TR-I-0112

会話文音声生成のための音声合成、および ニューラルネットワークの連続音声への適用

Speech Synthesis by Rule for Conversation, and Applying Neural Networks to Continuous Speech

宮武 正典
M. MIYATAKE

Aug. 1989

概要

昭和６１年（１９８６年）９月より，平成元年（１９８９年）８月まで，自動翻訳電話研究所において，音声の合成，認識の研究を行った。ここにその概要を報告する。

会話文音声生成のための音声合成

より人間らしい，多種多様な音声を合成するためには，概念の形成から実際の音声の合成までの間に解決すべき様々な問題が山積している。その手始めとして，種々の口調・発声様式が韻律（声の大きさ，高さ，速さなど）に及ぼす影響の分析を行い，韻律パラメータ間に強い相関関係があることを明らかにした。また，種々の発声様式に適応できる基本周波数の制御法を提案した。

ニューラルネットワークの連続音声への適用

従来より高い音韻認識率を示していたニューラルネットワークであるが，連続音声への適用方法が確立されていなかった。今回，時間遅れニューラルネットワーク（ＴＤＮＮ）を用いた音韻スポッティングの手法と，そのための効率的な学習方法とを提案し，単語音声において音韻抽出率98.0%と極めて高い音韻スポッティング技術を確立して，ニューラルネットワークによる連続音声認識の可能性を示した。

ATR自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

１．はじめに

昭和６１年（１９８６年）９月より，平成元年（１９８９年）８月まで，自動翻訳電話研究所において，音声の合成，認識の研究を行った。ここにその概略を報告し，３年間のまとめとする。

２．会話文音声生成のための音声合成

より自然な，人間らしさを重視したマン・マシン・インタフェースの実現のためには，音声の多様な入出力技術が不可欠である。しかし，このための研究はようやく途に着いたところであり，音声合成技術においても，人間が行う多種多様な発声については分析すら十分に行われていないのが現状である。

状況に応じた音声出力を可能にするためには，種々の段階で数多くの問題を解決せねばならない。そこでまず，考えられる種々の制御レベルの一例を簡単にまとめた（「会話文音声生成のための音声合成」，情報処理学会第35回（昭和62年後期）全国大会講演論文集，2H-3，1987.9）。これらの多くの問題の中で，我々はまず，異なる口調による物理パラメータ（韻律パラメータ）の制御を考えることにし，種々の口調を実現する上で基礎となる韻律パラメータ間の依存関係について調べた。

２．１　種々の発声様式における韻律パラメータの特徴について

男女各１名が７種の発声様式（高く，低く，大きく，小さく，速く，ゆっくり）で発声した単語をもとに，１モーラ当りの平均継続長，母音中心の平均基本周波数値，および母音中心の平均パワー値を比較した。その結果，基本周波数とパワーとの強い正の相関関係が明かになった。また，発声様式の違いが基本周波数パタンの形状に影響を及ぼすこともわかり，韻律パラメータの統一的な制御のモデル化の必要性が示唆された（「発声様式の違いが韻律パラメータに与える影響の分析」，日本音響学会講演論文集，2-6-3，1987.10）。さらに会話音声についても観察を行い，種々の発話口調（怒った口調，親切な口調，など）が韻律パラメータに及ぼす影響を示した（「種々の発声様式における韻律パラメータの性質について」，電子情報通信学会音声研究会技術報告，SP87-62，1987.10）。

２．２　種々の発声様式における基本周波数パタンの制御モデル

種々の発声様式を持つ音声の合成を目指し，韻律パラメータの規則化の手始めとして，最も顕著な特徴を示す基本周波数パタンの制御モデルを提案した。このモデルは，従来のモデルの特徴を活かし，モーラ長で時間正規化された点ピッチパタンを臨界制動２次線形系の応答によるモデルで記述しており，安定したパタンを定量的に精度よく近似できる。このモデルのパラメータは，多種の複数個サンプルに同時に適用したＡｂＳ法により求められるので，より規則に適したモデルパラメータが得られる。実際に，モーラ長，アクセント型の異なる単語データに対して同時にＡｂＳ（Analysis by Synthesis）を行い，パラメータをほぼ正しく推定できることを示した。さらに，種々の発声様式に対して当てはめた結果，ひとつのモデルにでき得ることが明かになった（「種々の発声様式における基本周波数パタンの制御モデルの検討」，日本音響学会講演論文集，1-1-16，1988.3）。

以上をまとめてＡＳＡ（Acoustical Society of America）で発表し，その内容をテクニカルレポートにまとめた（"Prosodic Characteristics and Their Control in Japanese Speech with Various Speaking Styles"，ＡＴＲテクニカルレポート，TR-I-0025，1988.6）。

３．ニューラルネットワークの連続音声への適用

高い音韻認識率を示している時間遅れ神経回路網（Time-Delay Neural Networks=ＴＤＮＮ）を連続音声に適用するために，ＴＤＮＮの時間方向に対するシフトインバリアントな性質を活かして，音韻スポッティングを試みた。日本語の全２３音韻（子音１８，母音５）を認識する音韻統合ＴＤＮＮを用いて，手始めに単語音声に適用したところ，９０％以上の音韻が正しく抽出された。さらにスポッティング性能の向上を目指して，無音カテゴリを加えて再学習させたところ，語頭や語尾，無声破裂音の閉鎖部などで見られた /h/ や /z/ の挿入誤りが減少した（「全音韻を統合した時間遅れ神経回路網（ＴＤＮＮ）による音韻スポッティング」，日本音響学会講演論文集，2-P-(24)，1989.3）。また，類似音韻を認識するＴＤＮＮを階層的に用いてスポッティングを行う方法との比較を行った（「時間遅れニューラルネットワークを用いた音韻スポッティング法」，電子情報通信学会春季全国大会講演論文集，SA-1-4，1989.3）。またこれらの結果をまとめ，研究会に報告した（「連続音声中の音韻スポッティングのためのＴＤＮＮ構成法」，電子情報通信学会音声研究会技術報告，SP89-32，1989.6）。

依然として存在する挿入，脱落の誤り傾向を検討したところ，学習用音韻データの切り出し位置に依存していると思われる誤りが多いことがわかった。これを改善するために，学習用音韻データの効率的な切り出し基準を提案し，この基準に基づいて再学習を行ったところ，全音韻の98.0%が正しく抽出され，挿入誤りは初期の約４分の１に削減され，極めて高性能の音韻スポッティングが実現された（「時間遅れ神経回路網（ＴＤＮＮ）による音韻スポッティングの改良」，日本音響学会講演論文集，1989.10発表予定）。

以上の成果をまとめ，電子情報通信学会論文誌および IEEE's ICASSP (International Conference on Acoustic, Speech and Signal Processing) に投稿中である。