日本音響学会誌 71 巻 4 号（2015），pp. 206–207

連載企画──音響学の温故知新──

Whither Speech Recognition?

──音声認識研究の展望──*

古 井 貞 熙（豊田工業大学シカゴ校）**

43.72.Ne

“Whither speech recognition?” [1] は，1969 年に米国音響学会誌に掲載された John R. Pierce のレターである。音声認識研究への批判が書かれており，要約すると次のようになる。

「一般用途の音声認識は当分できそうもないし，特殊用途の音声認識でできることは極めて限られている。到達目標をよく考えるべきである。…人間が他人の音声から実際に聞いている内容は極めて少なく，聞き取る内容の多くは記憶に基づいている。騒々しい電車の中で，英語が流暢な外国人でも聞き取れない会話を，英語を母国語とする人は聞き取ることができる。雑音中や不明瞭な発声に対しても，我々は何を話しているのか常に推測し，確信を持って内容を把握できる。このことは，英語を母国語とする人と同じ程度の知性と言語知識がなければ，一般的な音声認識は実現できないことを示している。…沢山の研究資金と時間が費やされているが，音声認識研究は欺瞞で，お金の無駄遣いである。…基本的で，明らかで，確かな知識は何も得られていない。やられているのは，実験（experiment）でなく，体験（experience）に過ぎない。」

これが音声認識研究は無意味という判断につながり，ベル研究所における研究をストップさせることになったが，数年のブランクを経て，NTT から派遣された板倉先生によって研究が再開されたことは，よく知られているとおりである。

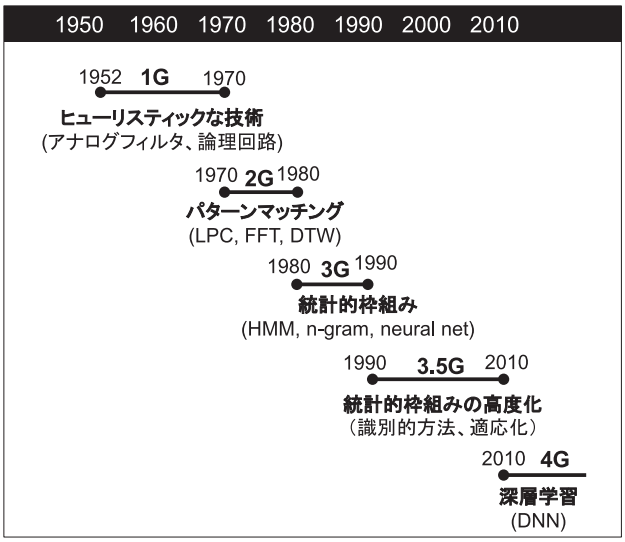
図–1 に，音声認識技術の変遷を示す。第 1 世代は 1950～1960 年代で，フォルマント周波数に着目し，アナログフィルタバンクと論理回路を用いたヒューリスティックな処理が用いられていた。これが Pierce の上記の批判を招くことになった。第 2 世代は 1970 年代で，あらかじめ蓄えておいたテンプレートとの，DP（動的計画法）を用いた時間軸整合マッチングが用いられた。第 3 世代は 1980 年代で，HMM（隠れマルコフモデル）と N-gram 言語モデルを用いた統計的枠組みが使われるようになった。このベースは，現在でも変わっていないが，1990～2000 年代にかけては，第 3.5 世代として，誤り率最小化に基づく識別的モデル，モデル適応，話し言葉音声データベースなどによる技術の高度化が進められた。第 4 世代は 2010 年以降で，深層ニューラルネットワーク（Deep Neural Network: DNN）による大幅な性能向上が実現されている。

私が 1970 年に就職した NTT 研究所では，1960 年代の後半から，基礎研究として音声認識研究を始めていた。私自身が研究を始めて 20 年たったときの研究所内の講演会で，人間並みの音声認識が実現できるまであと何年かかると思うかと質問されて，「あと 20 年待ってください」と答えた。それから今日まで 25 年，Pierce がレターを書いたときから 45 年がたって，統計的パターン認識，機械学習，ニューラルネットワーク，コンピュータ，スマートフォンなどの技術進歩により，役に立つ音声認識システムが多数使われるようになった。しかし，まだ人間並みとはとても言えない。

今後の音声認識研究を考えると，Pierce が書いたことを今一度よく考えてみる意味があるのではないかと思う。音声は，書かれた文を読み上げたものではない。原稿を書いてそれを読んでいる講演ほど退屈で分かりにくいものはない。人類が言葉（言語能力）でコミュニケーションを始めたのがいつかについては定説がないが，5～10 万年前と考えるのが妥当らしい。文字を持つようになったのは，紀元前 3500 年頃とされているから，音声言語は文字言語の 10 倍以上の長さの歴史があり，今でも文字を持たない文化が世界中に沢山あ

* Whither speech recognition? — Perspectives of automatic speech recognition research —.

** Sadaoki Furui (Toyota Technological Institute at Chicago, Chicago, IL 60637, USA)
e-mail: furui@ttic.edu



図–1　音声認識技術の変遷

る。音声言語と文字言語にはおのずから大きな違いがある。現在の音声認識では，言語知識として文字言語と同じ統計的言語モデルが使われ，認識性能の向上に大きな役割を果たしているが，音声言語のモデルとしては，明らかに限界がある。人の音声理解では，単語や文レベルでの予測が極めて重要で，知っている単語中の音素は早く知覚されることが実験で確認されている。また，音声では韻律情報が大きな役割を果たしており，音声の言語モデルでは，韻律情報を含む，離れた位置にある単語間の関係を表す必要がある。

調音結合の結果，音素のスペクトルは前後の音素の影響を受けて変化し，個々の音素のみでは精度よく認識することが難しい。この調音が不完全にしか行われない現象に対して，ある種の補正機構が人の聴覚処理系内に存在し，補正された特徴が知覚されているという研究があり，その工学モデルも研究されている。音声知覚実験によれば，スペクトルが最も大きく変化している区間が重要で，スペクトルの定常区間が削除されても，ダイナミックスからの予測として，音素が知覚できることが確認されている。工学的な音声認識では，今のところ，スペクトルやケプストラムの瞬時値にΔ特徴（時間微分）を用いると共に，トライフォンのように，音素コンテキストごとに別々のモデルを用意することで対応している。

人による音の動特性の知覚には，次のような3階層構造があると考えられている。

- 感覚・末梢過程（sensory process）：50～100 ms より短い時間で，知覚的現在（perceptual present）の中に入ってしまうので，その中での動的情報は動きとして感じるものでなく，時間形式を失って，音色として感じる。最低2 ms 以上の長さがあれば，その中の1個と2個のクリックを音色の違いとして感じることができ，時間感度（time sensitivity）と呼ばれている。20 ms 以上の長さがあれば，音刺激 AB 又は BA の時間順序の違い（逐次感）を，（順序は分からないが）音色の違いとして感じることができ，知覚瞬間（perceptual moment）と呼ばれている。
- 知覚過程（perceptual process）：100 ms 以上の時間で，音の変化（動き）を感じることができる。
- 認知過程（cognitive process）：数百 ms 以上の時間で，記憶を用いたトップダウン処理が行われる。音素や音節の移り変わりや，アクセントは，基本的に 100～数百 ms の長さの現象であり，イントネーションは数百 ms から数 s に及ぶ。

Pierce が書いたように，人は，多様な知識の記憶（雑音，部屋の反響，人による声や話し方のバリエーション，感情による話し方の変化，文法，意味，話題など）をもとに，それを状況に応じて適応的に用いて，相手が話している内容を予測し，聞き取っている。音声認識の高度化のためには，上記の知覚の3階層に対応した動的情報のモデル化と，その動的情報と記憶された多様な知識を用いた予測機能を実現することが必要である。人が音声を一見簡単に聞き取っているからといって，その仕組みが単純であるとは限らない。せっかく大量のデータや複雑な知識が DNN などの機械学習の仕組みで扱えるようになったのだから，人がどのような知識や情報を用いているかを真剣に研究し，それらを積極的に取り入れたシステムを構築すべきと思う。

文　献

[1] J. R. Pierce, “Whither speech recognition?,” *J. Acoust. Soc. Am.*, **46**, 1049–1051 (1969).