社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006−SLP−61 (1)
2006／5／11

音声認識技術の実用化に向けた自動車内実環境での評価実験

大淵 康成 †, 畑岡 信夫 †

† (株) 日立製作所 中央研究所

〒185-8601 東京都国分寺市東恋ヶ窪 1-280

E-mail: obuchi@rd.hitachi.co,jp, hataoka@crl.hitachi.co.jp

あらまし 音声認識技術の実用化を促進するためには、ユーザーが実感する性能を正しく再現しながら研究を進めることが不可欠である。そのため、対象となるアプリケーションをカーナビゲーションシステムに絞り込み、評価データベースの収集を行なった。さらに、実環境における性能を向上させるために有望な研究の方向性を探るため、音声認識システムを細かいモジュールに分け、それぞれに対する改良方式の評価実験を行なった。実験の結果、音声区間検出、特徴補償、話者適応、複数マイクによる並列認識などのアプローチが有望であるとの結果が得られた。

キーワード 音声認識、実用化、データベース、カーナビゲーション

Evaluation Experiments in Automotive Environments for Development of Practical Speech Recognition Technologies

Yasunari OBUCHI † Nobuo HATAOKA †

† Central Research Laboratory, Hitachi Ltd.

1-280 Higashi-koigakubo, Kokubunji, Tokyo 185-8601, Japan

E-mail: obuchi@rd.hitachi.co,jp, hataoka@crl.hitachi.co.jp

Abstract To realize practical applications of speech recognition technologies, it is necessary to develop the system while keeping an eye on its real performance experienced by the user. We focus on the car navigation system as the target application, and develop an evaluation database to reproduce the real performance of speech recognition. Next, we investigate the potential effectiveness of various research directions to improve the performance of speech recognition in real environments. We divide the system into several small modules, and carry out evaluation experiments of improving methods for each module. Experimental results show that among various methods, speech endpointing, feature compensation, speaker adaptation, parallel decoding are promising approaches.

Key words Speech recognition, practical application, database, car navigation

1 はじめに

音声認識研究の長い歴史を通じて、その性能は初期のものに比べて格段に向上している。にもかかわらず、実用化という観点ではいまだ低い水準に留まっている。研究レベルでは極めて高い認識率が得られ、より困難なタスクへと研究対象がシフトしつつあるのに対し、ユーザーの反応としては「音声認識はまだまだ性能が低いので使いにくい」というものが多い。このような違いが生じる原因として、ユーザーインタフェースの使い勝手の悪さに起因するものと、音声認識エンジンの実環境性能の不十分さに起因するものとが考えられる。前者は、特に語彙や文法が規定された音声認識システムに対し、「何と言ってよいのかわからない」といったユーザーの声が寄せられることなどに象徴されている。一方後者は、受理される単語や文については理解しているとしても、実際にそれを発話する際の様子が必ずしもシステムの設計者の想定通りになっておらず、大量の誤認識が発生するような現象に現れている。本研究では、後

者の問題を対象とし、第一にそのような問題の所在と比率を明らかにし、第二に既存の様々な改良アルゴリズムがそのような問題の解決にどのように寄与しうるかを調べることを目的とする。

実際のユーザーが感じる不満が、研究開発段階で十分に洗い出されないという問題が発生する原因として、評価データベースが実環境を十分に模擬していないということが挙げられる。特に、これまで研究分野で用いられてきた主要なデータベースの特徴として、

1. 発話が自然に行なわれていない
2. いわゆる“失敗発話”のデータが捨てられている
3. 実環境でアレイマイクロフォンを用いている例が少ない

などが、実環境との乖離を生んでいると考えられる。1は、発話内容リストを被験者に与えて読み上げているケースが典型的であり、ユーザーが自発的に話す場合に比べて、発話の明瞭性やタイミングなどが大きく異なっている。これに対し、「日本語話し言葉コーパス」[1]の開発など、近年ではより自然な発話のデータを収集することを意図したプロジェクトも起こりつつあるが、個別のアプリケーションに対応するまでには至っていない。2に関しても、そもそも読み上げによる収録では実際よりも失敗が起こりにくいということに加え、データベース整備の段階で“失敗発話”のデータは捨ててしまうことが多い。そのため、例えば発話のタイミングが極端にずれているデータや、酷い間違い、言いよどみなどのデータ、さらには発話と誤認されがちな雑音データなどが、十分に含まれていないことになってしまっている。3はこれらとはやや異なる観点であるが、近年のマイクロフォンアレイ技術の発展を実用化に活用しようとした場合、アプリケーションが実際に使われる状況にアレイを設置して音声を収録することが必要であるが、そのようなデータはあまり多くない。本研究では、これらの点をふまえた上で、実際のアプリケーションとしてはカーナビゲーションを選び、実環境での評価用データベースを収録する。次に、音声認識システムを構成する様々なモジュールの役割を分析し、それぞれにおける改良手法の影響を評価実験によって明らかにする。その際、標準的に用いられているいくつかの代表的なアルゴリズムを評価するだけでなく、時には各モジュールの役割に対する教師信号に相当するものを与えることにより、そのモジュールの寄与の上限値を見積もる。これにより、今後の研究の方向に対する指針が得られると期待される。

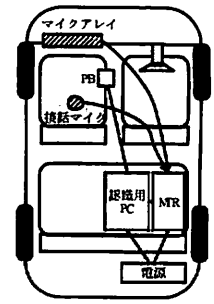

図1　データ収集装置の概要

表1　データベース収録の詳細

収録日	2005/10/24-11/01 (うち6日間)
収録環境	日産ウイングロード 助手席 / 都内一般道
話者	20代前半 男性11名+女性7名
タスク	目的地(POI)入力 (東京都内152地名)
マイク配置	直線状のマイクアレイ (7マイク)
データ量	全3620発話 (発話時間約4時間)

2　実車走行環境での評価用データベース

カーナビゲーションシステムでの目的地入力を対象とした音声収録を行なった。使用した機器の構成を図1に示す。助手席に座った被験者の前(ダッシュボード上)に、7つのマイクを直線状に並べたアレイを設置した。マイク間の距離は、端から10cm/5cm/5cm/5cm/5cm/10cmとした。この他に、参照用として接話マイクによる録音も同時に行なった。合計8つのマイクの出力は、マルチトラックレコーダに送られて録音される。それと同時に、接話マイクによる音声をノートPCに転送し、実際の音声認識システムを動作させた。音声

認識システムは、被験者に持たせたプッシュボタンにより起動させる。具体的には、被験者がプッシュボタンを押すと、「目的地を入力して下さい」というメッセージの後にビープ音が鳴り、その直後から音声認識システムが音声の取り込みを開始する。発声後には、音声認識システムによる認識結果が再生される。誤認識が生じた場合には、一回だけ再発声を行なうように指示した。

表1にデータベースの詳細を示す。被験者は20代前半の男性11名および女性7名で、それぞれ約1時間半のセッションのあいだの全音声を録音した。発話の自然性を保つため、被験者には、152個の目的地 (Point Of Interest: POI) のラベルを貼った地図帳を与え、そこから自分の意思でPOIを選択して入力するように指示した。この際、ラベルを読むことがないよう、いったん地図帳を閉じてから発声するようにした。また、次のPOIを選択する際には必ず異なるページを開く、同じカテゴリー (駅名、ホテル名、ホール名など) をなるべく連続して選ばないなどの条件を付加することにより、POI選択の際に被験者自身の主体的な判断が必要になるようにした。そのため、発話の頻度には被験者ごとのばらつきが生じ、最も多い被験者で326発話、最も少ない被験者で134発話のデータが得られ、合計で3620発話となった。152個のPOI別で見ると、最も多く発話されたもので48回、最も少ないものは10回ほど選択された。この中には、被験者の意図が自明と思われる読み間違い (「東京都葛西臨海水族園」を「東京都葛西臨海水族館」、「武蔵新田 (にった) 駅」を「武蔵新田 (しんでん) 駅」など) も数多く含まれる。その他、被験者の勘違いによる明らかな語彙外発声、プッシュボタンの誤操作による無発声 (雑音のみ) などが合計28個あった。

これらのデータに対し、信号対雑音非 (SNR) の推定を行なった結果を表2に示す。7つのマイクに対し、運転席側から窓側に向かって順に1から7の番号を振ってある。音声区間の平均パワーおよびその前後にある非音声区間の平均パワーの単純比較によって求めたSNRは、-2.7dBから-5.0dBの範囲に分布したが、必ずしも中央のマイクが最も高いということにはなっていない。更に、自動車内での雑音スペクトルは、音声認識に直接の影響の少ない低周波数帯域にピークを持っているため、

表2 SNRの推定

マイクID	全帯域 (dB)	バンドパス (dB)
1	-5.0	9.3
2	-2.8	12.1
3	-3.4	8.6
4	-3.0	9.2
5	-2.7	11.7
6	-3.8	8.5
7	-2.9	10.5

より正確な推定のため、400Hzから5500Hzの通過帯域を持つバンドパスフィルターを通した後のデータでも推定を行なった。その結果は、表の右側に示してある通り、8.5dBから12.1dBの範囲となった。

3 ベースライン評価実験

収録したデータベースの概略を知るため、152個のPOIを対象とした孤立単語認識のタスクで、ベースラインとなる認識実験を行なった。実験結果の信頼性を確保するため、二つの認識プログラム (デコーダ) をできる限り併用して行なうこととした。使用したデコーダの概略を表2に示す。片方のデコーダは、我々が本実験のために新規に開発した、孤立単語認識専用のエンジンである。以下これを「オリジナルデコーダ」と呼ぶ。オリジナルデコーダ用には、独自に収集した音素バランス文約16時間分の学習データを用いてトライフォンHMMを学習し、高速化のためにサブベクトル量子化を行なった。特徴量としては0次から12次までのMFCCおよびその一階・二階時間微分を用いた。一方、もう片方のデコーダとしては大語彙連続音声認識システムJulius (実際にはJulian rev.3.4.2) を使った。音響モデルとしては、システムに同梱される標準モデルのうち、Phonetic Tied-Mixture (PTM) モデルを用いた。特徴量は1次から12次のMFCCと対数パワー、それにこれらの一階時間微分を加えたものである。オリジナルデコーダはリジェクション機能を持たず、すべての入力に対して語彙内の単語のいずれかを出力するのに対し、Julianでは無音モデルだけのパスが選ばれることが可能で、そのような場合は入力が棄却されたものと見なす。ベースライン実験では、ビープ音を基準とした固定幅での音声区間検出を行ない、発話全

表 3　ベースライン認識実験

デコーダ	オリジナル（実験専用版）	Julius (Julian rev.3.4.2)
音響モデル	サブベクトル量子化 Triphone-HMM	PTM-HMM（標準モデル）
学習データ	独自収集音素バランス文 （120 名、合計 16 時間）	
特徴量	MFCC（0 次〜12 次）+Δ+ΔΔ	MFCC（1 次〜12 次）, logP +Δ
リジェクション	なし	簡易版（無音だけのパス）
区間検出	時間で固定（ビープ音後 0.9 秒〜4.9 秒）	
特徴補償	CMN（発話単位）	
サンプリング	16kHz	
タスク	東京都内 152POI 孤立単語認識	

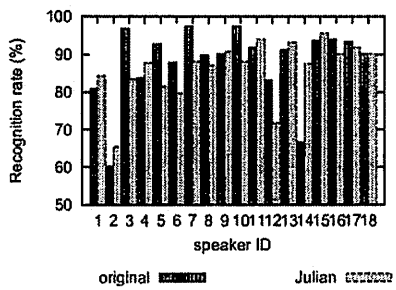

図 2　ベースライン認識実験の結果

全体を用いた Cepstral Mean Subtraction (CMN) を適用した。

図 2 にベースライン認識実験の結果を示す。横軸は個別話者の ID であるが、話者別に見ると、認識率は最高で 97.4%、最低で 60.3%であった（オリジナルデコーダの場合）。2 つのデコーダで得られた結果を比較すると、個人別には多少の上下が見られるが、異なる学習データを用いていることを考えると不思議には当たらない。全体の平均では、オリジナルデコーダで 87.3%、Julian で 86.0%と両者でほぼ同等の性能が得られており、ベースラインとして信頼できる値が得られたものと考えられる。

4　モジュールおよび改良手法の評価

次に、音声認識システムを構成するモジュールごとの評価実験を行なった。ここでは、音声認識システムを以下のモジュールに分けて考える。第一に、入力信号から音声認識の対象となる部分をおおまかに取り出す音声・非音声識別モジュール。次に、得られた信号をもとに、音声として照合すべき区間の始終端を正確に判定する音声区間検出モジュール。次に、得られた音声信号を強調したり、雑音除去を行なったりする音声強調モジュール。さらに、そこで得られた音声信号を照合にふさわしい特徴量に変換し、必要があればさらに補償・正規化などを行なう特徴抽出・補償モジュール。さらに、特徴量とモデルの照合を行なう音響モデル照合モジュール（話者適応などのモデル変換も音響モデル照合モジュールに含むものとする）。最後に、照合結果をもとに最終出力を決定する後処理モジュールである。個々のモジュールに対する評価実験の結果をまとめたものを図 3 に示す。

4.1　音声・非音声識別モジュール

音声・非音声識別モジュールの寄与を調べるため、非音声（語彙外発声を含む）入力が与えられた場合、強制的に棄却させる (oracle rejection) こととした場合の認識率の変化を見た。結果は、オリジナルデコーダで 88.1%、Julian で 86.4%となり、それぞれ 0.8 ポイントおよび 0.4 ポイントの改善であった。プッシュボタンによる起動方式を取っているため、そもそも非音声の入力が為される割合がその程度であり、改善の程度はさほど大きくない。

4.2　音声区間検出モジュール

詳細な音声区間情報を天下りにあたえた場合 (oracle endpointing) の認識率を調べた。区間情報は、接話マイクで同時録音したデータを利用し、発話内容との強制アラインメントを取り、単語の前後に付加した無音モデルとマッチした部分を非音声区間と見なして除去した。その結果、認識率はオリジナルデコーダで 91.1%、Julian で 90.2%とな

り、それぞれ3.8ポイントおよび4.2ポイントという大きな改善効果が得られた。これらの値は誤り削減率に直すと29.9%および30.0%に相当する。

4.3 音声強調モジュール

音声強調による効果を見積もるため、まず最初に、アレイマイクのデータと接話マイクのデータの重み付き和を取ることにより、SNRが改善された状況をシミュレートした。その結果、SNR14dBで90.2%（オリジナル）・89.3%（Julian）、SNR26dBで91.1%（両者とも）となり、大きな改善が得られることが分かった。次に、マイクロフォンアレイを用いた代表的な手法として、遅延和アレイおよび独立成分分析(ICA)[2]を適用して認識実験を行なった。遅延和アレイでは、中央のマイクの正面50cmの位置に話者の口があると想定し、マイクの配置から推定した遅延を用いた。ICAでは、中央のマイクの両側にある3番と5番のマイクを用い、時間領域と周波数領域の両方での適用を試した。ICAではマイク数と同じ数の出力信号が得られるが、その中から認識対象を選択する部分のエラーを取り除くため、すべてに対して認識を行ない最適なものを選択した。その結果、遅延和アレイでは88.2%（オリジナル）・86.7%（Julian）となり若干の改善が見られたが、ICAではむしろ僅かに認識率が低下するという結果となった。また、7つのマイクすべてを使ったICAも行なったが、認識率が更に低下するという結果であった。

4.4 特徴抽出・補償モジュール

本実験ではすべてMFCCを特徴量として用いているが、それに対する代表的な特徴補償方式として、MVN (Mean and Variance Normalization)、HEQ (Histogram Equalization)、DCN (Delta-Cepstrum Normalization)[3]の三つを適用する実験を行なった。それぞれを単独で用いた場合、認識率は86.2%(MVN), 84.0%(HEQ), 86.9%(DCN)となり、改善効果は得られなかったが、これを音声区間既知(oracle endpointing)の実験と組み合わせた場合、MVNで95.6%という高い認識率が得られた。(HEQでは91.0%, DCNでは90.1%) HEQやDCNは推定するパラメータが多く、本データベースのような短い発話では有効に働かなかったが、MVNにおいては、音声区間を正しく知ることにより分散の推定精度が上がり、高い性能が得られることがわかった。

4.5 音響モデル照合モジュール

音響モデルの品質が音声認識精度に強い影響を及ぼすことは言うまでもないが、寄与するパラメータの多様さと、学習に要する処理の大きさから、あらゆるパターンの音響モデルを包括的に評価することは困難である。ここではそれに代わり、話者適応によるモデルの精度改善が音声認識精度にどれくらい寄与するかを調査することにした。話者適応のアルゴリズムとしては、最も良く使われるものの一つとしてMaximum Likelihood Linear Regression (MLLR)があるが、その実装の一形態であるFeature-space MLLR (FMLLR)[4]を用いた。その結果、1単語を使った話者適応では認識率が劣化したが、2単語使用の場合でほぼ変化なし、5単語使用で顕著な向上が見られ、10単語使用でほぼ性能が飽和する様子が見られた。なお、教師信号として音声認識結果を用いるいわゆる教師なし話者適応でも、教師ありの場合とほぼ同等の認識率が得られたため、図には教師なしの場合の結果を示してある。

4.6 後処理モジュール

上記の各モジュールを組み合わせて単一の音声認識結果を得るという方式の他に、複数の候補を残しておき、何らかの後処理で最終結果を決めるというアプローチも可能である。代表的なものとして、7つのマイクを使って並列認識を行ない、得られた7つの候補単語から最終出力を選択するという方式がある。この方式で得られる性能の上限を知るため、正解単語既知の条件で、発話ごとに最適なマイクを選ぶことができる場合の認識率を調べた。その結果はオリジナルデコーダで91.6%、Julianで92.4%となり、このアプローチの可能性を示した。次に、代表的なチャネル選択手法として、デコーダから得られる尤度によるものと、単純な多数決によるものとを比較した。その結果、尤度による選択ではベースラインよりも低い性能しか得られないのに対し、多数決では88.1%（オリジナル），87.1%（Julian）という若干高い性能が得られた。更に、多数決方式による改善度合を高めるため、特徴補償のところで述べたMVN, HEQ, DCNによる認識結果も含めた7チャネル4方式・合計28仮説による多数決を行なったところ、オリジナルデコーダの認識率が90.2%まで向上した。最後に、上記28仮説にJulianによる7仮説まで含

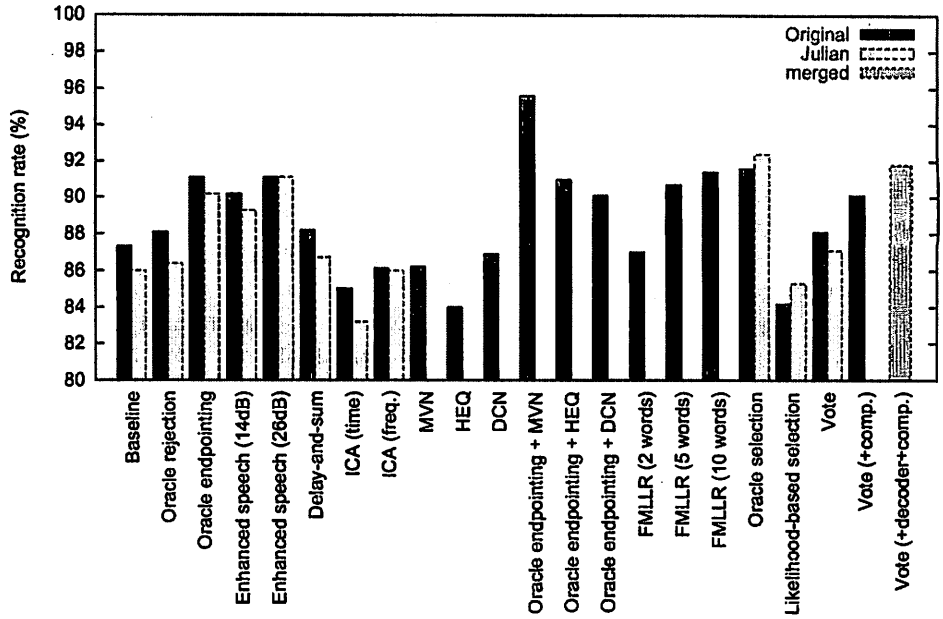

図 3　評価実験の結果のまとめ

めた結果、認識率は 91.9%となった。

5　まとめ

音声認識技術の実用化を促進するため、カーナビゲーションをビークルとして、実環境における音声認識性能の評価のためのデータベース収集および音声認識方式の評価実験を行なった。データベース収集においては、マイクロフォンアレイの活用を念頭に置きつつ、発話の自然性やデータの多様性を損なわないような方式を採用した。得られたデータを用いて、音声認識システムを構成する各モジュールの評価実験を行なった結果、音声区間検出や MVN による特徴補償、話者適応、複数マイクによる並列認識などで大きな性能向上が見られることを確認した。今後は効果の高そうな方式の詳細評価を進めるとともに、同様のアプローチにより他の実アプリケーションについての検討も進めていきたい。

謝辞

有益なご助言をいただいた、東京工業大学古井貞熙教授、早稲田大学小林哲則教授に感謝いたします。なお本研究は、(独) 新エネルギー・産業技術総合開発機構の「音声認識技術実用化に向けた先導研究」（早稲田大学受託）の再委託を受けて実施した。

参考文献

1) K. Mackawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation," *Proc. SSPR2003*, Tokyo, Japan, 2003

2) N. Murata, et al., "An Approach to Blind Source Separation Based on Temporal Structure of Speech," *BSIS Technical Report*, 00-6, 2000

3) Y. Obuchi, et al., "Normalization of Time-Derivative Parameters for Robust Speech Recognition in Small Devices," *IEICE Trans. Information and Systems*, Vol.E87-D, No.4, pp.1004-1011, 2004

4) M. J. F. Gales, "Maximum Likelihood Linear Transformation for HMM-based Speech Recognition," Technical Report, Cambridge University Engineering Department, 1997