テクニカルノート

砂時計型ニューラルネットワークの多段化による LSP パラメータ圧縮特性の改善

木本雅也[†] 清水忠昭[†] 吉村宏紀[†]
井須尚紀[††] 菅田一博[†]

我々は，多段接続 5 層砂時計型ニューラルネットワーク（CSNN(NL5)）を用いた日本語 5 母音の LSP パラメータの情報圧縮と特徴抽出を行う手法を提案した．男性話者 1 名による日本語 5 母音の発話資料を用いて CSNN(NL5) の有効性を実証した．これにより，1) CSNN(NL5) により 2 次に圧縮されたパラメータの分布は，第 1，第 2 フォルマントの分布と類似した分布を示すこと，2) CSNN(NL5) は，圧縮した LSP パラメータを音声合成に使用できる精度で復元できることを明らかにした．

Improvement of Compression Characteristic of LSP Parameters by Cascading Sandglass Type Neural Network

Masaya Kimoto,[†] Tadaaki Shimizu,[†] Hiroki Yoshimura,[†]
Naoki Isu[††] and Kazuhiro Sugata[†]

We proposed a new scheme that derives the characteristics of Japanese five vowels out of LSP parameters by compressing information in terms of cascaded five-layer-sandglass-type neural network (CSNN(NL5)). We have verified the ability of CSNN(NL5) by using five vowels pronounced by a male speaker . The followings were clarified, 1) the distribution of LSP parameters compressed by CSNN(NL5) is similar to the distribution of F_1–F_2 formants, 2) CSNN(NL5) can reproduce the LSP parameters from the compressed parameters usable for speech synthesis.

1. まえがき

音声符号化や音声合成では，計算機で比較的簡単に推定でき，高い合成音声品質を得られる LSP（Line Spectrum Pair）パラメータ[1] 等の線形予測系のパラメータが広く用いられている．しかし，LSP による音声の分析合成には 10 次以上のパラメータが必要であり，音韻との対応関係をとらえることが難しい．

一方，母音の特徴をよく表現するパラメータとしてフォルマントが古くから用いられてきた．フォルマントは，スペクトル領域において優勢な周波数成分であり，音韻を特徴付ける主たる要因となっている．日本語の 5 母音は，第 1 フォルマント周波数 F_1 と第 2 フォルマント周波数 F_2 を軸とする F_1–F_2 平面上で特徴的な 5 角形の分布を示す[2]．しかしフォルマントは，計算機による自動推定が難しく，音声符号化や音声合成に直接用いるには不向きなパラメータである．

我々は，音声の音韻特徴を反映した情報圧縮や，簡易で高品質な規則音声合成に適用可能な音声合成パラメータを得るために，砂時計型ニューラルネットワーク（SNN: Sandglass type Neural Network）[3] を用いる手法を提案した[4]．先行研究において，非線形特性を持つ SNN を用いて 2 次に圧縮した LSP パラメータは，F_1–F_2 分布と類似した分布形となり，音韻との対応性が良いことを示した．しかしこの方法では，中間層ユニットの対称性のために学習の任意性が残り，音韻分布は学習条件によりパラメータ平面上で不都合な回転を起こす．

本稿では，SNN による LSP パラメータ圧縮における前述の欠点を改善するために，中間層ユニットを 1 個とした単位 SNN を多段にカスケード接続し，中間層出力に順位付けした SNN を提案した．提案手法を

† 鳥取大学工学部
Faculty of Engineering, Tottori University
†† 三重大学工学部
Faculty of Engineering, Mie University

実験的に検証し，この手法により，1) パラメータ平面上で分布の軸の回転がなくなること，2) 単段 SNN と同等の圧縮/復元特性を得られること，3) 音声合成に必要な精度でパラメータを復元するには，SNN の接続段数は 2 段で十分であることを明らかにした．本稿では，特に 1) について報告する．

2. SNN による LSP パラメータの圧縮

2.1 砂時計型ニューラルネットワーク

本稿の主題である SNN の多段化の検討の前に，通常の SNN（単段 SNN）と先行研究[4] の結果について概説する．

SNN は，入力層と出力層のユニット数が等しく，中間層のユニット数が入出力層よりも少ない構造を持つ階層型ニューラルネットワークである．SNN の学習は，出力に与える教師信号として入力信号と同じ信号を与えて出力誤差が十分に小さくなるように行う．学習が成功すると，SNN は入力層から中間層までのネットワークで入力信号の情報圧縮を，中間層から出力層までのネットワークで情報の復元を行い，信号を再構成する能力を獲得する．

先行研究では非線形特性を持つ 5 層（Non-Linear, 5-layers）の SNN(NL5) を用いた．SNN(NL5) の第 2，第 4 層には，シグモイド関数を応答関数とする非線形ユニットを配した．実験の結果，1) LSP パラメータの復元誤差は音声合成に適用可能な精度であり，2) 中間層出力（2 次の圧縮パラメータ）はフォルマントのように母音を分離する分布を示すことを明らかにした．しかし，初期重みや学習データ系列を変えて SNN(NL5) の学習を行うと，中間層出力の 2 次元平面上で音韻の分布形が回転し，分析パラメータとしては好ましくない．

2.2 多段カスケード接続 SNN

前節で述べた SNN(NL5) を用いた手法の欠点を改善するために，本稿では SNN の多段化手法を SNN(NL5) に適用した多段カスケード接続 SNN（Cascaded SNN: CSNN）を提案した．

CSNN の構成を**図 1** に示す．CSNN は，中間層ユニットを 1 個とした単位 SNN を多段にカスケード接続して構成する．第 1 段単位 SNN は，出力に与える教師信号を入力信号と同じ LSP パラメータとして学習する．第 2 段単位 SNN は，第 1 段単位 SNN の出力と元の LSP パラメータとの誤差信号を 1 段目と同様な方法で学習する．第 3 段以降も同様な構成である．本稿では SNN(NL5) を単位 SNN とし，第 2，第 4 層のユニット数は最適値の 20 個とした[4]．

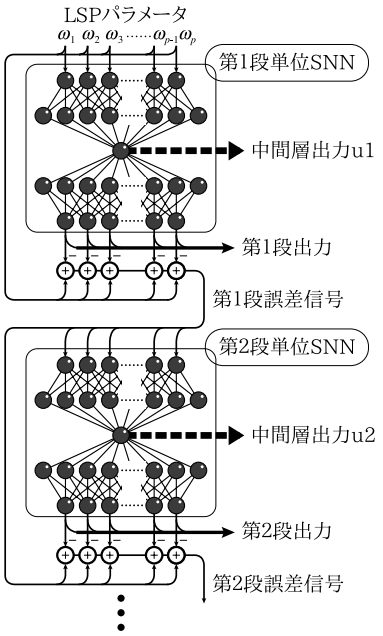

図 1 多段カスケード接続 5 層非線形砂時計型ニューラルネットワークの構成
Fig. 1 The structure of cascaded sandglass type neural network (non-linear, 5-layers).

表 1 音声資料の分析条件
Table 1 Condition of speech analysis.

A/D変換	
標本化周波数	11.025 kHz
量子化ビット数	16 bits/sample
LSP分析	
分析次数	14
フレーム長	256 samples(23.32 msec)
インターバル長	128 samples(11.61 msec)

3. CSNN(NL5) による LSP パラメータの学習実験

3.1 音声資料と CSNN(NL5) の学習データ

20 代の男性 1 名によって単独発声された日本語 5 母音を CSNN(NL5) の学習に用いた．各母音について 3 回発声した 15 個の音声資料を**表 1** の仕様でサンプリングし，LSP 分析した．CSNN(NL5) の学習データとして，各音声資料から中央部の音響特性が安定した 80 フレームを分析して得られた 80 組の LSP パラメータを用いた．このため，CSNN(NL5) に学習させる LSP パラメータは，音声試料 15 個 × 80 組=1,200 組となる．

以後，上記音声資料を用いた実験結果により議論を行う．ただし，すべての実験について 20 代の男女各 4 名の音声を用いて追加実験を行い，すべての話者で

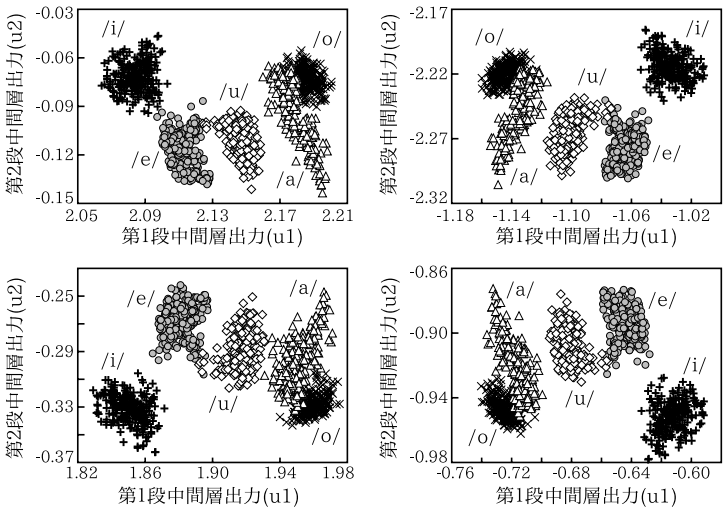

図 2 2 段 CSNN(NL5) の中間層出力の例
Fig. 2 Outputs from hidden units of 2-stages CSNN(NL5).

同様な結果を得ており，結論には一般性があると考えてよい．

3.2 CSNN(NL5) の学習実験

CSNN(NL5) の接続段数を変えて LSP パラメータを学習させ，音声合成に必要な精度で復元可能な段数を実験的に検討した．使用した LSP パラメータの次数は 14 次なので，CSNN(NL5) の段数は 1 段から 14 段とした．14 種類の CSNN(NL5) 各々に対し，提示順序の影響を抑えるためにランダムな順序に並べ替えた 10 通りの学習データ系列について，学習開始時の初期重みを変えた 10 試行ずつ，合計 100 試行の学習実験を行った．初期重みは，−0.01 から 0.01 の範囲の値をとる一様乱数で与え，100 試行すべてで異なる．実験の結果，単段 SNN と同様な圧縮/復元特性を持ち，2 段あれば音声合成に適用可能な精度で復元できた．

3.3 中間層出力の音韻分布

2 段 CSNN(NL5) による中間層出力の分布の代表的な 4 例を**図 2** に示す．u1，u2 は各々，第 1，第 2 段の中間層出力を表す．先行研究の単段 SNN(NL5) と同様な分布形が得られたが，分布形が 45 度や 90 度といった回転を起こしていない．u1 軸と u2 軸に対する正負の反転のみである．学習条件が異なる 100 試行で得られた CSNN(NL5) すべての u1–u2 平面を観察したところ，全試行で分布の軸の回転は見られなかった．ただし同図に見るように，分布形の平行移動やスケールの大小は，学習条件により異なる．

SNN の中間層は，大きな主成分を抽出する性質を持つ．中間層ユニットを 1 個とした単位 SNN を多段カスケード接続することで，大きな主成分から順に抽出することができる．中間層出力の分布が単段と多段で同形で，多段化により分布形の回転が抑止されるという実験結果は，上記の SNN の性質が本法でもうまく機能していることを示している．

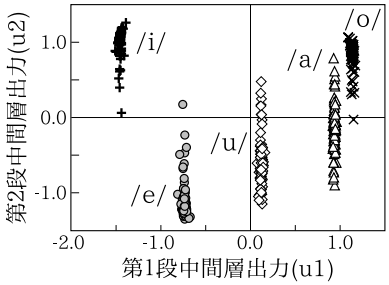

図 3 中間層出力の正規化を施した音韻圧縮パラメータの重心点
Fig. 3 Vowel centroids of normalized outputs from hidden units of 2-stages CSNN(NL5).

3.4 中間層出力の正規化

前節で示した問題を解消するため，以下の操作を施した．以後，この操作を中間層出力の正規化と呼ぶ．

1) スケール合わせと平行移動
u1 と u2 の 2 軸を各々1,200 個の学習データの平均と標準偏差で正規化する．

2) 軸の反転
フォルマントとの対応を良くするために/i/の分布が左上にくるように，軸の正負を反転する．

100 試行すべての u1–u2 分布に対して中間層出力を正規化し，音韻の重心をプロットしたものを**図 3** に示す．同図から明らかなように，u1 は正規化により全学習試行でほとんど一致することが分かる．u2 はどの音韻も学習ごとに変化するが，大まかな配置関係は

変わらない．

以上より，2段CSNN(NL5)を用いてLSPパラメータを2次に圧縮することで，フォルマントと同様な音韻クラスタを示すパラメータを得ることができた．また，圧縮パラメータを正規化することで，学習条件に依存しない安定なパラメータが得られる．本手法によって得られるパラメータは，フォルマントとLSPパラメータの利点を兼ね備えた有効なパラメータである．

4. おわりに

本稿では，2段CSNN(NL5)を用いたLSPパラメータ圧縮法を提案し，実験的に検証した．学習条件が異なる100試行の学習実験すべてにおいて，u1–u2平面上で分布形の回転をなくすことができることを示した．さらにu1–u2平面において正規化することで，異なる学習条件に対して安定なパラメータを得られることが分かった．

参考文献

1) 管村　昇，板倉文忠：線スペクトル対（LSP）音声合成方式による音声情報圧縮，電子情報通信学会論文誌（A），Vol.J64-A, No.8, pp.599–606 (1981).
2) 佐藤大和：男女声の声質情報を決める要素，研究実用化報告（NTT），Vol.24, No.5, pp.977–993 (1975).
3) Cottrell, G.W., Munro, P. and Zipser, D.: Image compression by back-propagation: An example of extensioal programming, Advances in Cognitive Science, Sharkey, N.E. (Ed.), Norwood, NJ: Ablex, Vol.3, pp.208–240 (1988).
4) 清水忠昭，木本雅也，吉村宏紀，井須尚紀，菅田一博：砂時計型ニューラルネットワークによる日本語5母音の特徴をとらえた音声合成パラメータの抽出，神経回路学会誌，Vol.11, No.4, pp.167–175 (2004).

(平成16年9月21日受付)
(平成17年1月 7 日採録)

木本　雅也（学生会員）

1977年生．2002年3月鳥取大学大学院工学研究科博士前期課程を修了．同年鳥取大学大学院工学研究科博士後期過程在籍．音声信号処理，ニューラルネットワークの研究に従事．神経回路学会の会員．

清水　忠昭

1963年生．1987年3月大阪大学基礎工学部生物工学科卒業．同年鳥取大学工学部助手．2002年鳥取大学工学部助教授．博士（工学）．音声信号処理，ニューラルネットワークの研究に従事．電子情報通信学会，日本音響学会，電気学会，神経回路学会の各会員．

吉村　宏紀

1970年生．1998年3月鳥取大学大学院工学研究科博士後期課程修了．同年九州工業大学情報工学部リサーチアソシエイト．1999年大阪府立大学工学部助手．2003年鳥取大学工学部助手．博士（工学）．ニューラルネットワーク，音声信号処理，ディジタル信号処理の研究に従事．電子情報通信学会の会員．

井須　尚紀（正会員）

1953年生．1978年3月大阪大学大学院基礎工学研究科前期課程修了．同年航空宇宙技術研究所研究員．1989年福井大工学部助教授．1992年鳥取大学工学部助教授．2003年三重大学工学部教授．医学博士．動揺病の生理工学・中枢神経系の生理学等の研究に従事．宇宙航空環境医学会，めまい平衡医学会，神経科学学会，生理学会の各会員．

菅田　一博

1938年生．1966年4月京都大学大学院工学研究科博士課程修了．同年同大学工学部電気工学科助手．1971年大阪大学基礎工学部助教授．1986年鳥取大学工学部教授．2003年停年退官．2004年近畿大学豊岡短期大学教授．工学博士．オートマトンと言語理論，計算の複雑さ，音声信号処理，ニューラルネットワークの研究に従事してきた．