～９日
1.5 参加者：約160名
1.6 全般的内容：

音声認識および合成技術を用いて、人とコンピュータシステムがコミュニケーションできるようにするための、当面の技術課題と、21世紀を目標とした長期的課題について、講演と討論が行なわれた。招待講演として、NECの加藤康雄氏が、NECにおける音声認識の研究の歴史について紹介し、自動通訳電話などの将来の夢について講演した。各セッションは、テーマに沿った２～３人（各30分程度）の講演と30分程度の討論の形式で進められた。

2. 各セッションの内容

以下に各セッションのタイトル、講演者、討論リーダ（[ ]で示す）および概要を紹介する。

(1) Session I: Scientific Bases of Human-Machine Communication by Voice

P. R. Cohen (SRI), J. L. Flanagan (Rutgers Univ.), [R. W. Schafer (Georgia Inst. Tech.)]

種々のタスク（語彙、構文）、環境（雑音、プライバシー、通信ハードウェア）、およびユーザが持つ制約（手や目がふさがっている、障害がある、ストレスがある）のもとで、音声を人とコンピュータの対話に用いる長所と短所について、他の方法と比較して論じた。音声と他の手段を組み合わせた（補い合う）マルチモーダルインタフェースを考えることが重要。(Cohen)

音声認識と合成は、どちらも音声信号に含まれる情報の表現法という意味で、共通の基礎的基盤（音声生成の物理、言語的制約、聴覚系における情報処理機構）の追及に基

づいて研究をしていくべき。自然な音声をどのように真似するかという観点から音声合成の研究をすることが必要。(Flanagan)

(2) Session II: Speech Synthesis Technology

R. Carlson (KTH), J. Allen (MIT), [M. Liberman (Univ. of Pennsylvania)]

最近の音声合成技術（フォルマント型、素片結合型、など）のサーベイ。現在の方法の限界と、今後の課題。声質、発声スタイル、アクセント、多言語などの観点からの "flexibility"が、今後の課題の一つ。(Carlson, Allen)

(3) Session III: Speech Recognition Technology

J. Makhoul (BBN), F. Jelinek (IBM), [S. E. Levinson (AT&T Bell Labs)]

最近の音声認識技術（HMM、ニューラルネット、統計的言語モデルなど）のサーベイ。雑音や歪みへの耐性が重要。(Makhoul, Jelinek)

(4) Session IV: Natural Language Understanding Technology

M. Bates (BBN), R. Moore (SRI), [L. Hirschman (MIT)]

自然言語 (NL) 理解における基本的問題（構文、意味、プラグマティックス、対話など）のサーベイと、現在の主なアプローチの紹介。

最近の方法の特徴：◆従来の例や直観に頼るアプローチから、データベースに基づく方法へ、◆NLシステムにおける有効性とカバー率を意識し、対話を扱える方法へ、◆分析的知識と統計的知識の組み合わせへ。

今後の課題：◇大量の実言語の処理ができる文法、◇構文や意味のモデルが（半）自動的に導出できる方法、◇自己適応システム、◇音声処理との融合（音声言語の処理は、音声認識技術と、テキストを対象としたNL処理技術の単なる組み合わせでは不可能）、◇コンテキストによって人の話し方がどう変わるかなどの研究、◇全理解か部分理解か、音声言語かテキストか、の要求に応じて適応できる技術。(Bates, Moore)

(5) Session V: Applications of Voice Processing Technology I

J. G. Wilpon (AT&T Bell Labs), H. Levitt (City Univ. of New York Graduate School),

[C. Seelbach (Seelbach Associates)]

通信の分野は、音声認識・合成技術が生きる最大の市場。例：オペレータサービス、番号案内、情報案内を中心とする種々の新しいサービス。このためには、性能向上（精度、耐性、信頼性）、使い易さ向上（音声技術がユーザから見えないこと；どんな言葉でも発声できる）、実用化が迅速にできること、が大切。ユーザインタフェースの難しさを、種々の具体例でデモ。(Wilpon)

障害者の音声コミュニケーションを助ける、種々の音声認識・合成利用技術の紹介と、問題点（背景雑音への対処など）の指摘。(Levitt)

(6) Session VI: Applications of Voice Processing Technology II

C. J. Weinstein (MIT Lincoln Lab), G. Doddington (SRI, DARPA), [J. Oberteuffer (ASR News)]

音声認識・合成技術の、種々の軍隊での応用例の紹介と、種々の課題（話者や環境への耐性、新しいタスクへのポータビリティ、種々の音声技術間および他のメディアとの融合、など）の指摘。応用システムを開発しながら、研究を進めることの重要性を主張。(Weinstein)

民生応用の分野も広いと予想されるが、最も難しい分野（処理のコストが高く、環境、ユーザの面からタスクとしても難しい）とも言え、ノウハウがほとんど分かっていない。通信の分野と違って、コストを多数のユーザで分けて負担することができない。今後は、音声認識技術と半導体技術の進歩により、展開が期待される。(Doddington)

(7) Session VII: Technology Deployment

R. Nakatsu (NTT Basic Res. Labs), C. Kamm (Bellcore), [D. Roe (AT&T Bell Labs)]

耐性と、優れたユーザインタフェースが、音声認識・合成技術の応用を成功させる上で最も重要。技術の不完全さと限界の多くは、後者によってカバーできる。ユーザとシステムの間に行き違い（誤解）の生じないように、対話の流れを制御できることが大切。(Nakatsu, Kamm)

(8) Session VIII: Technology in 2001

S. Furui (NTT Human Int. Labs), B. S. Atal (AT&T Bell Labs), M. Marcus (Univ. Pennsylvania), [F. Fallside (Univ. of Cambridge)]

2001年をターゲットとして、音声認識・合成技術の目標（性能と機能）をどこにおくべきか、その達成のために、どのような研究のアプローチをとるべきか、について講演と討論。(Furui, Atal, Marcus)

主な応用は、データベース（情報）アクセスと操作、商品やサービスの注文、ディクテーション、電子秘書、ロボット、自動通訳電話、障害者補助など。技術が広く使われるためには、人がやるのに比べてメリットがあることが必要。マルチメディア環境の中での役割を意識すべき。思考過程などに関する新しい知識を自動的に獲得する機能や、ユーザの意図を推論する機能が重要。今後は、人間の音声生成および知覚のメカニズムに基づいたモデル化の重要性が増す（現在の技術は、これらとの乖離が大きい）。(Furui：付属資料参照)

スペクトル分析よりも安定でコンパクトな音声の表現法？　パラメータの真の分布が既知であることを仮定しない認識法？　モデルの適応的学習法？　音声から直接、調音型合成器に学習させる方法？(Atal)

NL処理に関して、grammar-basedとexpectation-basedのアプローチの融合により、キーとなる情報を抽出できるアルゴリズム（概念スポッティング）を開発すべき。将来的には、annotationしていないデータベースから言語構造を学習する方法の研究が必要。(Marcus)

3. 現状認識と今後の展望

(1) 音声認識（理解）の研究面での進歩に関しては、米国のDARPAとLDC(Linguistic Data Consortium)が大きな牽引車となっているが、新しいアイディアの創出という点では、日欧を含めて、DARPAに独立な研究機関の貢献も大きく、全体として近年大きな技術的進歩がなされている。

(2) 近年の、確率モデルに基づく音声認識技術の大きな進歩と、LSIやワークステーションの着実な進歩によって、実際に実時間で動作させることのできる音声認識システムのレベルが上がってきた。

(3) 実用化された音声合成装置の合成音声の品質も、かなり上がってきた。

(4) AT&Tによる、音声認識を用いた自動オペレータサービスがトリガーとなって、米国の電話系会社を中心に、音声認識と合成を用いたシステムの実用化の機運が急速に高まっている。

(5) しかし、もっと広い複雑なタスクに音声認識を適用し、真に役に立つシステムを作るためには、上でも述べたように、解決しなければならない問題がまだ極めて多いのも事実である。

(6) 今後、音声認識・合成の最新技術とニーズとの接点において実際のシステムを構築して、これを用いたヒューマンファクタ等に関する実験的研究を進めるとともに、長期的な視野からのじっくりした研究を平行して進めることによって、技術的向上をはかっていく必要がある。

(7) 連続音声認識・理解の高度化を進める上で、当面の最も重要な技術的課題は、言語処理（対話処理、ロバストパージング、サーチアルゴリズム）とロバストな音響処理である。

Towards the Ultimate Synthesis/Recognition System

Sadaoki Furui

NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan

1. General Requirement: Speech synthesis and recognition systems are expected to play important roles in advanced user-friendly human-machine interfaces in 2001. Services using these systems will include database access and management, various order-made services, dictation and editing, electronic secretarial assistance, robots, automatic translation telephony, and aids for the handicapped (e.g. reading aids for the blind and speaking aids for the vocally handicapped). Ultimate speech synthesis/recognition systems should match or exceed human capability (that is, they should be faster, more accurate, more intelligent, more knowledgeable, less expensive, and easier to use.)

However, it is neither necessary nor useful to try to use speech for every kind of input and output. Although speech is the fastest and easiest input and output means for simple information to/from computers, it is inferior to other means in transmitting complex information. It is important to have an optimal division of roles and cooperation in a multimedia environment including images, characters, tactile signals, handwriting, etc.

The future interface should be able to automatically acquire new knowledge about the thinking process of individual users, to automatically correct user errors, and to predict the intention of users by accepting rough instructions and inferring details. A hierarchical interface that first uses figures and images (including icons) for expressing global information, and then uses linguistic expression for details will be a good interface that matches the human thinking process.

2. Future Speech Synthesizer: Future speech synthesizers should satisfy the following requirements: a) Highly intelligible (even under noisy and reverberant conditions, and transmitted over telephone networks), b) Natural, c) Prosody control based on meanings, d) Capable of controlling synthesized voice quality and choosing individual speaking style (voice conversion from one person's voice to another, etc.), e) Multi-lingual, multi-dialect, f) Choice of application-oriented speaking styles, including rhythm and intonation (e.g. announcements, database access, newspaper reading, spoken e-mail, conversation), g) Able to add emotion, and h) Synthesis from concepts.

3. Future Speech Recognizer: Future speech recognition technology should satisfy the following requirements: a) Fewer restrictions on tasks, vocabularies, speakers, speaking styles, environmental noises, microphones, and telephones, b) Robustness against speech variations, c) Capability of adaptation and normalization to variations due to environmental conditions and speakers, d) Automatic knowledge acquisition for phonemes, syllables, words, syntax, semantics, and concepts, e) The ability to process discourse (conversational speech) (for example, to analyze context and accept ungrammatical sentences), f) Naturalness and ease of use in the human/machine interaction, and g) Recognition of emotion.

4. Use of Articulatory and Perceptual Constraints: Although it is not always necessary or efficient for speech synthesis/recognition systems to directly imitate the human speech production and perception mechanisms, it will become more important in the near future to build mathematical models based on these mechanisms in order to improve performance.

For example, when sequences of phonemes and syllables are produced by human articulatory organs, such as tongue, jaw, and lips, these organs move in parallel, asynchronously, and yet systematically. Present speech analysis methods, however, convert speech signals into a single sequence of instantaneous spectra. It seems to be important to decompose speech signals into multiple sources based on the concealed production mechanisms. This approach seems to be essential for solving the coarticulation problem, one of the most important problems in speech processing.

Psychological and physiological research into human speech perception mechanisms reports that the dynamic features of the speech spectrum and the speech wave play crucially important roles in phoneme perception. However, almost all speech analysis methods developed thus far assume the stationarity of the signals. Discovery of a good method for representing dynamics of speech is expected to produce a substantial impact on the course of speech research.

Importance of Prompts

- Prompting users with synthesized commands or questions is crucial for smooth and successful conversation between computerized systems and users by means of speech recognizers and synthesizers.

- The percentage of isolated word responses elicited from users depends strongly on the prompts made by the machine.

- The intonation of prompted speech as well as the content can strongly influence user responses.

Table 1 - Projections for Speech Recognition.

Year:	1990	1995	2000	2000 +
Recognition capability:	Isolated / Connected words	Continuous speech	Continuous Speech	Continuous speech
	Whole word models, Word spotting, Finite-state grammars, Constrained tasks	Sub-word recognition elements, Stochastic language models	Sub-word recognition elements, Language models representative of natural language, Task-specific semantics	Spontaneous speech grammar, syntax, semantics; Adaptation, Learning
Vocabulary size:	10 - 30	100 - 1,000	5,000 - 20,000	Unrestricted
Processor requirements:	2 - 4 DSP's (25 Mips/Chip)	4 - 10 DSP's (50 Mips/Chip)	5 - 50 DSP's (200 Mips/Chip)	20 - 60 DSP's (1,000 Mips/Chip)
Applications:	Voice dialing, Credit-card entry, Catalog ordering, Inventory inquiry, Transaction inquiry	Transaction processing, Robot control, Resource management	Dictation machines, Computer-based secretarial assistants, Database access	Spontaneous speech interaction, Translating telephony

Table 2 - Main causes of speech variation.

Environment	Speech-correlated noise reverberation, reflection Uncorrelated noise additive noise (stationary, nonstationary)
Speaker	Attributes of speakers dialect, gender, age Manner of speaking breath & lip noise stress Lombard effect rate level pitch cooperativeness
Input equipment	Microphone (Transmitter) Distance to the microphone Filter Transmision system distortion, noise, echo Recording equipment

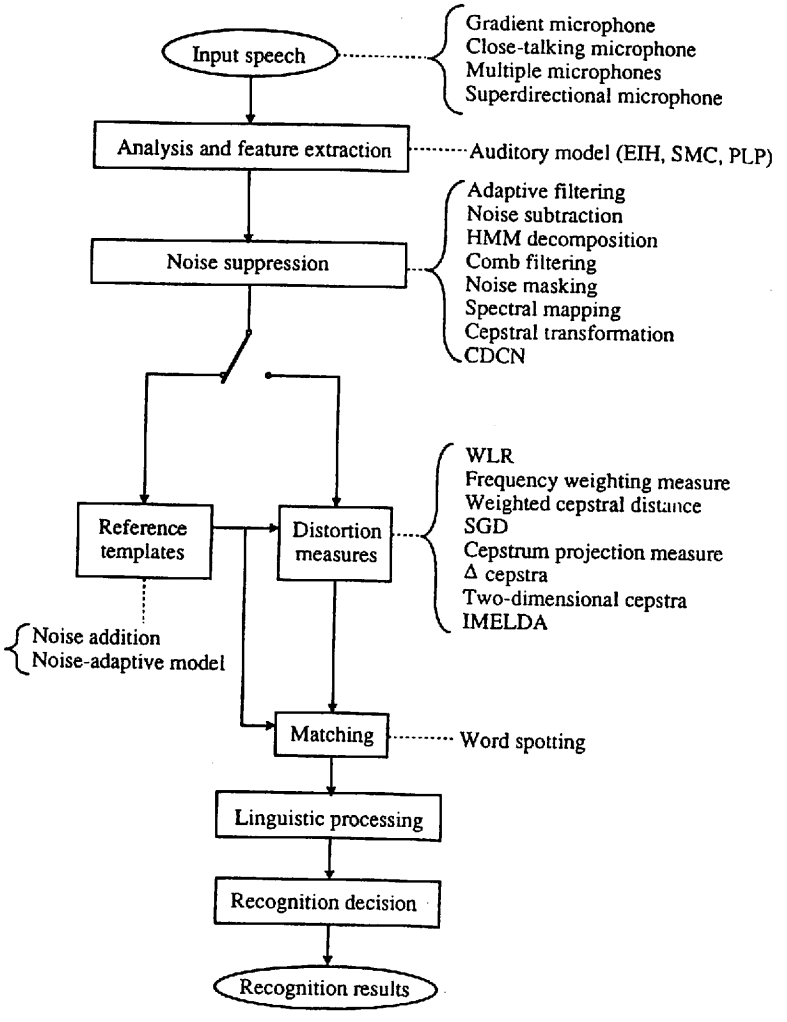

Fig. 1- Major methods for coping with noise problems in speech recognition.

Spontaneous Speech Recognition Problems

- Variations (extraneous words, out-of-vocabulary words, ungrammatical sentences, botched utterances, restarts, repetitions, and style shifts) → Robust and flexible parsing algorithms

- Instability in the detection of end-points and the necessity for the system to respond to the utterance as quickly as possible → Methods for detecting the time at which sufficient information has been acquired

- Contextual information → Prediction and focussing

Speech and Natural Language Processing

- New models that tightly integrate speech and language processing technology, especially for spontaneous speech recognition are important.
- New models should be based on new linguistic knowledge and technology specific to speaking styles.
- Proper adjustment of the methods of combining syntactic and semantic knowledge with acoustic knowledge according to the situation will be needed.
- How to extract, represent, and use concepts in speech are important issues in linguistic processing for spontaneous speech.
- Adaptation of linguistic models according to tasks and topics is also very important.