携帯電話試作機上で動作する旅行会話向け音声認識システム

花沢　健　　荒川　隆行　　岡部　浩司　　辻川　剛範
長田　誠也　　磯谷　亮輔　　奥村　明俊
NEC 共通基盤ソフトウェア研究所

1. はじめに

我々は、常に携帯していつでもどこでも手軽に使える通訳システムの実現を目指してきた[1,2]。今回、携帯電話の試作機上で単体動作する日英の旅行会話通訳システムを開発した。本システムは、サーバとの通信を必要とせず、端末単体で、日本語の音声を音声認識し、認識結果を機械翻訳して英語の翻訳結果を出力する、またはその逆を行うものである。本稿では、本システムの構成要素として開発した音声認識部分について報告する。

本音声認識システムの特徴は、数万語規模の大語彙連続音声認識をコンパクトかつ高精度に実行することである。市販されている携帯電話と同等のスペックを持つ試作機において、単体で快適なレスポンスを実現している。

2. 通訳システムの構成

携帯電話試作機向け通訳システムの構成としては、図１に示すように音声認識部と機械翻訳部を言語ごとに用意し、それらを統合部にて連携・制御することにより実現している[3]。

3. 音声認識システムの開発

3.1. 音声認識システムの構成

図１における日本語または英語音声認識部としては、図２に示すように、音声を入力として認識結果である単語列を出力する、大語彙連続音声認識システムを用いる。本システムは、音声認識エンジン本体と、音響的確からしさを与える音響モデル、言語的確からしさを与える言語モデル、および認識対象である単語の辞書とから構成される。ここで、音声認識エンジン自体は言語非依存であり、モデルや辞書のリソースを切り替えることで言語の切り替えが可能である。

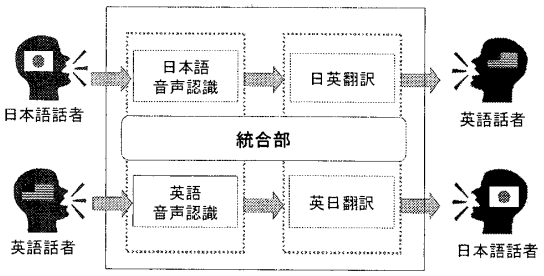

図１．通訳システムの構成

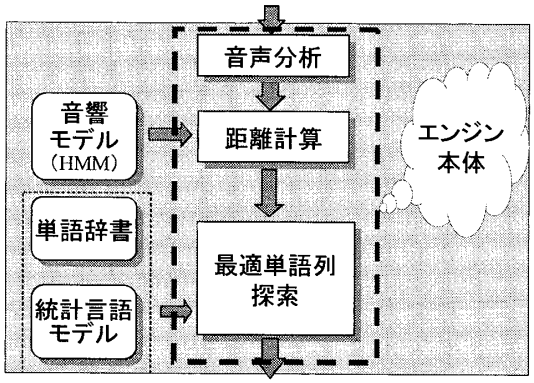

図２．音声認識システムの構成

3.2. 音声分析

入力音声を音声認識に適した特徴量系列に分析する音声分析部は、サンプリング周波数 11kHz の音声を、ケプストラムを含む特徴量に音声分析する。

さらにここでは、MBW 法[4]による雑音抑圧を行う。本手法は、雑音抑圧において、雑音だけでなく音声の知識を用いることにより、雑音の種類によらずに頑健に雑音の影響を軽減するものである。本手法により、街頭雑音を SNR10dB で重畳した場合に、従来法と比較して約 10%の誤り率削減が得られている。

3.3. 認識エンジン

音声認識エンジンとしては、当社独自開発のコンパクトで高速な音声認識エンジン[5]を使用する。コンパクト化のために、MDL 基準を用いた音響モデルの混合ガウス分布数の削減、ガウス分布の対角共分散行列の共有化、単語終端テーブルのガベージコレクションを行っている。高

Development of a Speech Recognition System Prototype for Cellular Phones
Ken Hanazawa, Takayuki Arakawa, Kouji Okabe, Masanori Tsujikawa, Seiya Osada, Ryosuke Isotani
NEC Common Platform Software Research Laboratories

速化のためには、木構造を利用した音響モデルの効率的な出力確率計算、単語終端における言語スコア計算結果の再利用を行っている。

さらに、コンパクトな環境での認識精度の劣化を低減するため、言語モデル先読み値の平滑化手法[6]を導入した。これにより、試作機においても速度の劣化なく、従来認識が難しいとされていた数種類の単語について精度が向上することを確認している。

3.4. モデル構築

音声認識エンジンで用いる音響モデル・言語モデルは、大量の音声あるいはテキストコーパスから統計的に学習する。日本語は標準語、英語は米国語を対象としている。

音響モデルは、日英とも 600 時間以上の音声コーパスを用いて不特定話者・性別非依存の状態共有 triphone モデルを学習した。

言語モデルは、日英とも数十万文規模の旅行会話テキストコーパスを構築し、単語 Ngram モデルを学習した。認識辞書は、テキストコーパスに出現するものをベースに、頻度情報を用いることでより少ない語彙でより広いコーパスカバー率が得られるよう工夫し、日英とも約３万語規模の辞書を作成した。

3.5. 携帯電話試作機搭載

携帯電話試作機にはミドルウェアとして実装した。各計算処理は固定小数点化を行うことで高速処理を実現している。また、言語モデルなどサイズが大きいモジュールは起動時の読み込みに時間がかかるため、プログラムに埋め込んでいる。

4. 評価

シミュレーションによる認識精度の評価を行った。評価データとしては、旅行会話の読み上げ発声データを用いる。評価データは男女各 10 名、合計約 900 発声である。評価の結果、日本語音声認識、英語音声認識ともに単語正解精度で９割以上の認識率が得られた[7]。

市販の携帯電話と同等のスペック（動作周波数 500MHｚ）の携帯電話試作機上で動作させたところ、動作速度については日本語音声認識、英語音声認識ともに、数秒程度の発声に対して発声終了から１秒以内に認識結果出力が確認できた。ほぼリアルタイムで動作していると言える。このとき、メモリ使用量は最大でも 10MB であった。また、音声認識システムの起動時間は約２秒となっており、高速な起動が実現できている。

さらに、日本語音声認識においてオンライン評価を行った。静かなオフィス環境において、8 名の話者が試作機を実際に操作しながら旅行会話文を発声した。評価文は、各自 20 文を２回ずつ、計 40 発声した。その結果、文レベルの完全一致では 73.4%、助詞の違いなど意味が同じ場合を正解とすると 79.4%の認識率が得られた。

5. まとめ

本稿では、携帯電話試作機上で単体動作する旅行会話向け日英通訳システムのための、音声認識システムについて説明した。本音声認識システムは、認識精度を高く保ちつつ、コンパクトで高速な数万語規模の大語彙連続音声認識を実現している。

シミュレーションによる評価を行ったところ、日本語音声認識、英語音声認識ともに９割以上の単語正解精度が得られることを確認した。さらに市販の携帯電話機と同等のスペックを持つ試作機上での評価を行ったところ、数万語規模の辞書を使用しながらリアルタイムに処理可能であること、使用メモリ量は 10MB 以下に抑えられることが確認できた。日本語音声認識をオンライン評価したところ、文レベルで７～８割の正解率が得られることを確認した。今後、さらなる実用可能性について検討したい。

参考文献

[1] 山端他，“PDA で動作する旅行会話向け日英双方向音声翻訳システム”，情処研報，2002-NL-150-9，2002.

[2] 山端他，“低消費電力マルチコアプロセッサで動作する日英自動通訳システム”，情処全国大会，4B-2，2006.

[3] 長田他，“携帯電話試作機上で動作する旅行会話向け自動通訳システムの開発”，情処全国大会，2D-2，2009.

[4] T. Arakawa, et al, "Model-Based Wiener Filter for Noise Robust Speech Recognition", Proc. of ICASSP 2006, Vol. I, pp.537-540, May 2006.

[5] 磯谷他，“話し言葉認識に向けた基本技術と応用”，情処研報，2005-NL-169，2005.

[6] 岡部他，“言語モデル先読み値の平滑化による探索誤りの改善”，音講論集（秋），1-1-15，2008.