聴覚情報保障のための生活下トリガー音認識システムの研究

研究代表者　　篠　崎　隆　宏　　千葉大学大学院融合科学研究科 助教
（現東京工業大学大学院総合理工学研究科 准教授）

1 はじめに

音声は人間の五感の中でも、視覚とともに特に生活上重要な感覚である。さらに、視覚とは異なり顔の向きにあまり依存せず、また対象物との間に障害物があっても感知可能であるなど、独自の特性を有している。この特性は、不定期かつ突発的な事象に対して、それまでの意識の集中や動作を中断し、必要な対応を行おうとする際に特に有用である。例えば、読書をしているときに自分の名前を呼ばれたりインターホンが鳴ったりした場合、読書を中断し、その対応をする必要がある。また道を歩いていて後ろから自動車にクラクションを鳴らされた場合、歩行を中断して脇によるなどする必要がある。これらは日常生活を行う上で非常に基本的で重要なことである。家電製品のアラーム音や自動車のクラクションなど日常生活でなんらかのメッセージを伝えるための音は、サイン音[1]と呼ばれている。本研究では、サイン音のうち即時的な行動を喚起させるものや呼びかけ声などの言語音声を合わせて、トリガー音と呼ぶこととする。トリガー音に対しては一般に迅速な対応が必要となるが、聴覚に障害があるとそのような反応・動作を行うことが不可能となり、生活の上で大きな支障となっている。そこで、音声認識技術を応用してマイクから取り入れた生活環境下における呼びかけ声や家電製品のアラーム音、自動車のクラクションなどの信号を小型計算機により自動認識し、光や振動、短い文書などに変換して利用者に知らせることが出来れば、非常に有用であると期待される。本研究では、そのようなシステムの実用化の可能性と課題を調査し、システムの実現に必要となる技術について研究を行った。

2 生活下トリガー音認識

生活下でのトリガー音を認識するシステムとしての困難は、収録室での録音などと異なり、管理されていない環境下で距離的に離れた発音源からの音や発話者からの音声を認識しなければならない点である。またトリガー音は不定期に発生するため、システムは常時稼働しながら稀に発生するトリガー音イベントを精度よく検出することが求められる。

3 トリガー音の選定と収集

コンパクトなシステムで高い認識精度を実現するためには、認識する必要性の高いトリガー音を選定する必要がある。どのようなトリガー音が重要なのかを明らかにするため、3 名の聴覚障がい者の方に協力いただきアンケートによる調査を実施した。実施したアンケートの内容は2 項ある。1 つは日常生活で重要と予想されるトリガー音 19 種類を予め提示し、それらの重要度を評価してもらうものである。もう 1 項は重要だと思うトリガー音を自由に記述してもらうものである。アンケート調査の結果から、重要と評価された音を認識対象として決定した。ただし、重要度が高いもののデータを集めることが困難なトリガー音については、認識対象から除外した。なお、名前の呼びかけに関しては対象とする名前の呼びかけだけに反応するようにシステムを構築し評価する必要があることから、3 種類の呼びかけをトリガー音として設定した。対象とする名前に関しては、無作為に選択した。

選定したトリガー音の種類を表１に示す。これらのトリガー音について、本研究に適した既存のデータベースが存在しないため、データの収録を行った。収録には IC レコーダを用いた。「救急車のサイレン音」は、道路を通過する救急車が近づいた際にそれにマイクを向けることにより集めることにより得た。「自転車のベル」は自転車のベルを収録用に鳴らし、それを録音することにより収集した。「ノック音」は扉の前にレコーダを設置し、扉の反対側からノックを行うことで収録した。「インターホン」は室内にレコーダを置き、インターホンを鳴らすことにより収録した。クラクションは停車した車の前方においてクラクションを収録した

ものである。「名前の呼びかけ」では、複数の話者に３名の話者の名前を発声してもらい収録を行った。「その他」は、生活下トリガー音認識システムが使用される状況を想定し、日常行動中に長時間レコーダを携帯して収録した、トリガー音以外のもの全てである。動物の鳴き声やドアの開閉音など様々な音が含まれている。収録したデータについて、人手によりラベル付を行った。

表 1　選定したトリガー音

トリガー音	イベント１回あたりの継続長
救急車のサイレン音	3.3-57 sec
自転車のベル	1.0 sec 前後
ノック	1.0 sec 前後
インターホン	2.0-7.5 sec
キッチンタイマー	1.0-60 sec 前後
クラクション	0.1-2.5 sec
名前の呼びかけ	1.0 sec 前後
その他	N/A

4 孤立トリガー認識実験

既存の技術を用いてトリガー音を認識した場合の認識精度を調べるため、手動で切り出しを行った孤立トリガー音認識により、基礎的な評価実験を行った。

4-1 実験条件

入力されたデータが何かを判断するために音響モデルが使われる。音声音響認識によく使用されるモデルとして、複数の正規分布を足し合わせ尤度を計算する混合ガウス分布:Gaussian Mixture Model (GMM) や、複数の状態を持つことで時間変化を考慮する隠れマルコフモデル:Hidden Markov Model (HMM) がある。本研究では HMM を使用した。HMM は不確定な時系列のデータをモデル化するための有効な統計的手法であり、確率的な状態遷移を備えたオートマトンである。 音響特徴量は 25 次元の MFCC を使用した。認識対象のトリガー音は、人手による切り出しを行ったものである。認識率の評価には、クラスごとにデータを３分割しそのうち 20 サンプルを学習に用い残りの 10 サンプルに対して認識を行う、3-fold cross-validation を用いた。モデルの学習や認識実験には、Hidden Malkov Toolkit (HTK)[2] を使用した。

4-2 実験結果

表 2　孤立トリガー音の認識結果

トリガー音	単語正解率	単語正解精度
救急車のサイレン音	97%	40%
自転車のベル	100%	80%
ノック	90%	83%
インターホン	90%	87%
キッチンタイマー	100%	63%
クラクション	100%	100%
名前の呼びかけ 1	97%	83%
名前の呼びかけ 2	93%	90%
名前の呼びかけ 3	97%	83%
その他	73%	60%
全体	94%	75%
その他を除いた全体	96%	79%

認識結果の正解率と正解精度を表 2 に示す。「その他」を除く正解率は全体で 96%と高かったが、誤挿入のため正解精が 79%まで低下した。救急車に対する誤挿入が多く見られたほか、呼びかけ音声間での誤認識が観察された。その他の認識率が低いのは、トリガー音以外の様々な雑音や無音などを含むためである。正解率が高いことからトリガー音が発生した際には高い確率で検出できるものの、単語正解精度が低いことから誤検出が多くなる傾向があると言える。より高い認識性能を得るためには耐雑音性の向上や認識方法の改良が必要である。

5 Denoising Autoencoder を用いた残響除去手法の提案

音声認識において、話者あるいは発音源とマイクの間が離れている場合、残響の影響が顕著となり認識精度が大幅に劣化する問題がある。このことは、数メートル離れた話者あるいは発音源からの音を認識しようとするトリガー音認識システムにとって、大きな問題である。そこで、残響重畳音声に対して高い精度の音声認識を行うため、Denoising Autoencoder を用いてパワースペクトルから残響の影響を除去する手法を提案する。さらに、音声認識に必要なサブ音素レベルでの時間分解能を維持しながら長時間にわたる残響の影響に効果的に対処することを目的として、長さの異なる 2 つの分析窓長を併用する手法を提案する。CENSREC-4 および JNAS コーパスを用いた認識実験により、提案法の有効性を示す。

5-1 はじめに

残響とは音源からマイクロフォンに直接到達する直接音よりも遅れてマイクロフォンに到達する、壁などで反射・減衰した音声である。残響は音声の明瞭度を著しく下げ、音声認識の性能低下を招く要因になる。残響を抑制する手法は古くからの課題であり、研究が行われてきた。これらの手法は一つのマイクロフォンを用いるものと複数のマイクロフォンを用いるものに、大きく分けられる。一般的に、一つのマイクロフォンで残響抑制を行うよりも、複数のマイクロフォンで残響抑制した方が良い性能が得られる。しかし、複数のマイクロフォンを同時に利用することは装置の大きさやコストなどの点で不利となる。そのため本研究では、一つのマイクを用いた手法に焦点を当てる。既存の耐残響音手法としては、長時間スペクトル減算や Frequency Domain Linear Prediction などの長時間分析窓を用いた正規化手法が挙げられるが、未だ十分な性能が得られているとは言えない状況である。

他方、ニューラルネットワークにおける近年の研究において、局所最適解に陥りにくい初期値決定法を用いた Deep Learning が提案され、音声分野においても応用が研究されている。 耐雑音音声認識への応用としては、Deep Learning の分野における一手法である Denoising Autoencoder (DAE) を用いて、加法性雑音を除去する手法が提案され、有効性が示されている[3]。そこで本研究では、一定長の特徴量フレーム系列を入力とする、DAE を用いた残響除去手法を提案する。さらに、時定数の大きな残響を効果的に処理するために複数分析窓長の音響特徴量を同時に入力して残響除去を図る手法を提案し、残響環境下の音声認識実験で性能評価を行う。

音声認識実験には残響下音声認識の評価環境データベース(CENSREC-4[4]) を用いた残響下における連続数字音声認識と、新聞記事読み上げ音声コーパス(JNAS[5]) に残響を畳みこんだデータセットを用いた残響下における連続単語認識を行う。

5-2 Denoising Autoencoder（DAE）

Autoencoder は小さく絞られた中間層を持った、多層ニューラルネットワークである。ネットワークは中間層を通して入力から同じ出力が得られるように、学習を行う。多階層のネットワークをバックプロパゲーションにより直接最適化することは、困難であることが知られている。このため、学習は初期値を設定する Pre-training と最終的なパラメタを求める Fine-tuning の 2 つのフェーズにより行う手法が提案されている。具体的には Pre-training では Restricted Boltzmann Machine (RBM) と呼ばれる二層構造のネットワークを順次積み上げる形で教師なし学習する。次に、積み上げた RBM のパラメタをコピーし入出力を反転させ、それらの出力と入力を接続することで図 1 に示すような上下対称のネットワークを得る。最後に Fine-tuning において、入出力に同じデータを与えたバックプロパゲーションにより、教師あり学習を行う。DAE は Autoencoder の拡張で、ノイズデータからクリーンデータを予測出力するニューラルネットワークである。学習は入力に雑音重畳特徴量、出力にクリーン特徴量を設定して行う。

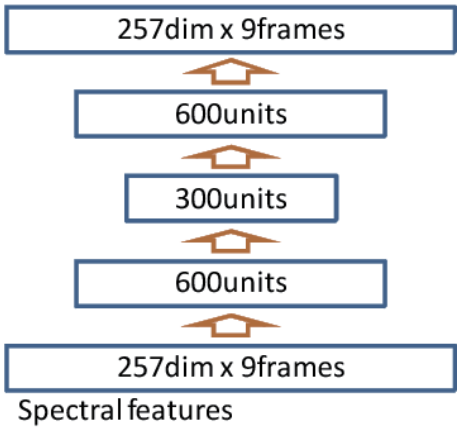

図 1 Autoencoder の例

5-3 提案手法

提案する耐残響音フロントエンドでは、DAE を用い残響が重畳したパワースペクトルから残響が除去されたパワースペクトルを推定する。DAE には連続した N フレームのパワースペクトルを入力し、出力として N フレームのパワースペクトルを得る。N フレームを同時に扱うことで音声の時間変化をモデル化し、それにより音声の知識に基づいた残響除去が可能となると期待される。

従来の耐残響手法で長時間の分析窓長を用いているのは、時定数の長い残響全体を捉え、正しく分析するためである。そのため提案法においても長時間周波数分析窓により得られる残響特性情報を加えることで、更なる性能の向上が期待できる。そこで、通常の音声認識において一般的な分析窓長の対数パワースペクトル特徴量に加えて、長時間分析窓のメルフィルタバンク特徴量を利用することを提案法の拡張として試みる。これにより音素のモデル化に必要な時間分解能を保持しつつ残響のモデル化に必要な長時間の特徴を取り込む事が可能となるため、さらなる性能向上が期待される。

残響を除去する際の具体的な動作としては、残響の重畳した N フレームの特徴量系列を 1 フレームシフトで DAE に入力する。そして、出力側において重なったフレーム時刻では対数パワースペクトル空間上で平均を求めることで、残響除去されたスペクトル特徴量系列を得る。音声認識で使用する MFCC などの特徴量は、そのスペクトルを元に求める。

5-4 CENSREC-4 を用いた数字認識実験

5-4-1 実験条件

評価実験には残響下での音声認識タスクを目的としたデータセット CENSREC-4 を用いる。CENSREC-4 は 8440 文のクリーンな連続数字読み上げ音声に 8 種類のインパルス応答(Office、 Elevator hall、 In-car、 Living room、 Lounge、 Japanese style room、 Meeting room、 Japanese style bath)を畳み込むことで室内残響環境音声をシミュレートしている。クリーン音声、インパルス音声ともに、サンプリング周波数 16kHz、量子化ビット数 16bit である。認識の対象は数字 11 種類 (1～9、 0 (マル)、Z (ゼロ)) で、各発話は 1～7 桁の連続数字音声となっている。 Autoencoder の学習音声はクリーン学習セット 8440 文と、それらをランダムに 4 分割して 4 環境のインパルス応答(Office、Elevator hall、 In-car、 Living room) を畳み込んだマルチコンディション学習セット 8440 文からなる。テスト音声はマルチコンディション学習セットと同じ 4 環境のインパルス応答(Office、Elevator hall、 In-car、 Livingroom) を畳み込んだ 4004 文 (TestA) と、それらとは別の 4 環境のインパルス応答(Lounge、 Japanese style room、Meeting room、Japanese style bath) を畳み込んだ 4004 文 (TestB) である。音声認識は CENSREC-4 のデフォルトの設定で行なっており、HMM に使用する特徴量はフレーム幅 25ms、フレームシフト 10ms の MFCC12 次元と対数パワーおよびその Δ と Δ Δ の 39 次元である。DAE の学習においては計算量の問題から、学習データの一部のみを使用した。Pre-training に用いたのは、クリーン学習セットからランダム選択した 2110 文とマルチコンディション学習セットからランダム選択した 2110 文の合計 4220 文である。また Fine-tuning では入力信号に同じ 4220 文を用い、教師信号にはインパルス応答を畳み込む前の対応するクリーン音声を用いた。Pre-training においては、Gaussian-binary RBM と binary-binary RBM の 2 つの RBM を順に学習した。Pretraining には Contrastive Divergence 法を用い、繰り返し数は 100 とした。Fine-tuning は二乗誤差

を目的関数として、共役勾配法に基づくバックプロパゲーションにより行った。 DAE の入力として使用するのはフレーム幅 25ms、フレームシフト 10ms の 256 次元対数パワースペクトルと対数パワーであり、セグメント長 N は 9 フレームである。すなわち DAE の入力次元数は(256+1) × 9=2313 である。これを「単一窓幅セグメント特徴量」と表記する。長時間分析窓を用いる場合はフレーム幅 500ms、フレームシフト 10ms の 24 次元対数メルフィルタバンクと対数パワーを追加特徴量とする。すなわち DAE の入力次元数は (256+1+24+1) × 9=2538 である。これを「複数窓幅セグメント特徴量」と表記する。

DAE を学習した後、それを全学習データおよび評価データに適用し、得られた特徴量をそれぞれ HMM の学習と認識評価に用いる。HMM は、日本語数字に必要な 18 種類の音素モデルと長さの異なる 2 種類の音素モデルの計 20 モデルである。長い無音とそれぞれの音素モデルは 5 状態のモデルで、短い無音は 3 状態のモデルである。各状態のガウス混合分布は 20 混合(無音モデルは 36 混合) とする。

5-4-2 実験結果

表 3 に CENSREC-4 における従来法による正解精度と提案手法による正解精度の比較を示す。表で"Baseline"は MFCC+Δ+ΔΔを用いた際の結果であり、"CMN"は cepstral mean normalization を行った際の結果である。これらにおいて、"Clean"はクリーン学習セットで HMM を学習した場合の結果、"Multi"はマルチコンディション学習セットで HMM を学習した場合の結果である。"Hdelta" は Hybrid delta 法でデルタ特徴量を取得した結果である。"DAE-S" は提案する単一窓幅セグメント特徴量を用いた DAE による結果であり、"DAE-SL"は複数窓幅セグメント特徴量を用いた結果である。従来法と２つの提案法を比べると、提案法の方が高い正解精度を与えている。また、提案法において単一窓幅セグメント特徴量を用いた DAE-S の場合（TestA：97.9%、TestB：96.4%）よりも複数窓幅セグメント特徴量を用いた DAE-SL の場合（TestA：98.4%、TestB：97.0%）の方が高い正解精度を与えていることが分かる。残響環境クローズである TestA だけでなく残響環境オープンである TestB に対しても頑健に残響除去できており、提案法が効果的であることが分かる。

表 3 CENSREC-4 を用いた数字認識実験の結果

Method	testA	testB
baseline(Clean)	83.8	82.8
baseline(Multi)	92.9	87.8
CMN(Clean)	86.5	88.6
CMN(Multi)	91.8	89.7
HDelta(Multi)	95.7	94.7
DAE-S	97.9	96.4
DAE-SL	98.4	97.0

5-5 JNAS 新聞読み上げコーパスを用いた認識実験

5-5-1 実験条件

CENSREC-4 を用いた実験では、数字認識における提案手法の有効性を確認した。本章では連続単語認識をタスクとした評価実験を行う。使用するデータセットは JNAS の新聞読み上げ音声に CENSREC-4 の環境残響を畳み込んだデータセットを使用する。JNAS は彙連続音声認識研究を目的とした毎日新聞記事を読み上げたコーパスであり、155 セット(約 100 文/セット、16176 文) の新聞記事を男女計 306 名の話者が読み上げたデータが含まれている。音声のサンプリング周波数は 16kHz、量子化ビット数は 16bit である。本実験ではこのうち 196 名の話者からの計 19895 文をトレーニングデータセット、20 名の話者からの 840 文をテストデータセットとして用いる。トレーニングセットとテストセットはそれぞれランダムに 4 分割し、トレーニングセットには前章の TestA の 4 種類のインパルス音声((Office、Elevator hall、In-car、Living room) を、テストセットには TestB の 4 種類のインパルス音声(Lounge、Japanese style room、 Meeting room、Japanese style bath)をそれぞれに畳み込んだ。

DAE の学習には計算量の制限から、上記学習セットから話者 80 名をランダムに選択しさらに話者ごとに 15 文をランダムに選択した計 1200 文を用いた。DAE の学習後、学習された DAE を全学習データに適用し、それにより得られた残響除去特徴量を用いて HMM の学習を行った。DAE の入出力としては前章と同じスペクトル特徴量を用い、また HMM についても前章と同じ MFCC 特徴量を用いた。言語モデルは JNAS から学習した 3-gram

を使用した。

5-5-2 実験結果

表 4 に JNAS を用いた単語認識実験の結果を示す。表で"Baseline"は MFCC+Δ+ΔΔを用いた際の結果である。"Clean"はクリーン学習セットで HMM を学習した場合の結果、"Multi"はマルチコンディション学習セットで HMM を学習した場合の結果である。表より、提案する DAE を用いた手法が連続単語認識実験においても有効であることが分かる。また提案法について、複数窓長を用いる方が単一窓長を用いる場合よりも高い性能が得られることが再確認された。

表 4 JNAS を用いた連続単語認識実験の結果

Method	Word accuracy (%)
baseline(clean)	44.3
baseline(multi)	59.8
DAE-S	64.7
DAE-SL	66.1

6 音声認識システムのパイプライン分解と遅延評価を用いた実装法

今日の音声認識デコーダは大規模な統計的モデルのもとに大規模な探索処理を行うことで動作している。そのソフトウエアは大規模で複雑なものである。特に、無限を含む任意長のデータを最小限の時間遅れで効率的に処理することと、様々なコンテキストの効果を取り入れた高精度な認識を行うことを両立させようとすると、実装は非常に込み入ったものとなりがちである。このことは、常時認識処理を行い音響イベントの発生と同時に即座の反応を求められるトリガー音認識システムにおいても、問題となる。

音声認識と同じく時系列を扱う分野として、線形時不変システムなどを対象とする伝統的な信号処理分野がある。そこでは、ブロックダイヤグラムを用いた設計法が広く用いられている。これは、単純な機能コンポーネントをパイプラインで接続することでシステムを構成するアプローチである。ダイヤグラムは視覚的で理解が容易であるとともに、データフローの構造定義と計算のタイミングが分離される利点がある。しかし通常のプログラミング言語と比べ一般のシステムを記述するには柔軟性に欠け、これまでのところ信号処理以外への応用は限られている。

他方、ソフトウエア工学の分野では純粋関数型プログラミングが手続き型プログラミングに代わる抽象度の高い記述を実現するパラダイムとして研究され、実用化されつつある。純粋関数型プログラミングは変数を一切持たない。そしてこの特徴から、遅延評価と呼ばれる関数引数の評価戦略との親和性がよいという手続き型言語には無い利点を有している。遅延評価は関数の結合関係と実際の計算のタイミングを分離し、関数に巨大なあるいは無限サイズのデータ構造を渡しながら同時に効率的な計算を可能とすることで、コードの抽象化に貢献する。しかし、プログラミングにおいて実際の計算タイミングの把握が難しく、全体の効率を最適化することが困難となることも多い。実際、我々はこれまでに純粋関数型言語 Haskell を用いてわずか 400 行程の非常にコンパクトなコードで大語彙連続認識デコーダを記述し動作することを示したが、大量のメモリや CPU コストが必要であり、一般的なパソコンで動作させることが困難であった。

この問題を解決するためには、純粋関数型言語において効率的に時系列データを扱う上で指針となるものが必要である。そのような研究として、純粋関数型言語をベースとしたデータフロープログラミングの研究が挙げられる。基本的なアイデアは、無限を含む任意長のデータ系列をリストで表現し、データ系列の操作を、リストを操作する関数として記述することである。それらを組み合わせたパイプラインとしてシステムを記述し遅延評価を用いることで、抽象度が高くかつ効率的なコードを見通しよく記述することが可能となる。従来のソフトウエア工学分野の研究では例題として低レベルの信号処理を対象とするに留まっていたが、純粋関数型言語は汎用のプログラミング言語であり、より広い応用が可能である。本研究ではこれが連続音声認識システムの記述に効果的に応用可能であることを示す。

6-1 純粋関数型言語によるパイプラインプログラミング

図 2 に一つの機能コンポーネント F1、入力 X および出力 Y を持つパイプラインの例を示す。X および Y は任意の長さを持つ時系列データを表すリストである。対応する Haskell コードは以下になる。

```
listY = funcF1 listX
```

このコードにおいて F1 に対応する funcF1 は、個々の入力要素ではなく、任意長のリストを引数として受け取りそれに対して処理を行った結果をリストとして返す関数である。また図の X および Y はコード中ではそれぞれ listX、listY として表現している。図 3 は 2 つの機能コンポーネント F1 と F2 をつないだパイプラインの例である。F2 も F1 と同様に任意長のリストを引数として受け取り、その結果をリストとして返す関数である。図は F1 の出力が F2 の入力に継ることを示している。対応するコードは以下になる。

```
listZ = funcF2 (funcF1 listX)
```

パイプラインは分岐やループを持つことが可能である。図 4 にループを持つパイプラインの例を示す。対応するコードは以下の様になる。

```
listY = funcF3 listY
```

これら純粋関数型言語によるコードでは、データフローの構造が直接に関数の依存関係に対応する。また遅延評価を用いることで、例え入力が無限であっても必要に応じたメモリ量と計算コストで効果的に動作する。他方、手続き型言語を用いた場合は引数が関数に渡されるタイミングで評価されるため、常に入力長に比例したメモリと計算コストがその場で必要となる。それを避けるためには、コンポーネント間でのデータ授受のための付加的コードが必要となる。これはコンポーネントの独立性を低下させコードを複雑化させる。純粋関数型言語を用いた場合のメモリや計算コストは、各コンポーネントの計算内容に依存する。例えば入力リストの各要素を定数倍する場合必要なメモリは最小であるが、リスト要素の順序を逆転させるにはリスト長に比例したメモリが必要となる。これは純粋関数型言語の制限では無く、時系列データ自体の性質によるものである。

図 2 １つのコンポーネントを持つパイプラインの例

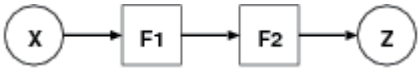

図 3 ２つのコンポーネントを接続したパイプラインの例

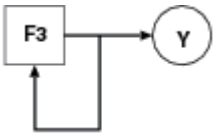

図 4 ループのあるパイプラインの例

6-2 音声認識システムのパイプライン分解

音声認識は音声信号 X を受け取り、単語系列 W を出力するシステムである。したがって、トップレベルでは図 5 に示すパイプラインにより表現される。図で O は X より抽出された特徴量ベクトルの時系列である。特徴量の抽出は古典的な信号処理の範疇であり、FeatureExtraction コンポーネントは FFT や DCT などの機能コンポーネントを直線状につなげたパイプラインとしてさらに分解して実現出来る。認識処理の中核は探索部 Search であるが、例えば WFST に基づくビタビビーム探索の場合は図 6 および図 7 に示すような分解が可能である。図 6 が前向き探索部であり、図 7 が漸進的な結果出力をサポートした逐次バックトラック部である。

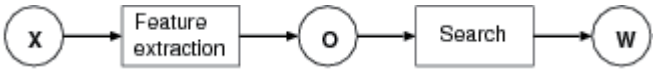

図 5 音声認識システムのパイプライン表現(トップレベル)

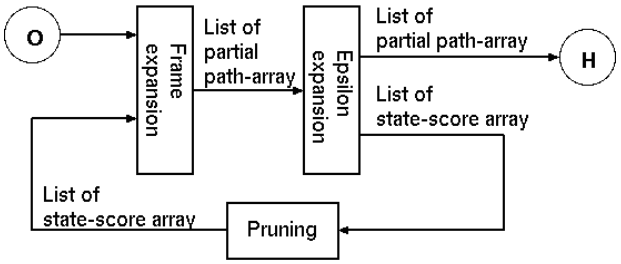

図 6 ビタビデコーダの前向き探索のパイプライン

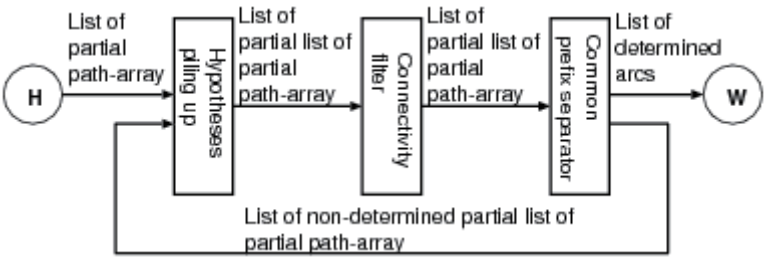

図 7 逐次バックトラッキングのパイプライン

6-3 実験条件と実験結果

パイプライン分解に基づき husky の再設計を行った husky2 を実装した。グラフ操作など、一部の機能については専用ライブラリを開発した。評価実験は、日本語話し言葉コーパス CSJ を用いて行った。連続音声認識実験において CSJ 標準テストセット 1 に対して 81.1%の高い単語認識精度を得た。これは同タスクでの T3 デコーダ[6]や旧版の Husky の最高認識率と同じであり、Julius[7]よりも高い精度である。必要メモリ量も、旧版の Husky と比べ大幅に削減された。

【参考文献】

[1] 岩宮眞一郎、”サイン音の科学-メッセージを伝える音のデザイン論-、”コロナ社、2012

[2] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, “The HTK Book,” Cambridge, U.K.: Cambridge Univ. Eng. Dept., 2009.

[3] A. Maas, Q. Le, T. O’Neil, O. Vinyals, P. Nguyen, and A. Ng,“Recurrent neural networks for noise reduction in robust ASR,” in Proceedings of INTER- SPEECH, 2012.

[4] T. Nishiura, M. Nakayama, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda, and S. Nakamura, “Evaluation framework for distanttalking speech recognition under reverberant environments—newest part of the censrec series—,” Proc. LREC’08, May 2008.

[5] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,”Acoust Soc Jpn E, vol. 20, no. 3, pp. 199—206, 1999.

[6] P. R. Dixon, D. A. Caseiro, T. Oonishi, and S. Furui, “The titech large vocabulary wfst speech recognition system,” in Proc. IEEE ASRU, 2007, pp. 443-448.

[7] A. Lee, T. Kawahara, and S. Doshita, “An efficient two-pass search algorithm using word trellis index,” in Proc. ICSLP, 1998, pp. 1831-1834.

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
音声認識システムのパイプライン分解と遅延評価を用いた実装法	日本音響学会 2012 年秋季講演論文集	2012 年 9 月
純粋関数型コンパクトデコーダ Husky2 の性能評価	日本音響学会 2012 年秋季講演論文集	2012 年 9 月
Pipeline Decomposition of Speech Decoders and Their Implementation Based on Delayed Evaluation	Proc. APSIPA	2012 年 12 月
複数分析窓長を用いた Autoencoder に基づく残響除去の検討	日本音響学会 2013 年春季講演論文集	2013 年 3 月
Reverberant Speech Recognition Based on Denoising Autoencoder	Proc. Interspeech (To appear)	2013 年 8 月