6P−6 自動通訳システム INTERTALKER における音声認識

吉田 和永　古賀 真二　磯谷 亮輔　塚田 聡　高木 啓三郎　畑崎香一郎　渡辺隆夫
(日本電気(株) C&C情報研究所)

1. はじめに

自動通訳における入力部としての音声認識は、誰もが様々な環境で自然に発声した声を受け付けられることが望ましい。このためには、不特定話者、大語彙、連続音声への対応、耐雑音化等の高度の音声認識機能が望まれる。

筆者らは大語彙・不特定話者向きの音声理解の実現を目指して、半音節を認識単位とする連続音声認識の検討[2]を行ってきた。今回これらの検討に基づき、自動通訳システム INTERTALKER[1]の認識理解部用に不特定話者連続音声認識システムを実験試作したので報告する。

2. 認識方式

試作した認識システムは、自動通訳のために望まれる音声認識の機能それぞれについて、できる限り高度に実現することを目標とした。以下に認識システムの仕様を示す。

1. 不特定話者・連続音声入力　誰でも自然に音声入力できるための最も基本的な機能である。
2. 限定タスク　タスクを設定し音声入力する文の表す意味内容を限定した上で、その意味内容の表層的な表現、即ち、言い回し、語順、語の省略等については極力自由度をもたせている。
3. 概念表現出力　認識結果として、多言語翻訳に適した概念表現を出力する。
4. タスク独立学習　タスクの設定を自由に行なえるよう、タスクに用いる語彙とは独立に音声の学習を行う。
5. 実時間認識　通訳を介した会話を円滑に行うため、実時間で認識処理を行う。
6. 耐雑音　実際の使用環境で起こり得る雑音下でも安定した性能が得られるようにしている。
7. リジェクト機能　不要な発話を認識対象外としてリジェクトする機能を搭載している。

認識方式は、半音節を認識単位とした構文ネットワーク制御による不特定話者連続音声認識方式[2]を基本としている。図1にシステムの構成を示す。分析部には耐雑音処理が組み込まれている。学習処理では、多数話者の学習データを用いてforward-backwardアルゴリズムによる学習を行い不特定話者の半音節HMMを求める。タスクを記述した構文ネットワークおよび単語辞書と半音節HMMとからHMMネットワークをあらかじめコンパイルしておく。同時に、概念生成に必要な情報を依存関係テーブル[3]に記述しておく。

Speech Recognition in an Automatic Interpretation System INTERTALKER, by Kazunaga Yoshida, Shinji Koga, Ryosuke Isotani, Satoshi Tsukada, Keizaburo Takagi, Kaichiro Hatazaki, and Takao Watanabe (C&C Information Technology Research Laboratories, NEC Corp.)

認識時には、Viterbiアルゴリズムを用いてネットワークの最適経路を求める。適応尤度補正、リジェクト判定を行った後、依存関係テーブルを参照して対応する概念表現を作成し、認識結果として出力する。

耐雑音

雑音除去のため、2段スペクトルサブトラクション(SS)法[4]を用いている。本方法では、音声入力用、雑音入力用の2チャンネルのマイク入力を採用し、スペクトル領域上で、定常雑音除去、非定常雑音除去の2段階の処理が行われる。

半音節HMM

少ない種類ですべての音素間の遷移を表現できる認識単位として「半音節」を用いる[5]。半音節は、図2に示すように、基本的に、音節をその母音中心で分割したものである。CV,VCセグメントに加えて、連続母音、促音、無音を表すセグメントの241種類のセグメントが用意されている。半音節単位は、音素や音節より種類が増えるが、すべての音素間の遷移を表現できる点で有利である。また、VCV, CVC等と比較すると種類が少なく、限られた学習データから効率的な学習を行うのに適している。

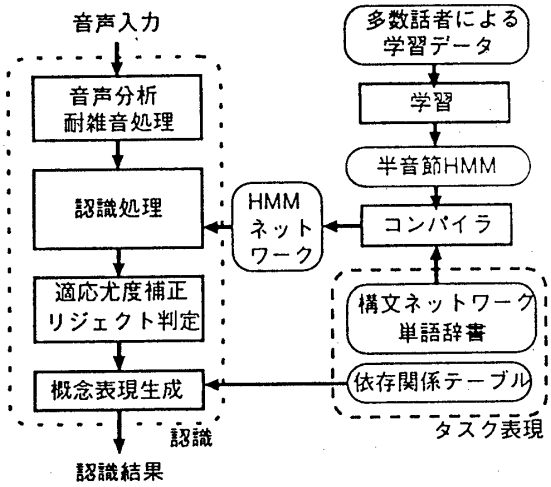

図1: システム構成

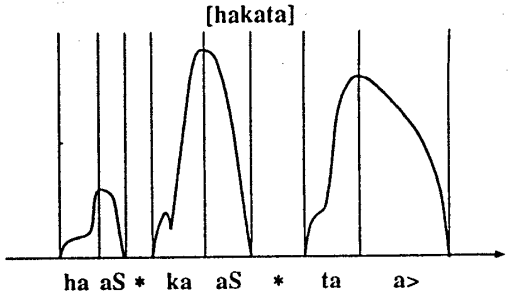

図2: 半音節セグメント例

各半音節は、left-to-right型のHMMモデルにより表現される。推定すべきパラメータ数を少なくして学習を安定化させるため、無相関ガウス分布とした。不特定話者における話者による変動を表現するため、少数混合数の混合ガウス分布を用いている。

連続音声認識

あらかじめコンパイル処理により、構文ネットワーク、単語辞書、半音節HMMとから単一のHMMネットワークが展開される。この時、単語中に無声化等の発声変形の可能性のある単語については、ルールに基づきネットワーク中に分岐として展開しておく。また、各単語の間には、単語間を滑らかにつなぐモデルとして単語間半音節モデルを挿入する。

認識は、展開されたHMMネットワーク上でViterbiサーチにより最適経路を求めることにより行われる。処理高速化のため、フレーム同期サーチを更に高速化したbundleサーチ[6]を用いている。これは、同一単語が構文ネットワーク中の複数箇所で出現するとき、その単語に至る累積尤度の初期値が最大の出現箇所に対してのみ単語内サーチを行い、残りの出現に対しては初期値をもとに単語終端における累積尤度を推定するものである。複数の出現に対する単語内サーチを1回にまとめることができるので、処理量を大幅に削減することができる。

リジェクト処理

認識対象外の入力に対する高精度なリジェクト機能の実現のため、適応尤度補正法[7]を用いている。これは、認識対象を限定しない条件、即ち、音節列としての認識を行った時の尤度を参照用尤度として用いて、認識タスクの構文制約のもとで得られた通常の尤度を適応的に補正し、この補正尤度を用いてリジェクト判定を行う。これによりリジェクト判定を困難にしていた、話者や発声環境の影響による尤度の変動を取り除くことができる。

3. 認識ハードウェア

前章で述べた認識処理を実時間で実行するための認識ハードウェア[8]を構築した。システムは、図3に示されるように、疎結合型バス結合の24個のマルチプロセッサを中心に構成される。各演算プロセッサは32bit浮動小数点演算DSPおよびローカルメモリ(512KW)を持ち、FIFOを介してバスと結合されている。FIFO間の転送を行なう転送プロセッサ、ブロードキャスト送信機能等により演算プロセッサ間のデータ転送の高速化を図っている。連続音声認識処理を、半音節HMMの各状態の出力確率の計算、単語内サーチ処理、構文ネットワーク上での最適経路のサーチの、入力フレームに同期してパイプライン処理可能な3つの処理に分割する。さらに、タスクの規模に合わせて、各処理を複数のプロセッサに搭載することにより、効率の良い並列処理が可能となる。

4. 評価実験

チケット予約タスク(単語数500、単語パープレクシティ5.5)を用いて評価実験を行なった。不特定話者の半音節HMMの学習には、85名が発声した音素バランス250単語を用いた。学習話者に含まれない10名の発声した日本語タスク文音声に対し、文認識率は平均83.0%で、文の概念表現が発話者の意図と一致した割合(意味理解率)は93.0%であった。Bundle処理の導入により、総状態数は3分の1に、サーチ処理時間は30%に削減されたが、これによる文認識率の低下はほとんど見られなかった。また、2段SS法の適用により、S/N15dBのノイズ下の文認識率を38.3%から71.3%に、適応尤度補正法により、認識対象外文の正リジェクト率を58.6%から95.2%に向上させることができた。さらに、認識ハードウェアにより、プロセッサ数19個で実時間認識処理が可能となっている。

5. おわりに

自動通訳システム用に不特定話者連続音声認識システムを試作した。単語数500語のタスクを用いて評価を行い、意味理解率93.0%の良好な認識性能と、リアルタイム応答を確認した。

謝辞　日頃ご指導いただく亘理メディアテクノロジ研究部長、ご協力いただいたメディアテクノロジ研究部、日本電気技術情報システム開発(株)の諸氏に感謝します。

参考文献

[1] 畑崎他,「日英双方向自動通訳システムINTERTALKER」,本予稿集,6P-5(1992.3)

[2] 渡辺他,「自動通訳のための不特定話者連続音声認識システム」,信学技報,SP91-(1992.1)

[3] 野口他,「概念表現を用いた自動通訳システムINTERTALKER」,本予稿集,6P-7(1992.3)

[4] 高木他,「2段スペクトルサブトラクションによる雑音下音声認識」,音学講論,pp.59-60(1991.3)

[5] 渡辺他,「半音節を単位としたHMMを用いた大語彙音声認識」,信学論,J72-D-8,pp.1264-1269(1989)

[6] 渡辺他,「Bundleサーチによる連続音声認識のための高速化手法」,音学講論,pp.125-126(1991.3)

[7] 塚田他,「未知語検出・リジェクションのための音声認識の尤度補正」,音学講論,pp.203-204(1991.3)

[8] 古賀他,「不特定話者連続音声認識用ハードウェアの開発」,信学技報SP91-(1992.1)

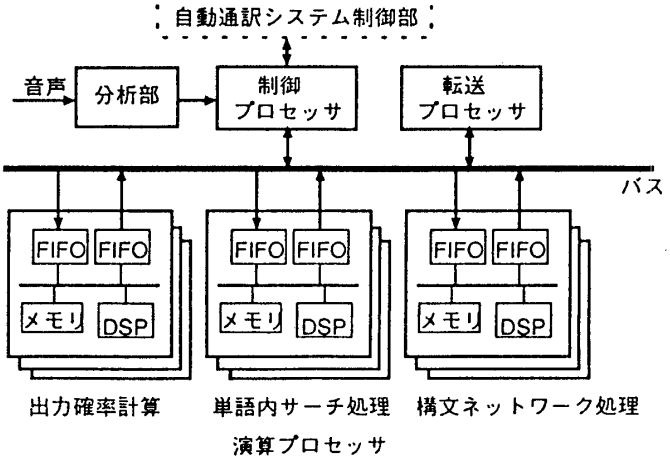

図3: 音声認識ハードウェア構成