Vol.2017-SLP-115 No.6
2017/2/17

進化戦略に基づいた単語検出ハードウェアのための DNN メタパラメータ最適化

王 健[1]　銭 博宇[1]　劉 溢[1]　篠崎 隆宏[1,a)]

概要：単語検出器を超低消費電力のハードウェアとして実現し生活空間に配置すれば、身の回りの様々な物を音声に応答して動作するようにできる。我々はこれまでにハードウェア実装のための、DNN 特徴量抽出器と DTW を組み合わせた単語検出器を提案した。しかし、この単語検出方式において高い検出精度を得るためにはいろいろなメタパラメータ、例えば DNN の層数とノード数などを最適化する必要がある。そこで、本研究では進化的アルゴリズムを応用し、これらのメタパラメータを最適化することを検討する。実験では、DNN の学習に日本語話し言葉コーパス (CSJ) を用いた。CSJ データを用いたパソコン上での進化実験と、FPGA 実装による実環境での評価実験の結果について報告する。

キーワード：キーワード検出、ディープニューラルネットワーク、DTW、ハードウエア実装、音声センサー

1. はじめに

単語検出器を超低消費電力のハードウェアとして実現し音声センサーとして生活空間に配置すれば、身の回りの様々な物を音声に応答して動作させるようにできる。このような目的のもと、我々これまでにハードウェア実装を目的としたディープニューラルネット (DNN) 特徴量抽出器と DTW を組み合わせた単語検出器を提案した [1]。DTW は HMM と比較して必要とするメモリや計算量が少ないことに加えて、音声テンプレートを登録するだけで学習を行うこと無く様々なフレーズの検出に対応できる利点がある。これは、音声センサーとして用いる場合に大きな利点である。さらに、特徴量抽出器に事前にパソコンやサーバーなどの高性能な計算機上で学習したディープニューラルネットワーク (DNN) を用いることで、DTW を用いながら話者や雑音に頑健な単語検出を行うことが可能となる。

しかし、この単語検出方式において高い検出精度を得るためには様々なメタパラメータを最適化する必要がある。これらを自動的に最適化する仕組みとして、我々は進化戦略を音声認識システムのブラックボックス最適化に応用する手法を提案している [2,3,4]。本研究では、このアプローチを音声センサー用の単語検出器の最適化に適用することを試みる。

[1] 東京工業大学
Tokyo Institute of Technology
[a)] www.ts.ip.titech.ac.jp

2. DNN 特徴量抽出器と DTW を組み合わせた単語検出器

図 1 に、DNN 特徴量抽出器と DTW を組み合わせた単語検出器における、DNN 学習および単語検出の処理フローを示す。図において最上段が DNN 特徴量検出器の学習プロセスを表している。DNN の学習には大規模な計算が必要となるが、このプロセスは検出対象の単語に依存しないためリソースの限られた音声センサー上で行う必要はなく、パソコン (PC) やサーバーなどを用いて行える。さらに、リソースの限られた音声センサーのハードウエア上への実装に適するように DNN の構造や計算精度などを時間をかけてチューニングすることも可能である。

図において 2 段目は、検出対象となるキーワードテンプレートの登録プロセスである。ここでは単語検出ハードウエア上のマイクロフォンで収録したキーワード音声または別の方法で別途収録した音声から MFCC 等の特徴量を抽出し、それを DNN 特徴量抽出器に入力することで話者の違いや雑音に頑健な特徴量へと変換する。そして得られた特徴量の系列を DTW のためのキーワードテンプレートとして用いる。DNN は上段のプロセスで得られたものをセンサー上に実装したものである。一度実装した DNN は、音声センサー上では固定して用いることを想定している。

そして最下段が、登録されたキーワードテンプレートを元に音声センサー上でキーワード検出を行うプロセスである。特徴量の抽出プロセスは、キーワードの登録の際と同

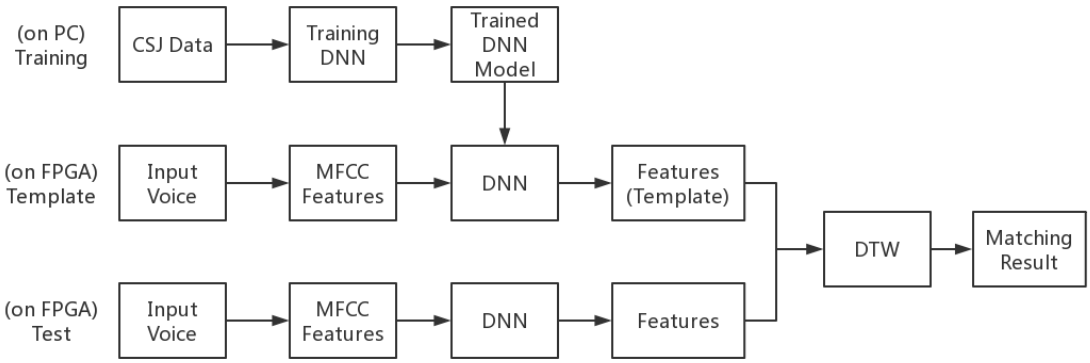

図 1 単語検出システム構成

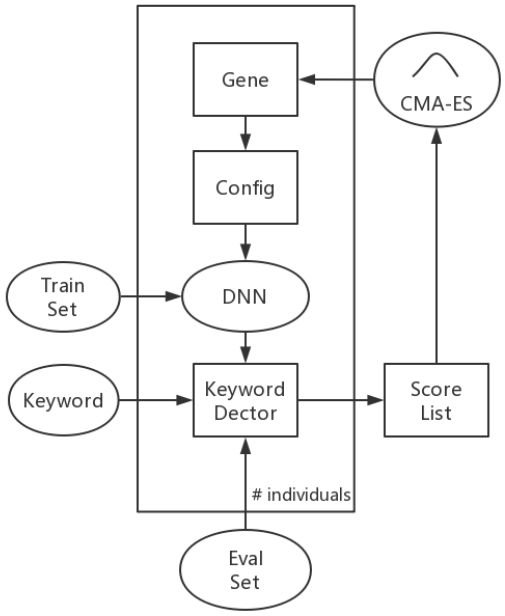

図 2 進化戦略を用いた DNN 特徴量抽出器の最適化プロセス

じである。使用時には、音声センサーは常時入力音声を監視し、対象とするキーワードが発声された場合いつでも即座に反応することが求められる。

3. 進化戦略を用いた DNN 特徴量抽出器の最適化

図 2 に、単語検出器用の DNN 特徴量抽出器の進化戦略を用いた最適化プロセスを示す。進化戦略としては、具体的には共分散行列適応進化戦略 (CMA-ES) を用いた。CMA-ES では、まず固定長の実数ベクトルとして最適化対象の問題を遺伝子表現する方法を決める。本研究では DNN の階層数および各階層でのニューロン数が最適化対象であり、これらを実数ベクトルに表現する。例えば最適化対象が自然数の場合、実数の指数をとったものを四捨五入し 1 を足す写像を定義することで、遺伝子中の実数表現を実際に必要な自然数に対応付けることができる。遺伝子の分布は、共分散行列を用いた多次元ガウス分布により表現される。ガウス分布の初期値は、適当に用意したシステム設定をエンコードした遺伝子を分布の平均、単位行列を共分散行列として指定することなどにより行う。初期設定の後、ガウス分布から N 個の遺伝子をその確率分布に従いサンプルする。このサンプル集合が第一世代の個体集合となる。また、N は一世代あたりの個体数である。

ついで、その世代における遺伝子集合中のそれぞれの遺伝子から、対応する DNN 学習のための設定ファイルを生成する。それぞれの設定ファイルは、それぞれの遺伝子の指定に従い DNN の階層数や各階層でのニューロン数を定義している。そして、それらそれぞれの設定ファイルに基づき、DNN の学習と単語検出率の評価を行う。各遺伝子の評価は独立しているため、並列計算が可能である。全遺伝子の評価が完了した後、それを元に遺伝子分布を表現する CMA-ES のガウス分布を更新する。ガウス分布の更新は、遺伝子に対応したシステム評価スコアの期待値が大きくなるように行う。分布の更新後、再度ガウス分布から N 個の遺伝子をサンプルする。これが、次世代の個体集合となる。以後このプロセスを何世代も繰り返すことで、より性能の高い個体の探索を行う。

表 1 進化実験の実験条件

世代数	5
個体数	20
遺伝子の次元数	11
初代 DNN の構造	60-80-60-60-42
学習データ	CSJ 8 時間
ターゲット音声	CSJ 10 講演 1.5 時間
テンプレート	オンセイ 290ms

4. 実験条件

DNN の学習に用いたのは、日本語話し言葉コーパス (CSJ) の音声データのサブセット 8 時間分である。特徴量

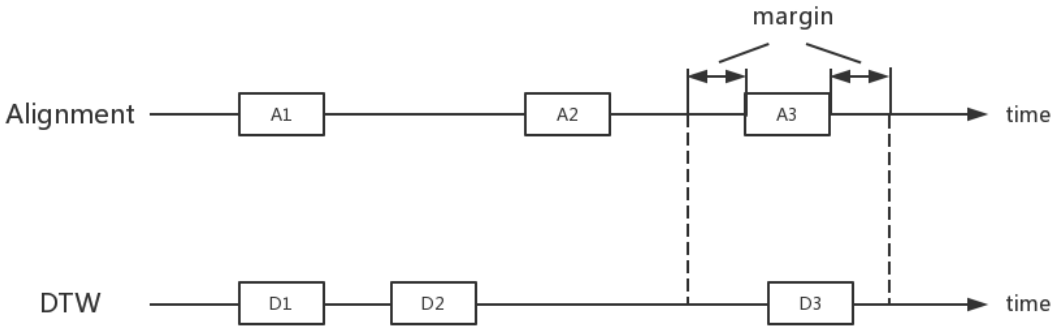

図 3 単語検出の採点基準。ターゲット音声中の検出対象音声の正解出現区間の両端に一定のマージンを儲け、DTW による検出区間がその区間内に含まれるときを正解とした。この図において、D1 と D3 は正しい検出であり、D2 は本来検出対象の単語が存在しないのにもかかわらず存在するとしてしまった誤検出である。また、A2 は存在しているのにもかかわらず検出に失敗している場合である。

表 2 遺伝子の内容

遺伝子の次元	内容
1	DNN の層数
2	第 1 層のユニット数
⋮	⋮
11	第 1 0 層のユニット数

は 12 次元の MFCC であり、DNN には前後 2 フレームを拡張した 60 次元を特徴ベクトルとして入力した。DNN の出力としては、42 種類の音素種別をターゲットとして用いた。これにより、不特定話者の条件でフレームごとに音素を認識する DNN が学習される。CMA-ES の初期化に用いた DNN の設定は 4 階層であり、各レイヤーのサイズは入力側から (60)-80-60-60-(42) である。この構造は事前実験により人手でおおよそ最適化したものである。DNN の学習は、パソコン上で行った。活性化関数として隠れ層はシグモイド関数を用い、出力層はソフトマックスを用いている。特徴量抽出器としては、学習の後最上段を取り除き、ボトルネック特徴量として用いている。表 1 に進化実験の実験条件、表 2 に遺伝子のエンコード内容を示す。

進化実験におけるシステム性能の評価は、本論文ではパソコン上で DNN 特徴量抽出器と DTW を実行することで行った。評価に用いた試験用キーワードは、CSJ の音声から切り出した「音声」の一単語の発声である。単語検出の性能は、式 (1) に示す F-Measure 方法を用いて評価した。

$$F = \frac{(1+\beta^2)PR}{\beta^2 P + R}, \quad \beta^2 = 0.5 \tag{1}$$

単語検出の成否は、図 3 に示すように正解単語の開始および終了時間に一定のマージンを儲けた幅を設定し、検出単語がその区間内にヒットした場合に正しい検出としてカウントすることにより行った。マージン幅は、3 秒に設定

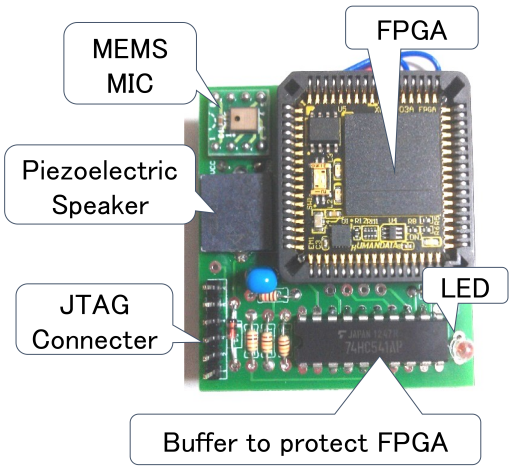

図 4 実環境評価に用いた FPGA 基板

した。検出単語は、フレームごとに得られる DTW の単語終端検出スコアが連続して一定値を下回った区間とした。しきい値は、遺伝子ごとに複数の値を試し最善のものを用いている。評価においてターゲット音声として用いたのは 1.5 時間の CSJ データであり、ターゲット中のキーワードの出現回数は 76 回である。

パソコン上での進化実験と評価に加えて、進化により得られた DNN を FPGA を用いた小型ハードウエアに実装した実環境評価についても行った。このハードウエアは、これまでに我々が開発した図 4 に示す FPGA 基板を用いたもので、FPGA の他に MEMS マイクと圧電ブザーおよび LED を搭載しており、電池で駆動する。FPGA には Spartan6 を用いている。表 3 に、FPGA 実装を用いた実環境の評価条件を示す。DNN は日本語データを用いて学習しているが、DTW を用いた単語検出方式であることから多言語の検出も可能であり、日本語に加えて中国語の

表 3 FPGA 実装を用いた実環境評価実験の実験条件

キーワード	ターゲット音声の話者 (人)	発話回数
オンセイ	4 (中国人 3、日本人 1)	20 回/人
パラメータ	4 (中国人 3、日本人 1)	20 回/人
Nihao	4 (中国人 3、日本人 1)	20 回/人

表 4 進化実験結果。スコアは単語検出の F 値

	gen0	gen1	gen2	gen3	gen4	gen5
best	0.334	0.405	0.404	0.379	0.382	0.399
mean	0.334	0.341	0.352	0.343	0.342	0.349

検出評価も行った。評価に用いたキーワードテンプレートは、日本語のキーワード「オンセイ」と「パラメータ」については CSJ より切り出した音声を用いた。また、中国語のキーワード「Nihao」については、ターゲット音声を発声した話者とは異なる中国人話者が発声した音声を用いた。評価時における FPGA 単語検出器上の MEMS マイクロホンと話者との距離は、1m である。

5. 実験結果

表 4 に進化実験の結果を示す。best は各世代での最良値、mean はその世代における平均値である。各世代の性能を見ると、初期個体と比べ僅かながら向上しているものの、あまり大きな効果は得られなかった。この原因としては、一つには今回の実験では評価に用いたキーワードが一つのみであり、また対象音声中のキーワードの出現回数も少なかったことから、学習された DNN の評価値が偶然に左右されやすく進化がうまく進まなかったことが考えられる。また、最適化対象のメタパラメタ数があまり多くなく、事前実験による最適化により概ね最適解が得られていたことも考えられる。今後はハードウエア実装の際の演算器のビット精度などより多くのメタパラメタを最適化対象とすることや、評価に用いるキーワードの種類やターゲット音声の分量を増やしてより正確な評価を行うことを予定している。

表 5 に FPGA を用いた実環境評価実験の結果を示す。進化実験の際の CSJ を用いた評価では連続発声を行っている発話音声中から検出対象単語と探す必要があるのに対して、このタスクではオフィス環境での背景雑音のもと孤立して発話されたキーワードを検出すればよいことから、CSJ 音声をタスクとした場合と比べて全体に F 値が非常に高くなっている。また日本語に加えて、中国語の単語についても高い精度で検出することができた。

6. まとめ

単語検出器に用いる DNN 特徴量抽出器の構造を進化戦略により最適化することを検討した。本論文では限定された実験条件を用いたため、十分な効果が得られなかった。今後、拡張した実験を行う予定である。また、最適化により得られた DNN 特徴量抽出器を FPGA を用いたハードウエアに実装した実環境評価を行った。日本語と中国語の検出実験を行い、DNN 特徴量抽出器は日本語で学習しているものの、どちらの言語の単語の検出も可能であることを示した。

表 5 FPGA の単語の実験結果

単語	ターゲット音声の話者	F 値
オンセイ	日本人	1.0
オンセイ	中国人	0.977
パラメータ	日本人	1.0
パラメータ	中国人	0.958
Nihao	日本人	0.964
Nihao	中国人	1.0

謝辞 本研究はマイクロソフトリサーチ CORE12 プログラムの支援を受けたものです。

参考文献

[1] 朱凱，李昊霖，篠崎隆宏，堀内靖雄，黒岩眞吾，“DNN 特徴量抽出器に基づく単語検出器の FPGA に実装と評価” 秋季音響学会, 2-q-6, 2015.9.
[2] T. Shinozaki and S. Watanabe, “Structure Discovery of Deep Neural Network Based on Evolutionary Algorithms”, Proc. ICASSP, 4979-4983, 2015.
[3] T. Moriya, T. Tanaka, T. Shinozaki, S. Watanabe and K. Duh, “Automation of System Building for State-of-the-art Large Vocabulary Speech Recognitio Using Evolution Strategy”, Proc. ASRU, 610-616, 2015.
[4] T. Tanaka, T.Moriya, T. Shinozaki, S. Watanabe, T. Hori and K. Duh, “Automated Structure Discovery and Parameter Tuning of Neural Network Language Model Based on Evolution Strategy” Proc. SLT, 665-671, 2016.