

外97-60

早稲田大学大学院理工学研究科

博士論文概要

論文題目

A Study on Speech Coding for Asynchronous
Transfer Mode Networks and Speech Recognition
for the World Wide Web
ATM 網用音声符号化および WWW 用音声認識に
関する研究

申請者

近藤 和弘

Kazuhiro Kondo

1998 年 / 月

第1章 序章

音声は人間が持つ最も基本的な意志伝達手段である。技術の進歩により、人間対人間のみでなく、計算機とも音声を通じて通信できつつある。また、人間対人間の音声通信でもより遠くに、より便利に通信できるようになりつつある。本研究では音声信号処理を利用した上記のシステムの研究を行った。特に、①パケット伝送に適した音声符号化・多重化方式の検討②音声認識における単語間音素環境の高効率モデル化手法の検討③音声認識を用いた音声により制御可能なWWWブラウザの検討を行った。

第2章 ATMネットワーク用音声パケット伝送システムの開発

現在盛んに検討が進められている次世代通信網においては、各種情報がセルと呼ばれる固定長パケットで統一的に多重化・伝送される Asynchronous Transfer Mode (ATM)が用いられる。ATMではセルを一旦バッファに格納しておく必要があるため、セル発生量が多い場合はバッファから溢れて情報損失が生じる場合があり、また伝送遅延も大きく変動する。ATMの公衆網への本格導入は、初期投資が大きいこともあり、少なくとも4、5年先になると思われる。私設網においては投資規模を限定でき、かつATMの利便性、経済性を追求できるため、まず私設網でATMの有効性が試されるものと思われる。しかし、私設網ではトラフィック規模が小さいため、トラフィックの変動が平均化されにくく、特定ノードにトラフィックが集中した場合は、バッファオーバーフローによりセルが失われる場合がある。また伝送遅延も変動するため、音声のような実時間情報では再生に間に合わずに廃棄せざるをえないセルが生じる。本研究では、このように廃棄の多い状況下でも劣化の少ない音声符号化・伝送方式を検討した。

パケット状の伝送では比較的簡単に伝送優先度付けができるため、2レベル程度のものを仮定した。この様な多優先度伝送・高廃棄率条件下で劣化の比較的小さい符号化方式としてはエンベデッド Adaptive Differential PCM (ADPCM)方式が知られている。そこでエンベデッドADPCM出力符号を半分に分け、予測に用いる上位半分と予測に用いない下位半部分を別セルで伝送することを検討した。しかし、以下に考察する理由で劣化が多すぎる事が判明した。

- 差分符号の内、上位2ビットで予測を行ったが、特に有声部分で予測精度が不足する。

- 下位部分を収めたセルが欠落すると、バースト状にノイズが増加するため、主観的に目立つ。

これに対応するため、以下の可変レート符号化方式を提案した。

- 有声部分は6 bit/sample, 48kbpsで符号化し、上位3 bitは予測フィードバックかつ高優先度セルで伝送する。下位は低優先度セルで伝送する。

- 無声部分は4 bit/sample, 32kbpsで符号化し、上位、下位2 bitづつセル化する。

- 有声/無声の判定には予測利得を用いる。

- 無音部分は2 bit/sample, 16kbpsで符号化し、低優先度で伝送する。

以上の符号化・伝送方式を主観評価したところ、10%の廃棄率でもほぼ原音と同等の音質が保たれることが分かった。また、デジタル信号処理プロセッサを用いて実時間で会話可能なシステムを試作し、良好な音質を確認した。

第3章 通信ネットワーク用音声認識システムの開発

1. 単語間音素環境の高効率モデル化手法の開発

柔軟な語彙の音声認識を可能にするには、音素単位のモデルを用意しこれを連結することにより必要な単語モデルを作成する必要がある。しかし、音素の特性は隣接する音素により強い影響を受け変化する。このため隣接音素の組み合わせ毎に別々にモデルを作成することが一般的である。これを音素環境依存モデルと呼ぶ。音素環境依存性は単語内のみでなく、単語間でも見られる。一般に単語間の音素環境の組み合わせは単語内の数倍にもなり、必要となるモデル数が膨大になる。本研究では、このモデル数を大幅に削減する方法を検討した。

一般に単語間の隣接音素の影響は単語内より少ないと見ることができ、よって単語間の音素環境は単語内ほど詳細な区別が必要ないと考えられる。そこで単語間の音素環境をポーズ、子音、母音と極端に少ない種類のクラスにクラスタ化したモデルを作成した。これによりクラスタ化されない場合に比べ、必要な音素環境依存モデルは半分に削減された。認識精度はいずれのモデルも単語間音素環境を考慮しない場合に比べ32%の向上が見られ、モデル間の有意差は認められなかった。

また、学習セットと異なる任意の語彙の音声認識に利用する場合を考慮し、単語間音素環境に可能な全組み合わせに対するモデルをあらかじめ作成しておく手法も検討した。いずれの場合も単語内の音素環境も単語間環境と同等とみなし、単語間音素環境の学習に利用した。①クラスタ化していない音素環境モデル・セットから、クラスタ内環境に相当するモデルを選び、学習セット内での発生頻度で重み付けして平均し、クラスタ化モデルを得る方法、②学習セット内の全音素環境をクラスタに分類し、改めてクラスタ化音素環境モデルを学習し直す方法を比較した。いずれの方法も有意差はなく、クラスタ化しないモデルに比べ10%程度の精度劣化で抑えられることを示した。

2. 音声認識を用いた World Wide Web(WWW)ブラウザ

WWWのページ数、トラフィックが爆発的に増大している。急激にユーザ層が広がったため、大多数のユーザは初心者である。そのためキーボード、マウスの操作に不慣れなユーザも多い。このようなユーザや一部の身障者にとって、音声はWWWブラウザをより使い易いものにする可能性がある。またコンピュータに

慣れたユーザにとっても音声を用いたハンズフリー操作、マルチメディア・プレゼンテーション等高度な使い道を提供できる可能性がある。このような背景から、日本語音声で制御可能なWWWブラウジング・システムを検討し、試作評価した。

本システムは、以下の機能をWWWブラウザに与える。

- 音声コマンド：ブラウザのコマンドを音声で与える（ページめくり、スクロール等）。

- 音声リンク：WWWページ上のリンク名を読み上げることで、そのリンク先に進むことができる。

- スマート・ページ：WWWページに音声認識用文法へのポインタを埋め込んでおくことにより、そのページで柔軟な文型を用いた制御が可能になる。

- 音声ブックマーク：ユーザ好みのWWWページをユーザ定義のキーワード（キーワード）を随時読み上げることによりこのページへ進むことができる。

音声リンクを実現するためにはWWWページのハイパーテキスト言語 (HTML) を解析し、ハイパーリンクが付与されているアンカーテキストとリンク先のアドレスを指す Uniform Resource Locator (URL) を抽出する必要がある。リンク名は単語単位に分割した上で、音素に変換する必要がある。この処理において、以下の点を考慮した。

- 日本語アンカーテキストの分割には本格的な形態素解析を用いなくても、辞書に元々ある単語単位の項目に加え、紛らわしい形態素単位の項目をある程度追加登録すれば、冗長ながら、実用範囲精度の分割がほぼ実時間で行えることが分かった。

- 日本語のページでも4割以上は英語を含む。よって英文にも対応する必要がある。本システムでは英文も辞書に登録・検索できるようにし、小語彙ながら頻出する英単語を登録することにより、ある程度の英文リンクにも対応できるようにした。

- 音素変換においては、広範囲な例外処理が必要である。特に数字に関するものが多い。これらは全て規則として登録した。また、記号も読み飛ばしを許すものを規則として登録し、読み方に変化があるものは全て辞書に登録した。

本システムを UNIX ワークステーション上に試作し、12名の被験者に各々30分程度ブラウジングさせたところ、91%のタスク達成率を測定した。ユーザが操作不慣れなため発生するエラーを除くと、達成率は94%以上になった。

第4章 結論

通信網への適用を目指した音声符号化・伝送システム、および音声認識システムを検討し、評価した。今後は音声のみならず、画像等を含めた各種信号の統合による利点を活かしたマルチモーダルなシステムを目指したい。