インターネット上で利用可能な音声認識サーバの構築*

4D－1

田岡典尚　中村 哲　鹿野清宏
奈良先端科学技術大学院大学 情報科学研究科

1　はじめに

多くの人々がインターネットを通じ、情報を公開したり、外部の情報を簡単に得ることができるようになった。その提供される情報の媒体は、テキスト文書であることが多いが、それ以外にも音声や、画像のような多種類のメディアが用いられている。

しかし、メディアは多いが、音声の利用に限ってみれば「音声を流す/聞く」という利用しかされておらず、音声を処理して新たな情報を生み出す、といった利用はされてない。

そこで、「音声を用いたサービス」として音声認識を題材として選び、インターネット上で利用するための音声認識システムを構築した。

2　音声認識システムの構成

2.1　条件

インターネット越しで利用可能な音声認識システムを構築するためには次のような条件が考えられる。

1. クライアント・サーバ型である
2. データ転送プロトコルは、転送順番を保証するものでなければいけない
3. クライアントに情報を返すまでの時間は速くなければいけない

本提案システムでは、上記の1,2,3に対し、以下の手法を用いて対処した。

2.2　システム構成

2.2.1　クライアント・サーバ

本システムでは、インターネット上の計算機から利用可能とするために、Tied Mixture HMM を利用した音声認識システムをクライアント・サーバ型で構築している。

複数要求に対する処理

通常サーバは通常クライアントから要求が来ると、fork 関数を用いて新しくプロセスを生成し、処理にあたらせる。しかし、多数のプロセスが同時に走っていた場合、メモリ不足やCPUの割当時間の減少による処理時間の遅延が発生してしまう可能性がある。

本システムでは、認識要求がある一定数を越えるような場合には、新たなプロセスの生成を行わず、一つのプロセスが各クライアントの要求を時分割で処理するように切り替わる。

サーバのプロセスの様子および、データの流れを図1に示す。

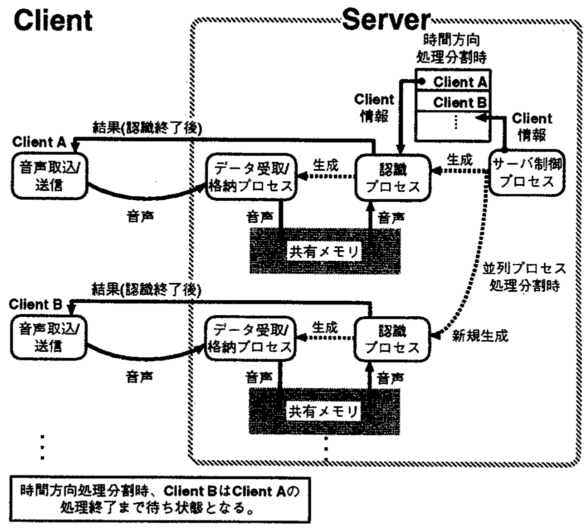

Fig. 1: クライアント・サーバ間処理

2.2.2　通信プロトコル

クライアント・サーバ間のデータ転送のための、通信プロトコルはTCP/IPを使用している。

2.2.3　処理の高速化

逐次転送処理

本システムでは、認識プロセスとは別にslaveプロセスを立ち上げている。このプロセスはクライアントからのデータを受け取り、共有メモリに格納する処理を行っている。これによって、認識プロセスは必要となった時に、ローカルメモリから素早くデータを読み出すことができる。

ビームサーチ

ビームサーチを使用することで、最大尤度ノードの値より設定値以下の値になるようなノードを刈り取っている。

Tied Mixture の制限

認識処理を高速化させるために、出力確率計算に使用するTied Mixture全分布256個全てを使わないで、個数の制限を行っている。

---

*Speech recognition server for internet application, Norihisa Taoka, Satoshi Nakamura and Kiyohiro Shikano, Graduate School of Information Science, Nara Institute of Science and Technology;

フレームの間引き

パラメータのデルタ項の値が小さい時は、前フレームと同じものであるとみなして、認識処理を省略している。

3 実験・結果

3.1 データ条件

音声認識サーバへ転送する音声は、サンプリング周波数12000Hz、窓長32ms、シフト幅8msである。特徴パラメータは、MFCC 16次元、デルタ項16次元、パワー項1次元の計33次元で構成されている。学習モデルは、ATRデータベースにおける男性2名、女性2名の各2560単語で学習したものを使用している。

実験では、クライアントからサーバへ転送するデータとして、学習外の500単語を用いている。

3.2 実験内容

まずはじめに、ビームサーチの値の最適値の測定を行った。次にビームサーチで求められた値を使用して、Tied Mixtureの使用個数に制限を与えて測定した。なお、測定時間はDEC 3000/300のマシン上で、音声データ受け取りから、特徴分析、Viterbi計算、ビームサーチまでの1フレームあたりの認識処理時間である。

3.3 結果、および考察

実験結果を図2に示す。

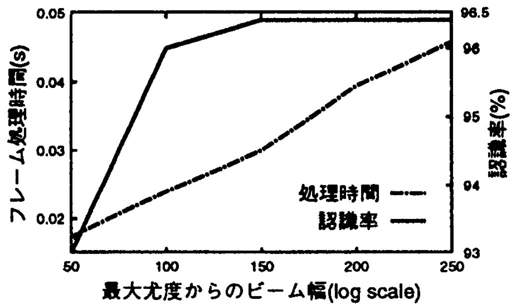

Fig. 2: ビームサーチの幅と処理時間

値を150に設定すれば、認識率をあまり下げないで、処理時間の短縮をはかることができた。しかし、この状態のままであれば、8msシフト幅の音声に対して、処理時間が30msを越えている。

この実験では、Tied Mixtureの計算において一定の閾値よりも出力確率値の小さい分布は除いている。この分布数は平均で120程度であるが、さらにこの分布の使用個数を制限することを試みた。

実験結果を図3に示す。

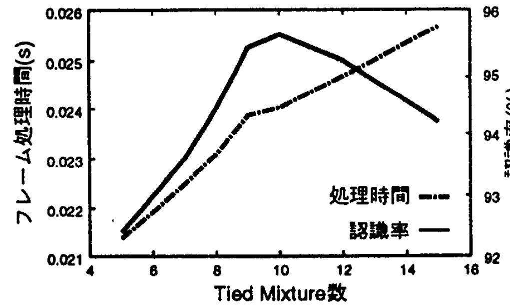

Fig. 3: Tied Mixtureの使用個数と処理時間

この結果から認識率、処理時間を考慮すると、Tied Mixtureの使用個数の値を10個に設定することが適当であると考えられる。しかし、認識時間はシフト幅8msの3倍かかっているため、さらにフレーム間引きを導入した。認識率1%程度の劣化で半分近いフレーム間引きが可能となっている。これにより、1フレームあたりの認識処理時間は、シフト幅のほぼ1.8倍となった。最後に、全体の時間的流れを図4に示しておく。

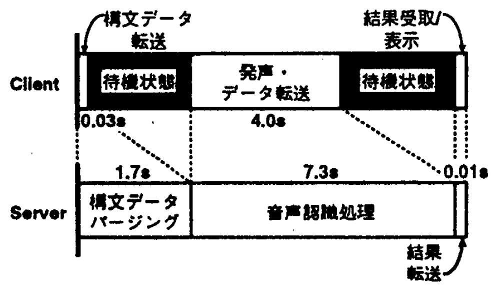

Fig. 4: 時系列フロー (CPU time)

4 おわりに

インターネット上で音声認識を利用してもらうためには、さらに認識速度、認識性能の向上をはからなければいけない。また同時に、音声認識を利用するための、様々なAPIの設計も行っていく必要がある。

参考文献

[1] H.Singer,T.Beppu,A.Nakamura,Y.Sagisaka, "A Modular Speech Recognition System Architecture", 音学講論, 2-8-1 (1994-10)

[2] 山田, 野田, 嵯峨山, "実時間動作を考慮した音声認識サーバ", 音学講論, 2-8-2 (1994-10)