音声の知覚と認識　－人は脳で音声を聞く．機械は？－*

○赤木正人，羽二生篤（北陸先端科学技術大学院大学）

１．まえがき

機械による音声認識について，この数十年来数多くの手法が提案されてきた．しかし，まだ十分な性能は得られていない．本講演では，現状の音声認識装置ではこんなこと（もちろん人にとっては簡単なこと）は出来ないだろう，というものをならべたて，改めて人による音声知覚と機械による音声認識を対比してみることで，今後期待される音声認識のメカニズムを探る．

２．音声知覚のユニークな例

入力音声が歪んでいるあるいは存在しないにもかかわらず，音声が知覚される例を紹介する．これらの例の一部は，日本音響学会誌 61 巻 5 号に掲載された特集[1]でも取り上げられている．

2-1 正弦波音声（sine-wave speech）

正弦波音声は，周波数と振幅が可変である 3 本あるいは 4 本の正弦波の重ね合わせで合成された音声である．すなわち，低次のフォルマントの周波数と振幅だけの情報を持つ音声である．Remez ら(1981)によって，Science 誌に紹介された[2]．スペクトログラムを図 1 に示す．

最初にこの音声を聞くと音声とはとても思えない音色で，内容自体も了解が難しい．しかし，一旦元音声を聴取するあるいは文字情報として内容を提示されると，この正弦波音声が聞き取れるようになり，もはや「音声とは思えない音色で内容自体も了解が難しい」状態には戻らない．

2-2 劣化雑音音声（noise-band vocoded speech）

雑音音声は，音声信号を 3 ないし 4 つの帯域に分けそれぞれの帯域の振幅包絡情報は残したまま，キャリアを帯域雑音に置換した音声である．その知覚特性は，Shannon ら(1995) によって，Science 誌に紹介された[3]．日本では，力丸らが詳細な研究を行っている[4][5]．スペクトログラムの例を図1(c)に示す．

この音声を初めて聞くと，音声ではなく振幅変調された雑音にしか聞こえない．しかし，正弦波音声と同じように，一旦元音声を聴取する，あるいは，文字情報として内容を提示されると，音声知覚が容易になる．

2-3 音韻修復 (Phonemic restoration)

音声の一部が削除されているにもかかわらず，削除された部分に別の音が挿入されると，あたかも削除された音声が存在するように知覚されることがある．しかも，音声を削除せず雑音を付加させた場合との違いを聞きわけできない．この現象を音韻修復という[6][7]．図 2 に，子音の先頭から 100 ms を操作した音声のスペクトログラムの例を示す．

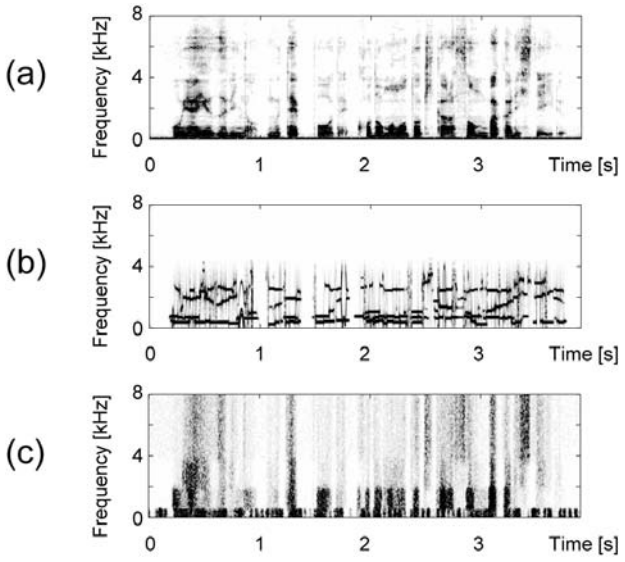

図 1　歪みを伴う音声の例．(a) 元音声，(b) 正弦波音声の例（正弦波が 3 本の場合），(c) 劣化雑音音声の例（4 帯域の場合）

4 種類の音声を聞き比べた場合，(c)と(d)では雑音が聞こえるものの音声の明瞭度は(a)と同等である．(b)は明らかに了解度が低下する．これは，音声を知覚する情報が 100 ms を超えて分布しており，その情報をトップダウン的に統合しているために生じる．

2-4 混合音声

複数の音声が混合して提示される場合，一つの音声にのみ注意を向けて聞くことはさして難しいことではない．しかもその発話内容が容易に想像される，たとえば，図 3(a)の混合音声の中に，同図(b)の音声が含まれることが分かっている，あるいは，文字情報として発話内容が分かっているとき，その音声に注意を向ければ，より目的の音声の聴取はたやすくなる．

３．人は脳で音声を聞く

上記四つの例すべて，耳から入る物理量（音響特徴）だけではなく，脳からのトップダウンの情報が強力に働いている例である．正弦波音声と劣化雑音音声は，提示された文字情報がトップダウンの情報となり，脳内で音声を作り上げている．音韻修復は，残された近隣の音声特徴をもとに，脳からのトップ

* Perception and recognition of speech: Humans hear speech by brain. How are machines?
By Masato AKAGI and Atsushi HANIU (JAIST)

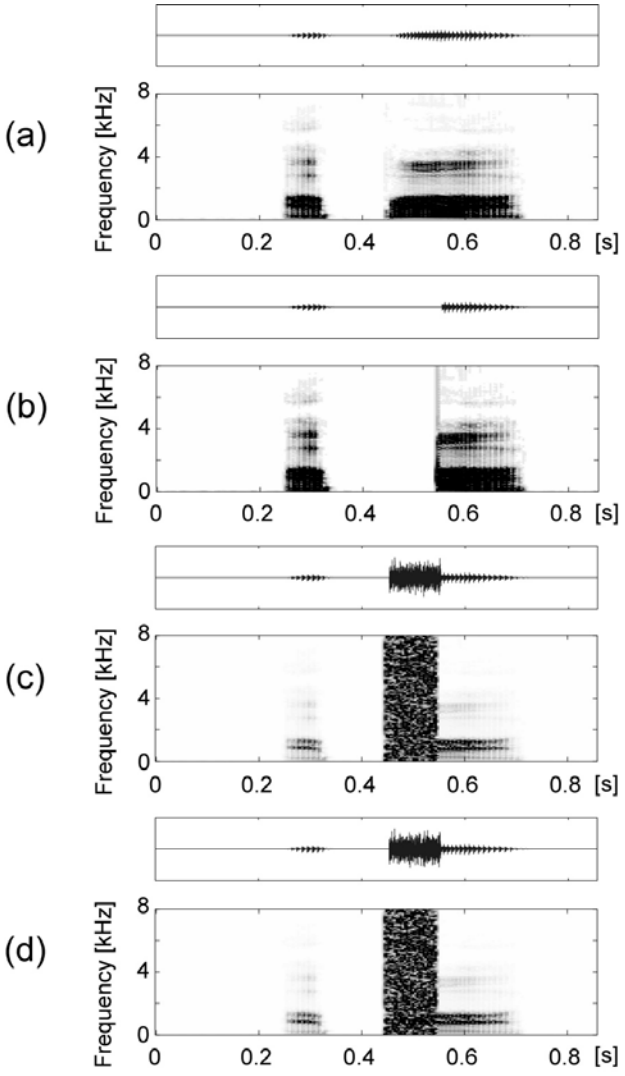

図 2　音韻修復の例．(a) 元音声の波形とスペクトル（以下同じ），(b) 一部（100 ms）が削除された音声，(c) 雑音置換音声，(d) 雑音付加音声

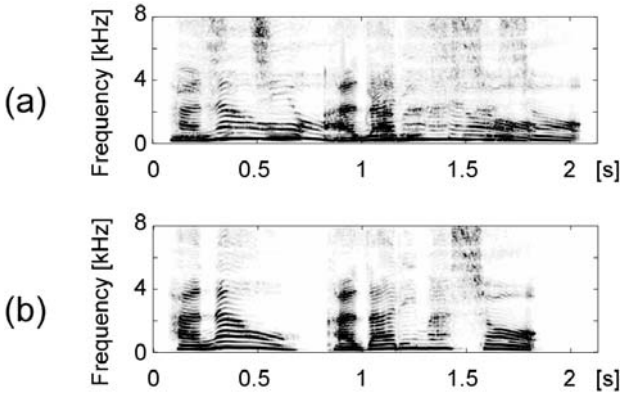

図 3　混合音声の例. (a) 混合音声, (b) ターゲットの音声.

ダウン処理により音声を再構成している．そして，混合音声では，提示されたきれいな音声の情報をトップダウン的に用いて，混合音中から目的音を検索している．すなわち，“人は脳で音声を聞く”と言っても過言ではない．

では，機械（現有の音声認識システム）ではどうであろうか．当然，上記の例にあるような劣化した音声の音響特徴は歪んでおり，現有の音声認識装置では認識は困難と思われる．しかし，音響特徴が壊れているのは，人に対しても機械に対しても同様である．ではどこが異なるのか？　異なるのはトップダウン情報の使い方であろう．

現有の音声認識システムの言語モデル，音響モデルでは，

Parsing：記号処理の領域でのトップダウン情報を記述

N-gram，N-phone，HMM：抽出された音響特徴のつながりを確率的に記述することにより，トップダウン情報を数十 ms の範囲で記述

と考えられるが, 時間的に広範囲にわたる, しかも, 音響特徴の抽出・再構築まで及ぶトップダウン情報は見受けられない.

４．音声認識の新たな展開：聞き耳のモデル

4-1 カクテルパーティ効果：聞き耳

二つ以上のメッセージが混在していても一方を選択的に聴取可能であるような聴覚上の効果を「カクテルパーティ効果」と呼んでいる[8].

カクテルパーティ効果の重要な要素として，聴覚的な「情景解析」(scene analysis)がある．人の聴覚情報処理過程では，周囲の音がすべて重畳された状態となった音を一旦部品に分解し，その後で強く関係する部品同士をまとめ，それらから周囲の状況を把握する．この聴覚の一連の情報処理過程は「聴覚情景解析」(ASA: Auditory Scene Analysis) [9]と呼ばれている．

Bregman によれば，人の耳に届いた音は，聴覚情報処理過程での初期的な表現により分解された後に, これを個々の音源から生じた音の一連のまとまり (音脈) を形成するように群化 (グルーピング) を行い ASA を行っていると，数多くの心理実験に基づいて述べている[9]. さらに, 音脈を形成する際には, 一つの音源から生じた音がもつ物理的な性質 (開始・終了時刻の同期，漸進的変化，調波構造，成分変化の同期性) を利用しているとされている[10]. 分解，群化，音脈形成の一連の働きを分凝とよぶ.

「聞き耳」は, 意図的に目的音に対して注意を向け, まさにこのような効果（分凝）を利用して，トップダウン的に目的音に関する事前情報から目的音を聞き分ける行為である.

4-2 聞き耳モデルを用いた認識システムの例

ここでは，目的音に対して注意を向けることによりトップダウン的に目的音に関する「事前情報」を積極的に利用し，混合音の中から物理的に妥当な音を選択的に聴き取る「聞き耳」の能力を模擬するモデルを提案する．そして,「聞き耳」のモデルを用いた音声認識システムを示す[11][12].

「聞き耳」のような場合には，目的音に対するトップダウン的な注意が向けられた音脈は「図」となり，その他は「地」となるような「地と図の分離」を行っていると考えられる．したがって,「聞き耳」のモデルでは, 独立成分分析 (ICA) のように全音源

に対して個別の音脈を形成する必要性はなく，最終的に「図」に相当する注意を向けた目的音の音脈が形成されればよい．

人の知覚では，「マスキング可能性の法則」[13]として知られるように，知覚される音がそこに存在したとしても物理的に何ら矛盾がない状況でない限りその音の知覚は生じない．つまり，人の聴覚情報処理過程においては，耳から入ってきた混合音中における目的音の「物理的存在」を評価しているわけではなく，目的音としての「存在の妥当性」を評価し，矛盾がないようであれば，それが「図」に相当する注意を向けた目的音として知覚される．したがって，目的音の音脈を形成する場合，目的音としての「存在の妥当性」を Bregman の発見的規則などを用いて評価する．

例えば，途中経過と最終出力が外部から検証可能な選択的音源分離モデル[14]に対して，図 4(a)に示したように目的音を含む混合音が入力された場合，選択的音源分離過程は何ら問題なく処理を行い，過程の妥当性と結果の評価に問題はない．一方，図 4(b)に示したように目的音を含まない混合音が入力された場合，選択的音源分離過程は途中で Bregman の発見的規則に対して矛盾を蓄積していき過程の妥当性が低くなるか，本来含まれていた音でも注意を向けて聴き取ろうとした音でもないものを分離し，結果の評価が低くなる．

以上のことを踏まえて，図 5 に示す選択的音源分離を用いたモデルを提案する．図中では，モデルへ入力されるモノラルの混合音を X_N と表記し，C_v は目的音のモデルから生成される目的音仮説，v は目的音仮説を区別するための添え字である．仮説には，音声知覚の Motor Theory[15]を念頭に，可能性のある音響特徴量系列を合成して使用する．モデルの出力は，目的音が存在したか否かの判定結果を出力する．

提案モデルの概念式を次に示す．

$$v_{\max} = \arg\max_{v}\left[\mathrm{Evaluation}\left\{\mathrm{Segregation}\left(X_N, C_v\right)\right\}\right]$$

本稿の「聞き耳」のモデルの概念式をみると，これは入力音 X_N に対して $\mathrm{Evaluation}[\cdot]$ が最大となる C_v を決定することと等価であり，これは音声認識器と読みかえることが出来る．

本モデルを用いた，雑音が付加された音声の簡単な認識実験の結果を示す．認識対象は日本語孤立発話数字，雑音は NOISEX-92 の中から，“pink”, “babble”, “machinegun”, “destroyerops”, “leopard”, “white”を使用した．

用いた ASR システムは，**A**: 前処理や雑音環境への適応を行わない ASR システム，**B-1**, **B-2**: 前処理を持つ ASR システム，**C-1**, **C-2**: 雑音環境へ適応を行った ASR システム，**D**: 提案手法に基づく ASR システムである．**B-2**, **C-2** は，雑音が既知な理想的なシステムとし，目的音に加算した雑音をそのまま処理に用いた．**B-1**, **C-1** では，雑音が未知で雑音推定が行われるシステムとし，混合音作成に用いた雑音の平均スペクトルを処理に用いた．結果の一例を図 6 に示す．

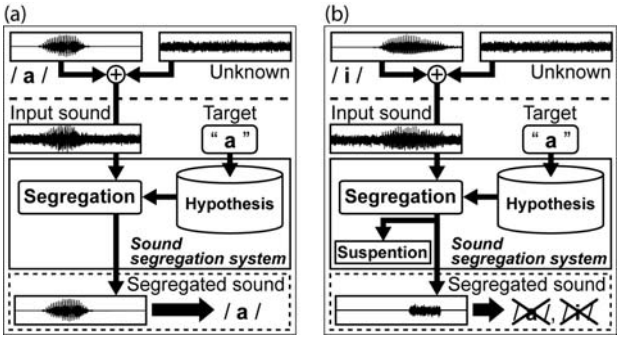

図 4　音源分離モデルの挙動 (a) 入力音中に目的音が存在した場合 (注意を向けて聴き取る音は “a”，入力音は /a/ と雑音が加算されたもの): 目的音と類似した音が分離される，(b) 入力音中に目的音が存在しなかった場合 (注意を向けて聴き取る音は “a”，入力音は /i/ と雑音が加算されたもの): 目的音と異なる音が分離されるか処理が一時中断する

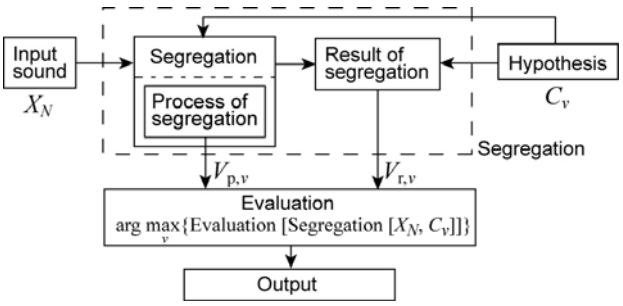

図 5　「聞き耳」モデルによる音声認識の概念

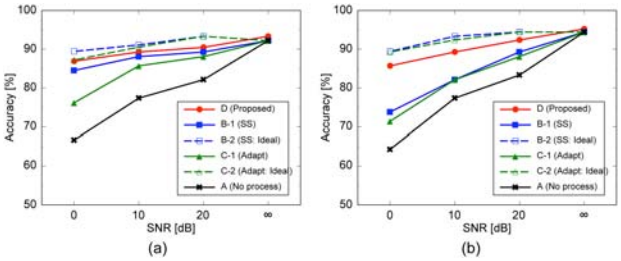

図 6　認識結果．(a) “babble”，(b) “machinegun”

「聞き耳」のモデルを音声認識モデルと捉えた場合，仮説を文字情報ではなく特徴量情報として予測生成して用いる点，混合音が入力されることが前提となっている点，認識の評価尺度が「存在の妥当性」という既存のものとは異なる点，既存の音声認識では推定精度の問題を含んでいた雑音モデルを内包していない点が特徴としてあげられる．

既存の音声認識では，リファレンスと入力音との類似性が認識尺度となっていたが，提案モデルでは「存在の妥当性」が認識尺度となっている．類似性は，クリーンな環境で使われることが前提であって雑音に対して頑健な評価尺度ではない．一方，提案モデルの評価尺度である「存在の妥当性」は，混合音であり音響特徴量が歪んでいることが前提で設け

られた尺度であることから，歪に対して頑健であると考えられる．また，既存の音声認識であれば混合音を処理する場合，雑音に関するモデルを用いて前処理や適応処理を行う必要があるが，混合音の入力が前提である提案モデルではその必要性はない．

５．HMM への朗報

図 7 は，横軸：単語対における音韻・音節の相違数，縦軸：データベース中の単語対の総数分布及び人による聴取誤りと HMM を用いた認識での認識誤り率を示している[16]．図を見れば，相違数 1 の場合は HMM の方が人に比べて誤り率は低い．これは，時間幅が短い範囲（音素 1 つ分）では HMM が優れていることを示している．ところが相違数 2, 3, 4 では HMM が劣っており，相違数が 2, 3, 4 となるにつれて単語の分布密度が多くなるので，総合すると HMM が人に比べて劣ってしまう．

HMM は，抽出された音響特徴の関係を数十 ms の範囲で記述する能力を有する．今後，時間的に広範囲にわたる情報を統合して扱えるようになれば，周りの情報から認識誤りの修正を行うことがより容易くなると思われる．

５．まとめ

本稿では，音声知覚のユニークな例のデモ，これらの効果の一端を取り入れた聞き耳モデル, および, 聞き耳モデルを用いた雑音中の音声認識を紹介した．他にもたくさん音声知覚の知見はあるが, ここでは，「トップダウン情報の有効な利用」に絞って解説を行った．

話す・聞くは人間の営みである．人を知り，そして，営みを記述することで、高度の音処理システムの実現が可能ならば，その方法を試してみる価値はある．

謝辞：本研究の一部は，SCOPE（071705001）の援助を受けて行われた．

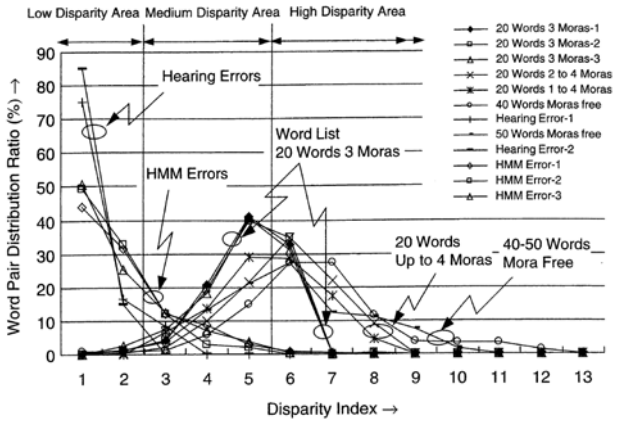

図 7　人による聴取誤りと HMM を用いた場合の誤り率

文献

[1] 日本音響学会(2005). “小特集-音と映像で体験できる聴覚の不思議な世界-”，音響学会誌，61, 5, 260-294.

[2] Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carell, T. D. (1981). Speech perception without traditional speech cues. Science, 212, 947-950.

[3] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. Science, 270. 303-304.

[4] 小畑，力丸(1999). "継時的振幅変化に着目した周波数成分劣化音声知覚の検討".聴覚研究会資料 H-99-6

[5] 橘，力丸(2005). "劣化雑音音声知覚に関する脳内活動：functional MRI による研究".聴覚研資料 H-2005-12

[6] Warren, R. M. (1970). “Perceptual restoration of missing speech sounds,” Science, 167, 392–393.

[7] Kashino, M. (2006). “Phonemic restoration: The brain creates missing speech sounds,” Acoust. Sci. & Tech. 27, 6, 318-321.

[8] Cherry, E. C. (1953). “Some experiments on the recognition of speech, with one and with two ears,” J. Acoust. Soc. Am., 25, 3, 975-979.

[9] Bregman, A. S. (1990). “Auditory scene analysis: The perceptual organization of sound,” MIT Press.

[10] Bregman, A. S. (1993). “Auditory scene analysis: Hearing in comples environments,” in Thinking Sound (Eds: McAdams and Bigand) Cp. 2, Oxford Univ. Press.

[11] Haniu, A., Unoki, M. and Akagi, M. (2005). “A study on a speech recognition method based on the selective sound segregation in noisy environment,” NCSP2005, 403-406.

[12] 羽二生，鵜木，赤木(2009). “ヒトの聴覚情報処理過程を考慮した音声認識モデル”，電子情報通信学会技術報告，SP2009-33.

[13] Warren, R. M., Obusek, C. J., and Ackroff, J. M. (1972). “Auditory induction: Perceptual synthesis of absent sounds,” Science, 176, 1149-1151.

[14] Unoki, M., Kubo, M., Haniu, A., and Akagi, M. (2006). “A model-concept of the selective sound segregation: A prototype model for selective segregation model of target instrument sound from the mixed sound of various instruments,” J. Signal Processing, 10, 6, 407-417.

[15] Liberman, A. M. and Mattingly, I. G. (1985). “The motor theory of speech perception revised,” Cognition, 21, 1-36.

[16] Ebukuro R. (2006). “Auditory ergonomics,” In International Encyclopedia of Ergonomics and Human Factors, Vol. 1,” (Ed. Karwowski, K.), CRC, 277-286.