ニューラルネットワークを用いた基本周波数パターンのモデル化*

☆田中勇翔，森大毅 (宇都宮大)

1　はじめに

これまで、基本周波数 (F0) 軌跡などの韻律パターンからパラ言語情報に関連する特徴を抽出するためには発見的な手法が採られてきた。一方、深層学習を利用した特徴抽出が発見的な特徴を凌駕する性能を持つことが、数多くの分野で報告されている。本研究は、発話の F0 パターンを深層学習により階層的にモデル化することを目指し、その最初の段階として、まずは発話の局所的な F0 パターンについての知識を獲得することを目的とする。

その具体的な問題として、本稿ではニューラルネットワークを用いた F0 の復元問題を取り上げる。音声の無声区間では F0 が抽出されない。また自発音声から F0 の抽出を行うと creaky voice の区間で F0 が不連続となる場合がある。このような不連続な F0 軌跡から潜在的な連続した F0 パターンを推定できれば、音声分析や音声合成等で有用である。

本稿では、creaky voice 区間の F0 の誤抽出や無声子音の区間を潜在的な連続した F0 パターンに対する雑音とみなし、Denoising Autoencoder[1, 2] を用いて F0 パターンを復元するニューラルネットワークを構築する。

2　F0 パターンの復元方法

本研究で提案するニューラルネットワークでは、音声から抽出した F0 パターン (誤抽出および無声を含む) に対して修正または補間を行う。音声から人手により推定した藤崎モデルのコマンド列から生成した F0 パターンを潜在的な連続した F0 パターンとし、Denoising Autoencoder によって音声から抽出した F0 パターンから藤崎モデルの F0 パターンに復元するようなネットワークを学習する。

音声から抽出した各発話の対数 F0 系列をオーバーラップした固定長のフレームに分割する。次に、縦軸が周波数、横軸が時刻、画素値が F0 の存在位置を表すグレイスケール画像に変換する。フレーム毎に得られた F0 パターン画像をニューラルネットワークに入力し、出力層が誤抽出および無声区間が修正または補間された F0 パターン画像を得る。

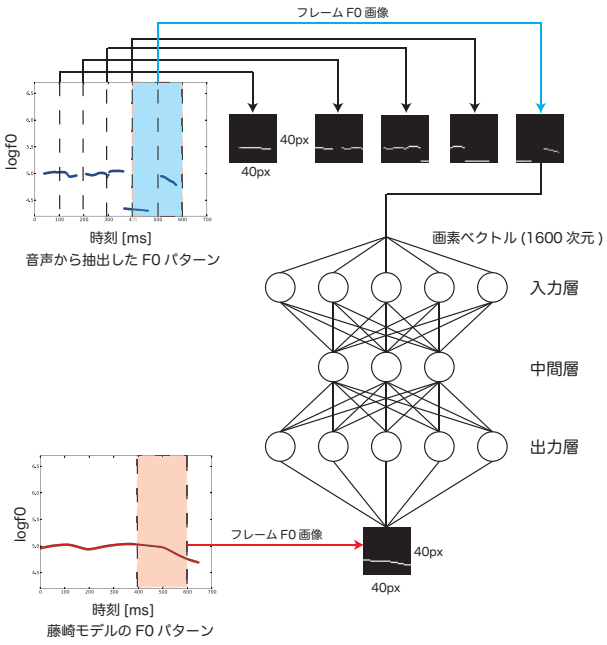

Fig. 1　ニューラルネットワークの学習方法

ニューラルネットワークの学習方法を図 1 に示す。音声から抽出した F0 パターン画像を入力し、ネットワークの出力画像と対応する正解の F0 パターン画像の画素ベクトルの二乗誤差を評価し、誤差逆伝播法により学習する。

3　F0 パターンの復元実験

本実験では Denoising Autoencoder により F0 パターン画像を復元するようなニューラルネットワークを学習した後、テスト用の F0 パターン画像を入力してニューラルネットワークの処理能力を評価する。宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB) に収録されている話者 6 名の 1307 発話から自己相関法によりサンプリング周期 5ms で抽出した対数 F0 系列に対して、藤崎モデルのパラメータである基底周波数が 4.45 lnHz になるように話者ごとに正規化し、フレーム長 200 ms、フレームシフト 100 ms で分割する。対数 F0 の下限を 4.2 lnHz、上限を 6.7 lnHz とし、F0 に対応する画素には白色

*Modeling of the fundamental frequency pattern using a neural network.
by TANAKA, Hayato, MORI, Hiroki (Utsunomiya University)

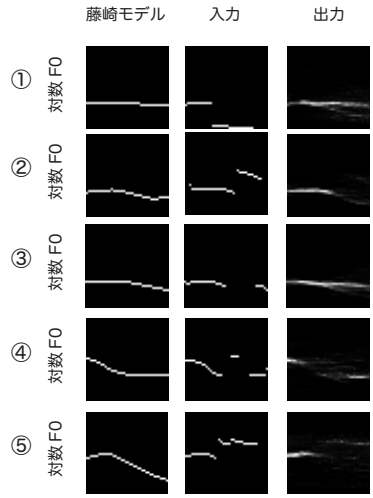

Fig. 2　正解と入力と出力画像の比較

を、それ以外には黒色を与えて 40 × 40 ピクセルのグレイスケール画像を作成した。本実験では 1307 発話から得られる 10623 フレームの画像を学習データ、2000 フレームの画像をテストデータとした。同様に、正解として藤崎モデルから作成した F0 パターン画像も作成した。なお、藤崎モデルのパラメータは文献 [3] で付与したものを用いている。

本実験では図 1 に示す 3 層ニューラルネットワークを用い、入力層と出力層のユニット数は 1600、中間層のユニット数は 200 とした。また、中間層と出力層共に活性化関数としてはシグモイド関数を用いている。学習回数は 5000 回で 200 フレーム毎に逐次学習を行っている。

実験結果の一部を図 2 に示す。図 2 は左からそれぞれ、藤崎モデルにより作成した F0 の画像、音声から抽出した F0 の画像、ニューラルネットワークが出力した F0 の画像である。図 2 の①及び②では、音声から抽出した F0 パターンにおいて誤抽出している区間をニューラルネットワークがおおよそ修正していることが分かる。また、③のように無声区間が補間されている F0 パターンも確認できた。しかし、抑揚のある F0 パターンに対しては④や⑤のように出力が鮮明でなかった。抑揚の無い F0 パターンが学習データの 7 割程度を占めているため、ニューラルネットワークが抑揚のあるパターンを十分に学習出来なかったものと考えられる。

本研究の F0 パターンの復元精度を定量的に評価するために、テストデータに対するニューラルネットワークの出力画像から得た対数 F0 系列と藤崎モデルから得た対数 F0 系列の RMS 誤差を求めた結果を表 1 に示す。有声区間における

Table 1　F0 パターンの復元精度 (RMS 誤差)[st]

	有声区間	有声区間+無声区間
RMSE	1.10	1.62

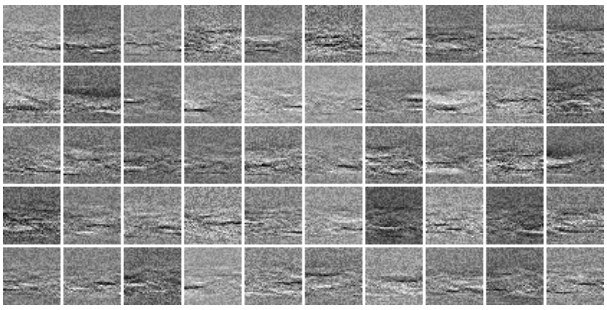
Fig. 3　Denoising Autoencoder 学習後の中間層表現

F0 パターンの修正精度は 1.10 と比較的小さいのに対して、無声区間の補間が難しいことが分かった。

図 3 は一部の中間層のユニットに結合する重みベクトルを画像にして並べたものである。図 3 から中間層では直線成分、下降成分といった F0 パターンの局所的な特徴を獲得していることが確認できる。抑揚のある F0 パターン全体を表現する特徴を得るためには、層を増やす等ネットワーク構造を変える必要がある。

4　おわりに

本研究ではニューラルネットワークを用いて局所的な F0 パターンのモデル化を行った。F0 パターン画像に対して誤抽出及び無声区間の修正、補間をある程度行うネットワークが得られた。また、中間層では F0 パターンの局所的な特徴を表現する重みが得られた。今後は、CNN 等の多層ニューラルネットワークを用いて発話全体の F0 パターンをモデル化し、局所的な特徴を組み合わせた抽象的な表現、つまり大域的な F0 パターンを表現するようなネットワークの構築について検討を行う。

参考文献

[1] Vincent et al., J. Mach. Learn. Res., **11**, 3371–3408, 2010.
[2] 小宮山他, 情報処理学会研究報告, SLP–97–1, 2013.
[3] 渡邉, 森, 音講論 (春), 511–512, 2013.