音声言語情報処理 45－22
（2003．2．8）

汎用音声認識エンジンJulius/Julianの PDAへの移植と性能評価

原 直* 河口 信夫† 武田 一哉‡ 板倉 文忠‡

*名古屋大学工学部
†名古屋大学情報連携基盤センター
‡名古屋大学大学院工学研究科

〒464-8603 名古屋市千種区不老町1
*hara@itakura.nuee.nagoya-u.ac.jp †kawaguti@itc.nagoya-u.ac.jp
‡{takeda,itakura}@nuee.nagoya-u.ac.jp

あらまし 携帯情報端末(PDA)での音声認識の研究を進めるために，汎用音声認識エンジンJulius/JulianをPDA上に移植し，その性能評価を行った．その結果，孤立単語音声認識では，Linux上で動作させた場合に比べて認識率はわずかに低下するが，PDAで実時間の約1.9倍で90%の認識率を得ることができた．連続数字音声認識ではPDAで収録した音声で学習することにより，約99%の認識率を得た．

Implementation and evaluation of Julius/Julian on the PDA environment

Sunao HARA*, Nobuo KAWAGUCHI†, Kazuya TAKEDA‡ and Fumitada ITAKURA‡

* School of Engineering, Nagoya University
† Information Technology Center, Nagoya University
‡ Graduate School of Engineering, Nagoya University

Furo-cho 1, Chikusa-ku, Nagoya 464-8603, JAPAN
*hara@itakura.nuee.nagoya-u.ac.jp †kawaguti@itc.nagoya-u.ac.jp
‡{takeda,itakura}@nuee.nagoya-u.ac.jp

Abstract In order to develop an open source platform of the speech recognition on Personal Digital Assistant(PDA), a general-purpose speech recognition engine Julius/Julian is ported to the PDA environment. From the experimental evaluations the following results are obtained. In the isolated word recognition, 90% accuracy is obtained by about 1.9 times of real time, that in about 73 times to a standard PC environment. In the connected digit recognition, 99% word accuracy is obtained using HMMs trained by sentences recorded by PDA.

1 はじめに

近年，PCの小型・軽量化が進んでおり，その一つの到達点として情報携帯端末(Personal Digital Assistant, PDA)があげられる．PDAはノート型PCよりも小型・軽量であり，また起動時間も早いことから主にPIM(Personal Information Manager)用途として用いられている．しかし，PDAはその大きさのためにキーボードは基本的に付属せず，文字の入力のためにはソフトウェアキーボードや手書き文字認識などを用いる必要がある．だが，これらの入力手段は通常のキーボードでの入力に対して，入力速度の点で劣る．PDA上で音声認識が可能となれば，有力な入力手段として利用できるであろう．

現在，PDA上での音声認識の研究も進んでおり，実際に製品化した例もあるが[1][2]，研究用として容易に利用できる物はない．そこで，今後PDAにおける音声認識に関する研究を進めるために，汎用音声認識エンジンJulius/Julian [3]をPDAに移植した．本報告ではこのPDA用に移植したJulius/Julian(以下，Pocket Julius/Julian)の性能を評価する．

また，現在PDAにおける音声認識に関する研究を進めるために，連続数字音声をPDAを用いて収録したデータベースを作成している．本報告ではデータベース作成の現状を報告するとともに収録・整理済みのデータを用いて，Pocket Julius/Julianの性能評価を行う．

2 システムの概要

Julius/Julianの実装には，Compaq iPAQ PocketPC H3970（以下，H3970）を用いた．OSはMicrosoft®Pocket PC 2002 (Windows CE 3.0)，CPUはIntel®XScale™PXA-250 400MHz，内蔵メモリは64MBとなっている．

音声収録用のPDAには，Compaq iPAQ PocketPC H3870（以下，H3870）を用いた．OSはMicrosoft®Pocket PC 2002 (Windows®CE 3.0)，CPUはIntel®StrongARM®SA-1110 Processor 206MHz，内蔵メモリは64MBとなっている．

また，Juliusの動作比較対象としてLinuxをインストールしたデスクトップPCを用いた．スペックは，OSはRedHat®Linux 7.3，CPUはIntel®PentiumIII 1.2GHz，メモリはSDRAM 512MBである．

3 Julius/JulianのPDAへの移植

バージョンRev 3.3p2について，PDA上に移植した．なお，移植・最適化の際にCELib(Build 3.12)[1]，Intel®Integrated Performance Primitives(Intel IPP) 2.1[2]をライブラリとして用いた．

今回，Juliusの最適化を行ったのは，音声から特徴量に変換する処理である．内部計算が浮動小数点演算で行われているものを固定小数点演算とし，処理速度の向上を図った．

また，本実装はIPPを使用しているため，CPUはStrongARM(SA-1110)もしくはXScale(PXA-250)のみで動作する．実際にH3870(StrongARM)，H3970(XScale)で動作することを確認した．IPPライブラリは，PocketPC2002用，Windows CE.Net用，Linux用の3種があるため，今回報告するPocketPC2002以外にも上記CPUを搭載したPDAならば他のOSに移植することは可能であると考えられる．

4 音声データベースの作成

収録用のPDAとして，H3870を用いた．収録は防音室（名古屋大学1B情報館4階板倉研究室）で行った．被験者は机上におかれた数字の書かれた原稿を読み，その音声をPDAと接話マイク(Sennheiser HMD410)で同時に録音を行った．PDAは防音室内の机の上に置いて収録している．被験者とPDAの間は約50cm離れており，頭の位置・向きについて細かい指示はしていない．収録の様子を図1に示す．

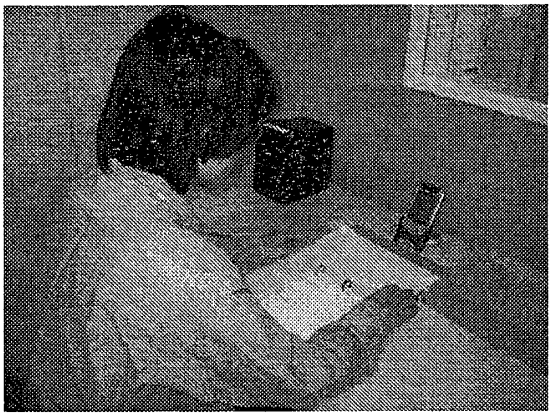

図1: 音声録音の様子

[1] http://www.rainer-keuchel.de/wince/celib.html
[2] http://developer.intel.com/design/pca/applicationsprocessors/swsup/IPPv2_1.htm

収録内容は1桁～12桁までの連続数字(Aurora2J [4])であり，被験者数は220名(男性110名、女性110名)である．また，各話者について，学習用セットは88文，評価用セットは49または50文である．

5 PDA内蔵マイク収録音の音響特性

5.1 PDAとコンデンサマイクロフォンの収録特性の比較

収録に用いたPDA(H3870)とコンデンサマイクロフォン(Sony ECM-77B)の周波数応答を求め，その対数振幅スペクトルのスペクトル歪み(Spectral Distortion, SD)を算出する．SD値は以下の式1で与えられる．

$$SD = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(20\log_{10}\frac{|H_1(f_k)|}{|H_2(f_k)|}\right)^2} \ [\mathrm{dB}] \quad (1)$$

測定は可変残響室(名古屋大学工学部7号館410号室)で行った．空中に固定されているスピーカ(BOSE ACCOUSTIMAS)よりTSP信号(5.94s)を放射し，マイクスタンドに固定したPDAにて，TSP応答を収録した．スピーカとPDAの画面との距離は300mmとした．参照音として，同条件下でコンデンサマイクロフォンを用いて測定したインパルス応答を使用した．測定条件を表1に示す．

表1: インパルス応答測定条件

測定用信号	TSP信号
信号長	262144 point (5.94s)
サンプリング周波数	44100Hz
同期加算	1回
室温	17.9℃(測定前) 19.1℃(測定後)
暗騒音レベル	18.1 dB(A)(測定前) 19.6 dB(A)(測定後)
音圧レベル	76.1 dB(A)

同様の収録をH3970，Fujitsu PocketLoox（本体内蔵マイク及び外部ヘッドセットマイクSony Ericsson HBH-30）についても測定しSD値を算出した．

5.2 実験結果

収録したTSP応答よりインパルス応答を求めた．PDAのマイクとコンデンサマイクのそれぞれのインパルス応答を256ポイント切り出し，それぞれのパワーで正規化し，256ポイントのFFTを行い対数振幅スペクトルを求めた．H3870で収録した場合の対数振幅スペクトルを図2に示す．また，式(1)

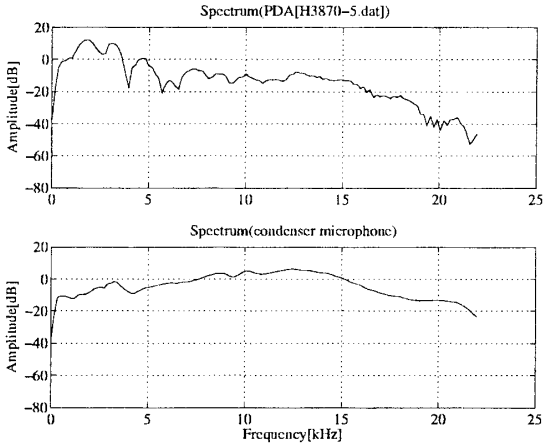

図2: PDA(H3870; 上)とコンデンサマイク(下)の対数振幅スペクトル

よりSD値を求めた．ただし，計算時には周波数帯域を344.53Hzから4134.38Hzに制限(N=23)しSD値を求めた．

表2: 各マイクにおけるSD値

マイクロフォン	SD [dB]
iPAQ H3870	7.3442
iPAQ H3970	4.1210
Pocket Loox	2.9010
HBH-30	9.8785

図2から，H3870で収録した場合，対数振幅スペクトルは大きく異なっていることがわかる，実際，H3870で収録した場合，SD値は非常に大きくなっている．

6 PDAを用いた認識実験

6.1 孤立単語音声認識の性能の評価

Julius を用いて孤立単語音声認識を行い，PDA(H3970) で動作させた Pocket Julius の性能を認識時間と単語認識率について評価する．比較対象として，Linux で PDA とほぼ同じ条件となるようにコンパイルし (configure 時に `--enable-setup=fast --enable-lowmem` とした)，時間測定ルーチンを組み込んだ Julius を用いた．

評価用音声は，CIAIRの車内音声データ[5]より，孤立単語発声(車内通常環境，アイドリング)のデータを用いた．50単語セット(表4)を話者20名(男性10名，女性10名)が発声した計1000語である．音響モデルは，IPAの標準日本語音響モデル[3]より，状態数129，性別非依存，monophoneモデルを用いた．混合数は2,4,8の3つに関して計算した．

言語モデルは，一文一単語の制約を課した擬似N-gramを用いて単語認識をするようにした[6]．

認識時間は，第一パス（ビームサーチ）のビーム幅（ノード数）を変化させ12回計測した．$(10,15,\cdots,45,50,60,80,100)$また，認識処理を，音声から特徴量への変換(WAV2MFCC)，第1パスの処理(PASS1)，第2パスの処理(PASS2)，の3つに分割して考え，その時間を計測した．

分析条件は，表3である．

表3: 音声認識のための分析条件

サンプリング周波数	16 kHz
分析窓	ハミング窓
フレーム長	25 ms
フレームシフト	10 ms
特徴パラメータ	MFCC(12次) Δ MFCC(12次) Δパワー(1次)

表4: 孤立単語音声認識に使用する単語

デジタルロッカー / 認証開始 / ２００１年１月１日 / 山田太郎 / 検索終了 / 暗証番号 / 0123 / 4567 / 8901 / 2345 / 6789 / コンテンツ / 映画 / 羊たちの沈黙 / サウンドオブミュージック / ゲーム / パックマン / 音楽 / ＪＰＯＰ / 今週のトップテン / ジャンル別検索 / ポップス / ロック / ビートルズ / 選曲 / イエスタデー / レットイットビー / 配信開始 / フェリー案内 / 時刻表 / 第二便を予約 / ネットニュース / トピックス / 音声読み上げ / 天気予報 / 交通情報 / 神奈川県 / 横浜市 / 中区 / 東京都 / 世田谷区 / 首都高速 / 東北自動車道 / セブンイレブン / ユニクロ / スターバックス / ホテル一覧 / パシフィックホテル / 予約表 / サービス終了

6.2 連続数字音声認識の性能の評価

Julianを用いて連続数字音声認識を行い，Pocket Julianの性能を単語正解精度(Accuracy)を用いて，評価する．Accuracyの算出には式(2)を用いた．

$$Accuracy = \frac{N-S-D-I}{N} \times 100 \ [\%] \qquad (2)$$

ただし，N:全単語数，S:置換誤り数，D:脱落誤り数，I:挿入誤り数である．

音響モデルは次の3種について計算を行った．

1. IPAの標準日本語音響モデル[3](性別非依存)
2. PDAの音声を用いて学習したモデル
3. 接話マイクの音声を用いて学習したモデル

3つとも，状態数129，混合数2のmonophoneである．IPAの音響モデルは、音素バランス文で学習しており、後者2つに関しては，男性14名，女性15名の連続数字発声 計2552文を用いて学習を行い音響モデルを作成した．音響モデルの作成にはHTK3.0[3]を用いた．

文法は12桁までの数字を受理するものとした．

評価用セットは男性10名，女性10名の計992文を用いた．

分析条件は，表3である．

7 認識実験結果

7.1 孤立単語音声認識の性能の評価

話者20名，50単語での孤立単語認識を行った．音声波形の平均の長さは，1678.484 [ms]であり，これを実時間とし，認識時間は実時間比として算出した．第一パスのビーム幅と認識時間の関係を図3(混合数2)に，認識時間と認識率の関係を図7に示す．また，第一パスのビーム幅とAccuracyの関係を図5に示す．

同様に，Linuxにおいて計算した結果を，図4，図8，図6に示す．

図3より，PDAとLinuxでは認識時間に相当の差があることがわかるが，PDAでは音声から特徴量への変換にかかる時間(WAV2MFCC)が割合として小さくなっている事がわかる．図5，図6を見ると，混合数2の場合には，PDAとLinuxの間にそれほど大きな認識率の差は見られない．しかし，混合数を大きくした場合，PDA環境ではLinux環境ほどの認識率の向上が見られない．近似計算及び固

[3]Hidden Markov Model Toolkit
http://htk.eng.cam.ac.uk/

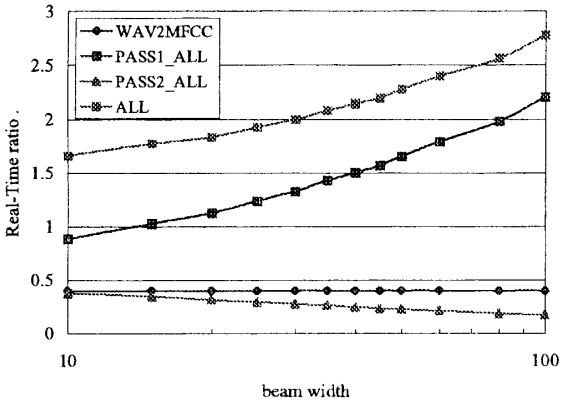

図 3: 認識時間 (混合数 2, PDA)

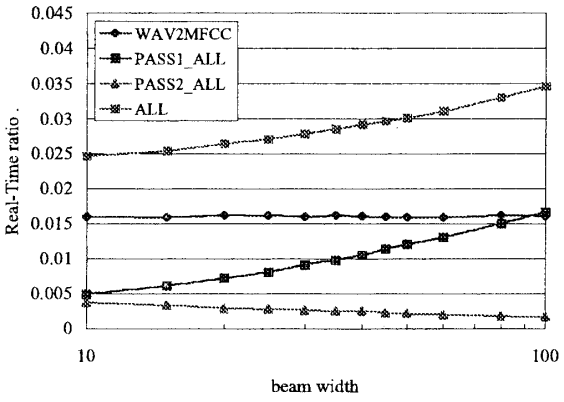

図 4: 認識時間 (混合数 2, Linux)

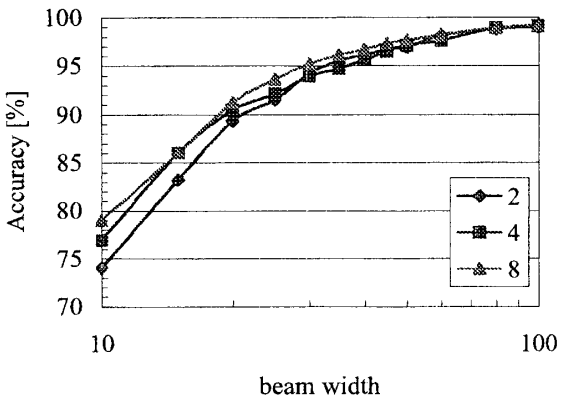

図 5: 単語正解精度 (PDA)

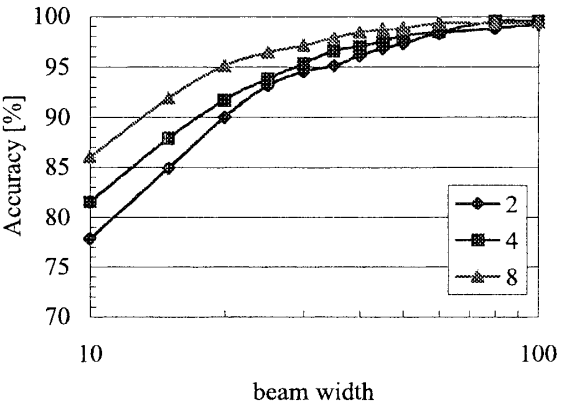

図 6: 単語正解精度 (Linux)

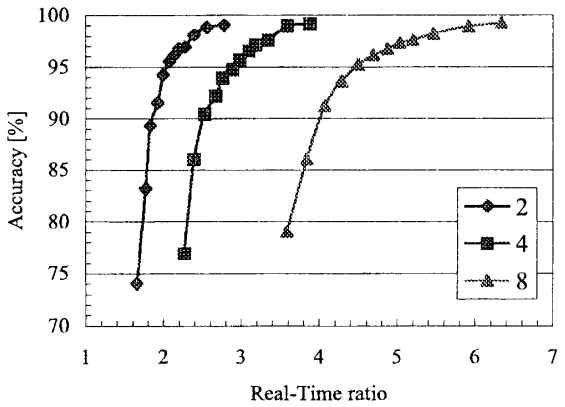

図 7: 実時間比と認識率 (PDA)

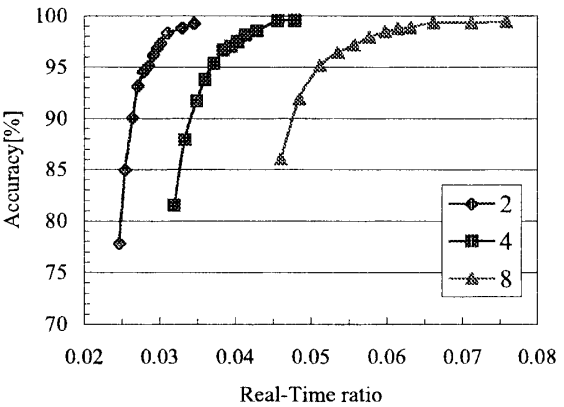

図 8: 実時間比と認識率 (Linux)

定小数点演算による誤差のために音響モデルの高精度化の効果が薄れてしまったと考えられる．認識時間に関しては，PDA では，実時間の約 1.9 倍で認識率 90%，実時間の約 2.0 倍で認識率 95%となっているのに対して，Linux では，実時間の約 0.026 倍で認識率 90%，実時間の約 0.028 倍で認識率 95%となっており，PDA での認識時間は Linux 環境での認識時間には到底かなわない．PDA での実時間動作はまだ難しい状態である．

7.2 連続数字音声認識の性能の評価

連続数字音声認識を PDA で実行した場合の認識率は表 5 のようになり，Linux 上で実行した場合の認識率は表 6 となった． 表 5 と表 6 を比較すると，

表 5: PDA 上での認識率 (Accuracy[%])

テスト音声	音響モデル		
	IPA	接話	PDA
接話	95.19	99.09	77.61
PDA	93.69	97.80	99.31

表 6: Linux 上での認識率 (Accuracy[%])

テスト音声	音響モデル		
	IPA	接話	PDA
接話	95.83	99.32	82.42
PDA	94.21	98.19	99.56

各組み合わせにおいて，PDA で認識した際にやや認識率が落ちている事がわかる．これは，孤立単語音声認識の場合と同様である．

それぞれ，表の中で比較してみると，IPA の音響モデルを用いた場合に比べそれぞれの音声を用いて学習した場合の方が認識率は高い．これは，IPA のモデルが音素バランス文を用いて学習した物であるのに対して，それぞれの音声を用いた場合のモデルはそれぞれの環境に適応した物であるということが原因である．

PDA 音声の認識率は，接話マイクで学習した音響モデル・PDA で学習した音響モデルともに IPA の音響モデルよりも高い．しかし，接話マイク音声の認識率は，PDA の音声で学習した音響モデルの場合，IPA の音響モデルよりも低くなる．無音部でのノイズが多い PDA の音声で学習した音響モデルでは，ノイズの少ない接話マイクの音声認識には適していないといえる．

8 まとめ，今後の課題

本報告では，Julius を PDA 用に移植し，その性能を評価した．孤立単語認識，連続数字音声認識共に数%の認識率の低下となっており，一般的な PC 環境で動作させた Julius/Julian と比較しても十分な性能であると言える．しかし，認識率 90%を出すためには，実時間の約 1.9 倍ほどの時間がかかっておりまだ速度的に十分ではない．今後は，特徴量分析のみではなく探索アルゴリズムに関しても検討する必要がある．

参考文献

[1] 山端潔，磯谷亮輔，安藤真一，花沢健，石川晋也，江森正，磯健一，服部浩明，奥村明俊，渡辺隆夫．”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”．信学技報, NLC2002-18, 2002-07

[2] 石川晋也，江守正，三木清一，大西祥史，磯谷亮輔，磯健一．”コンパクトなディクテーションの開発”．音講論，pp165-166,2002-3

[3] 河原達也，李晃伸，小林哲則，武田一哉，峯松信明，伊藤克亘，山本幹雄，山田篤，宇津呂武仁，鹿野清宏．”日本語ディクテーション基本ソフトウェア (98 年度版)”．日本音響学会誌 56 巻 4 号, pp.255-259, 2000

[4] 中村哲，武田一哉，黒岩真吾，山田武志，北岡教英，山本一公，西浦敬信，藤本雅清，水町光徳．”SLP 雑音下音声認識評価のための WG:評価データ収集について”．情報処理学会研究報告，SLP45-9(本研究報告内)

[5] 河口信夫，松原茂樹，武田一哉，板倉文忠，稲垣康善．”実走行車内音声対話データベース”．情報処理学会研究報告，SLP39-24，pp.141-146，2001

[6] 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹夫．”IT TEXT 音声認識システム”．pp168, 2001