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

2003－ＳＬＰ－49　（６）
2003／12／18

# 大規模車内音声データベースにおける認識性能変動要因調査

藤村　浩司†　　伊藤　克亘††　　武田　一哉††　　板倉　文忠†††

† 名古屋大学大学院 情報科学研究科 メディア科学専攻
†† 名古屋大学大学院 情報科学研究科
††† 名古屋大学大学院 工学研究科
〒464-8603 愛知県名古屋市千種区不老町
E-mail: †fujimura@itakura.nuee.nagoya-u.ac.jp, ††{itou,takeda}@is.nagoya-u.ac.jp,
†††itakura@nuee.nagoya-u.ac.jp

**あらまし**　異なる対話モードの対話発声の特徴を分析した。分析には異なる３つの対話モードでの，800 人の発話者のスピーチコーパスが用いられている。３つの対話モードとはヒューマンナビゲータとの対話，WOZ システム，ASR システムとの対話である。文の複雑さや声の大きさのような特徴が対話モード間で意義深い特徴として見られた。線形回帰分析により認識率とこれらの特徴との相関の重要性を明確にした。
**キーワード**　音声認識　車内対話データベース　SNR　評価

# Research on the factor that influence recognition accuracy of large scale In-Car Spoken Dialogue Database

Hiroshi FUJIMURA†, Katsunobu ITOU††, Kazuya TAKEDA††, and Fumitada ITAKURA†††

† Graduate School of Engineering, Nagoya University
†† Graduate School of Information Science, Nagoya University
††† Graduate School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8603 japan
E-mail: †fujimura@itakura.nuee.nagoya-u.ac.jp, ††{itou,takeda}@is.nagoya-u.ac.jp,
†††itakura@nuee.nagoya-u.ac.jp

**Abstract**　The dependency of conversational utterances on the mode of dialogue is analyzed. A speech corpus of 800 speakers collected under three different modes, i.e., talking to a human operator, an WOZ system and an ASR system, is used for analysis. Some characteristics such as sentence complexity and loudness of the voice are found to be significantly different among the dialogue modes. Linear regression analysis results also clarify the relative importance of those characteristics on speech recognition accuracy.
**Key words**　speech recognition In-Car Spoken Dialogue Database SNR evaluation

## 1. 序　　論

話し言葉は運転中もっとも便利なインターフェースとして期待されているが，現在の技術では運転者が複雑な操作中に気をそらすことなく，自然かつ確実にその認識システムを活用することはできない。車内音声認識のもっとも重要な問題のひとつに運転状況の変化があげられる。すなわち，運転者のシステムに対する姿勢が頻繁にかわることである。現在の音声認識技術は語彙や文の複雑さ，話速や声の大きさの変化に敏感であるため，運転状況によってめまぐるしく変化する発声に対応できない。車内対話音声システムの構築において運転状況の変化による発声の変化をモデル化し，かつ予測することは欠くことのできない要素である。

この研究の目的は運転者が機械を意識することなく，その運転状況の変化に対応し，自然に車内での情報検索を行える

ような音声インタフェースを構築することが目的である。そのような自由度の高いインタフェースを構築するための第一歩として，多人数の話者が公道を運転しながら3つの異なる対話システムを使って，実際に情報検索を行ったデータを，様々な特徴量から分析し，認識性能を悪化させる要因を導き出す。

名古屋大学CIAIRでは車内対話やさまざまな行動を信号化した大規模なコーパス[1]を収集している。本紙ではCIAIR車内対話コーパスを用いて運転者の発声が対話モード（人間との対話，WOZシステムとの対話，機械の情報検索システムとの対話）によってどのように変わるかを明確にする。

## 2. CIAIR車内対話コーパス

CIAIR車内コーパス[1]は800人を超えるドライバのデータが記録されたマルチメディア信号1.4TB以上から構成されている。特にデータ収集を目的として作られた，データ収集車（DCV）はマルチチャンネルオーディオデータ，マルチチャンネルビデオデータそして運転データを同時に収録することができる。音声信号は接話マイクと4つのマイクロフォンアレーを含む車内に配置された16個のマイクロフォンで収集した。運転席のバイザー部分に取り付けられたマイクを以後遠隔マイクとよぶ。このマイクは話者の口唇から40cm以内にある。車内の画像は運転者の顔と車外の前面の様子がビデオデータとして収録されている。運転データは5つの特徴量（エンジン・ブレーキペダル，エンジン回転数，スピード，ハンドル角度）が収録されている。

名古屋大学のキャンパス周辺の一時間程度の運転中に運転者は3つの違ったシステム（HNシステム，WOZシステム，ASRシステム）との対話を行う。HNシステムは，助手席に座っている人間（ナビゲータ）がドライバの情報検索に関する質問に応えるシステムである。3つのシステムの中でドライバがもっとも機械を意識せずに対話できるシステムである。WOZシステムは，DCV内にインストールされたタッチパネルを用いた検索システムを用いて人間のナビゲータが検索を行い、その応答を音声合成で出力することによって対話を行うシステムである。性質としてはHNシステムとASRシステムの中間にあたるものである。ASRシステムは，1500語からなる語彙のバイグラムを用いた音声認識システムJulius[2]によってドライバの発話を認識し、その認識結果に基づいて対話処理を行うシステムである。質問の受け応えは全て機械が行う。よって3つのシステム中もっともドライバが機械を意識する必要があるシステムである。図1から図3にそれぞれのシステムの対話例を示す。図中のDはドライバの発話を表し，Nはナビゲータの発話を表す。図1より，HNシステムではドライバ音声，ナビゲータ音声共にフィラーやいい間違いが存在する。またその対話内容は豊富なバリエーションを持ちドライバは機械を意識することなく自然な会話で情報検索を行うことができる。図2より，WOZシステムでは聞き取るのが人間なのでドライバは比較的自由な発話で検索を行うことができるが，あらかじめデータに入っていないような質問には応えることができない。図3より，ASRシステムでは聞き取り，応答と共に機械の判断によるので例のように会話が成り立たないことがある。

D ： この近くにコンビニってありますか？
N ： お店の指定はございますか？
D ： できたらサークルケーがいいんですけど
N ： はい えー この近くですとサークルケーがございます
D ： えーと この辺りで評判のいい麺類のお店はありますか？
N ： は はい
N ： えー ラーメン うどん そば パスタなどがございますが
D ： そしたらうどんでお願いします

図1 HNシステムの対話例

D ： お腹が空いたので何か食べたい
N ： どういったジャンルがよろしいでしょうか？
D ： そうですねえ
D ： オープンカフェみたいなとこがいいです
N ： はい
N ： 近くにアトリエシェルブランがあります
D ： ほかにはないですか？
N ： 申し訳ございません
N ： データがありません

図2 WOZシステムの対話例

N ： レストラン検索システムです
D ： はい
N ： 検索条件を入力してください
D ： 和食が食べたいんですけど
N ： 和食 洋食 中華のどれがいいですか？
D ： 和食がいいです
N ： ジャンルは和食でよろしいでしょうか？
D ： はい

図3 ASRシステムの対話例

## 3. コーパス分析

### 3.1 コーパスサイズ

この研究で使われるコーパスの特性を各対話セッションごとに表1に示す。分析には435人分の発声の32000から40000文を使用する。文は200ms以上の無音区間で切り出された音声である。無音区間と音声区間の境界は手動で与えられた。文の平均の長さは継続時間においても形態素の数に関してもHNセッションが他の2つのセッションよりも長い。これら2つの平均はASRセッションの約2倍である。

表 1　コーパス分析

(a) size

| | | HN | | WOZ | | ASR | |
|---|---|---|---|---|---|---|---|
| Total(sec) | | 101430 | | 73116 | | 80978 | |
| Driver | Navigator | 0.40 | 0.60 | 0.39 | 0.61 | 0.21 | 0.79 |
| 文数 | | 40560 | | 32883 | | 40149 | |
| Driver | Navigator | 0.44 | 0.56 | 0.42 | 0.58 | 0.40 | 0.60 |
| 文毎の継続時間 (sec) | | 2.26 | | 2.05 | | 1.07 | |
| 形態素数 | | 353875 | | 195513 | | 262354 | |
| Driver | Navigator | 0.34 | 0.66 | 0.44 | 0.56 | 0.19 | 0.81 |
| 文毎の形態素 | | 8.72 | | 5.95 | | 6.53 | |
| Driver | Navigator | 6.84 | 10.2 | 6.27 | 5.71 | 3.17 | 8.74 |
| 話速 (msec/mora) | | 144.3 | | 143.5 | | 149.1 | |

(b) 複雑さ (perplexity)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| bigram | trigram | 18.1 | 7.7 | 14.1 | 7.1 | 9.1 | 6.6 |
| 語彙サイズ | | 5001 | | 3216 | | 1839 | |

(c) 音響特性

| | 接話 | 遠隔 | 接話 | 遠隔 | 接話 | 遠隔 |
|---|---|---|---|---|---|---|
| SNR[dB] | 23.0 | 10.6 | 24.0 | 11.3 | 26.0 | 12.9 |

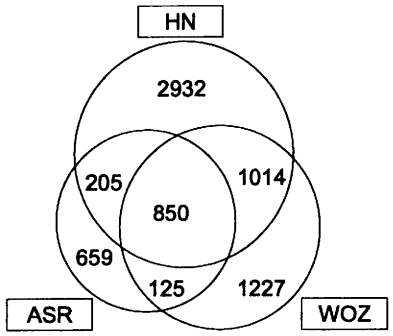



図 4　異なるセッション間の語彙の重なり具合

対話は質問と応答の対からなっているのでドライバとナビゲータ間の文の発声率はセッション間にあまり違いはない。大体ドライバが 40-45%の文を発声している。

### 3.2 話　速

HTK を使い，音素ごとにアライメントをした結果から，ドライバ発話の話速を算出した。話速は 1 モーラの平均継続時間を基準とした。その結果，話者ごとの平均継続時間に関しては HN セッションと WOZ セッションではあまり大きな差は見られなかったが，ASR セッションは他のセッションと比べて，5msec/mora 以上大きいことがわかった。これは ASR セッションが他のセッションよりゆっくり話していることを意味する。

### 3.3 発声の複雑さ

表 4 より HN セッションの語彙サイズは ASR セッションの約 2 倍であり，WOZ セッションはその中間ぐらいである。違うセッション間の語彙の重なり具合を図 1 に示す。HN セッションが他のセッションより語彙が多くなるのは “出張”，“誕生日パーティー”，“空腹” などのタスク外のものを含むためである。

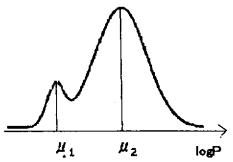



図 5　音声と雑音の log パワー分布

表 2　言語モデルのトレーニングデータ

| | HN | WOZ | ASR |
|---|---|---|---|
| 人数 | 535 | 586 | 575 |
| 文数 | 22240 | 19044 | 21289 |
| のべ形態素数 | 149213 | 117250 | 66612 |
| 異なり形態素数 | 5532 | 3694 | 2083 |
| bigrams | 35095 | 22277 | 9850 |
| trigrams | 67972 | 44322 | 18403 |

パープレキシティはほぼ語彙サイズに比例する，それは各々HN，WOZ，ASR セッションの順で 18.1，14.1，9.1 である。

### 3.4 音響状態

各セッションにおいて話者の声の大きさの変化を調べるために，また SNR と認識率との関係を調べるために，音声と雑音の比である SNR(signal-to-noise ratio) を算出する。以下にその算出方法を示す。

各文章全体の log フレームパワーの分布を，雑音の分布と音声の分布という二つの混合ガウス分布 (図 5) であると仮定する。EM アルゴリズムを用いてこのガウス分布を推定し，平均値の小さい分布の平均値 ($\mu_1$) を雑音 ($\log N$)，平均値の大きい分布の平均値 ($\mu_2$) を音声 ($\log S$) として後述の式の SNR を計算する。

$$\mathrm{SNR} = \log \frac{S}{N} \tag{1}$$

発話時間の短い文章では，認識に用いるフレーム幅，フレームシフト (表 3) ではフレーム数が少なすぎるためガウス分布で近似することができない。よって今回はフレーム幅 4msec, フレームシフト 2msec でフレームパワーを算出した。また EM アルゴリズムの初期値として，各ガウス分布の平均値には，log フレームパワーの最小値から累積頻度 10%のところと，累積頻度 90%のところを用いた。この算出方法は，SNR が負のときにはうまく計算することができない [3]。ASR セッションの SNR は HN，WOZ セッションにくらべ 2dB ほど高い。3 セッションとも同様な運転状態であるので混入する雑音はほぼ同じである。よって ASR セッションは運転者自身が大きな声を出しているといえる。

## 4. 認識実験

### 4.1 実験準備

セッション間での認識率の性能比較を行うために，言語モ

表3 分析条件

| | |
|---|---|
| サンプリング周波数 | 16kHz |
| 分析窓 | ハミング窓 |
| フレーム長 | 25ms |
| フレームシフト | 10ms |
| 特徴パラメータ | MFCC(12 次) |
| | ⊿ MFCC(12 次) |
| | ⊿パワー (1 次) |
| 使用帯域 | 250-8000Hz |

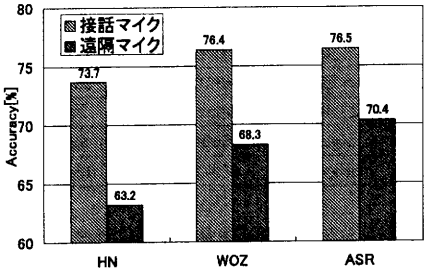



図6 3つのセッションの認識性能

デルを各セッションごとに構築した。言語モデル構築のためのテキストデータの詳細を表2に示す。テキストデータにはいい間違いやフィラーが混在する。形態素解析には ChaSen を用い，その後手動で修正した。トレーニングデータには認識実験に使うデータが全て含まれる。認識にはカットオフなしでバックオフスムージングした前向きバイグラム，後ろ向きトライグラムを構築し使用した。

音響モデルは接話マイクの音声で学習したものと，遠隔マイクで学習したものの2種類を作成した。認識時にはそれぞれのマイクに対応した音響モデルを使用した。トレーニングセットは10音節以下の文 (ほとんどが はい/いいえ などの応答や，フィラー，いい間違いで構成されている。) を除き，全てのセッションを用いて構築した。トレーニングデータの量は HN セッションの文 11746 文 (11.4 時間分)，WOZ セッションの文 8550 文 (7.84 時間)，ASR セッションの文 4878 文 (3.10 時間) の計 25174 文 (22.4 時間) である。テストセットの話者はすべてトレーニングデータに含まれているが，話者によってその量は違う。帯域は 250-8000Hz を使用している。音響モデルは3状態トライフォン HMM を用い，状態数が 2000 状態，ガウス分布の混合数が 32 混合のものを作成した。HMMの学習には HTK を用いている。ショートポーズのモデルは初期状態から最終状態までの飛び越し遷移を許す，スキップモデルとして学習している。デコーダには Julius [2] を用いた。また表3に音声の分析条件を示す。

### 4.2 セッション間の認識率の比較

音声の認識性能を図6に示す。接話マイクのみで比較するとその性能は HN セッションが最も悪い，そして WOZ セッションと ASR セッションとの間にそれほど差は見られない。この結果から，音質に差がないときに認識性能を劣化させる要因は主にパープレキシティが占めているといえる。

一方，遠隔マイクのみで比較すると，WOZ と ASR セッションの認識性能が 2%以上違う。SNR15dB 以下の騒音環境下においては 2dB の改善は認識性能において致命的な差となる。

### 4.3 SNR 基準の認識性能の比較

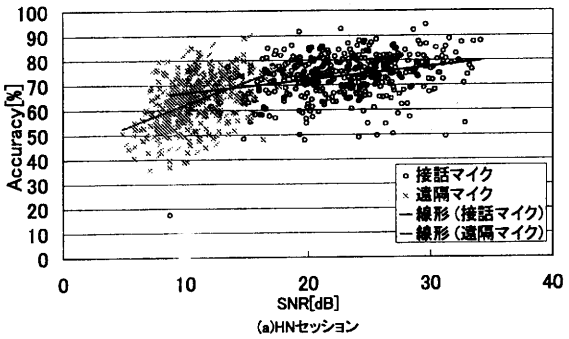
話者ごとの平均認識率と平均 SNR の関係を図7に示す。また二乗誤差を最小にするような回帰直線も重ねて示してある。接話マイクではどのセッションでも SNR と認識率との相関はあまり高くなく，約 0.54%/dB である。しかし，遠隔マイクでは 1.88%/dB であり，SNR と認識率との間に相関が見られる。この結果から，ある程度 SNR が得られるところではあまり SNR と認識率に相関はなく，SNR があまり得られないところでは (本研究では SNR15dB 以下程度) 認識率との相関が高くなる。

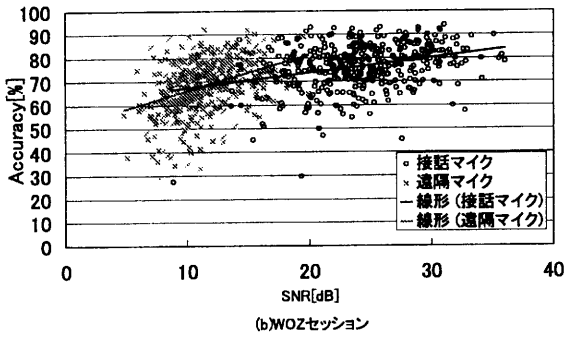
### 4.4 パープレキシティ基準の認識性能の比較

話者ごとの平均認識率と平均パープレキシティの関係を図8に示す。また二乗誤差を最小にするような回帰直線も重ねて示してある。どのセッションにおいても，接話マイクと遠隔マイクで傾向はほぼ同じであり，パープレキシティが大きくなるほど認識性能は悪くなる。

### 4.5 話速基準の認識性能の比較

話者ごとの平均認識率と話速の関係を図9に示す。また二乗誤差を最小にするような回帰直線も重ねて示してある。相関係数は 0.10 である。独話の講演音声では話速が速いほうが認識率の性能が悪い [4] と言われているが，本研究ではその傾向は見られなかった。図9は ASR セッションのものであるが，実験の結果どのセッションでもその傾向は見られない。単純に話すスピードだけではなく，発話中に挿入されるポーズの頻度や継続時間，話速の変化が関係している可能性もある。また人によっては語尾を延ばすような形で1文内でもモーラ継続時間が大きく変化する。そこで図10に話者ごとの平均認識率と話速分散をプロットした。結果相関係数は 0.25 となり，話速そのものよりも認識率との相関は上昇した。相関係数は最小二乗誤差基準で計算されているため，外れ値に対して頑健ではない。図10には話速分散が高いところに数点の外れ値が存在する。この点を詳細に分析することで話速分散との関連性をさらに調査する必要がある。また今回は話速基準としてモーラごとの平均継続時間を使用したが，分散値を見るとわかるようにあまり安定した値が得られない。よってもっと大きい単位 (単語のモーラ毎平均継続時間等) で平均化した後，認識率との相関を出す必要がある。



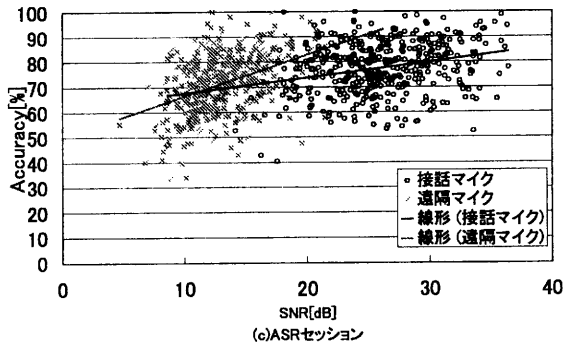
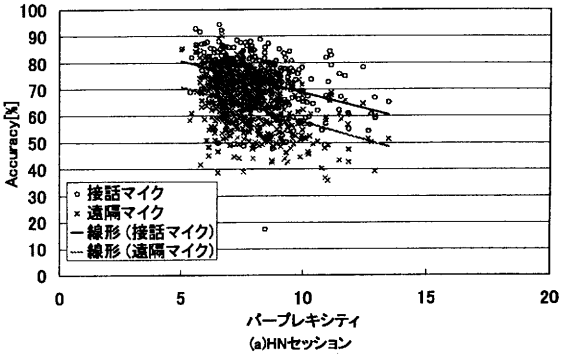
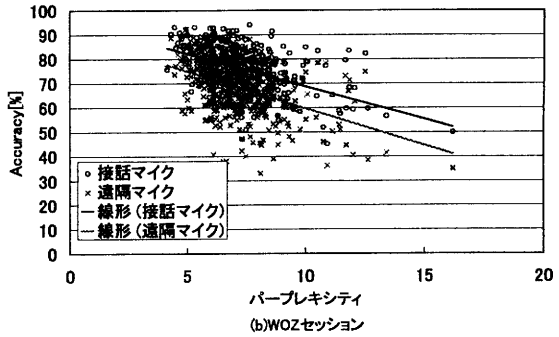




図 7　話者ごとの認識率と平均 SNR を示したもの。435 人のデータがプロットされている。(a)HN セッション (b)WOZ セッション (c)ASR セッションの結果がそれぞれ示されている。○は接話マイクの結果を示しており，× は遠隔マイクの結果を示している。



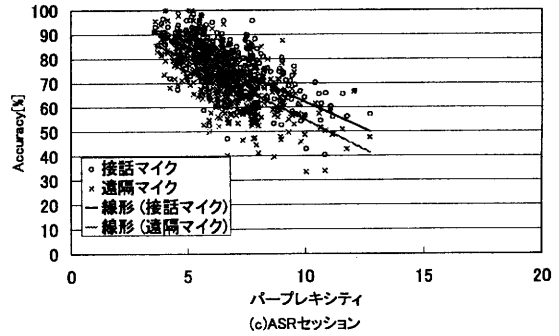




図 8　話者ごとの認識率と平均パープレキシティ(trigram) を示したもの。435 人のデータがプロットされている。(a)HN セッション (b)WOZ セッション (c)ASR セッションの結果がそれぞれ示されている。○は接話マイクの結果を示しており，× は遠隔マイクの結果を示している。

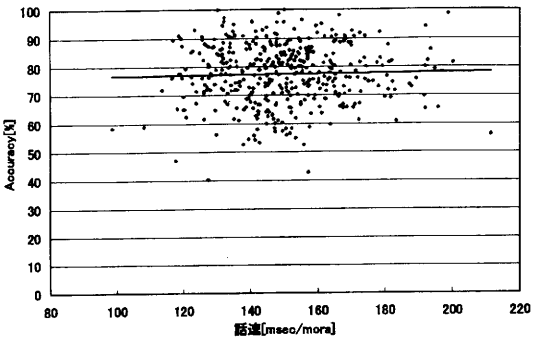



図 9　話者ごとの認識率と話速 [msec/mora]

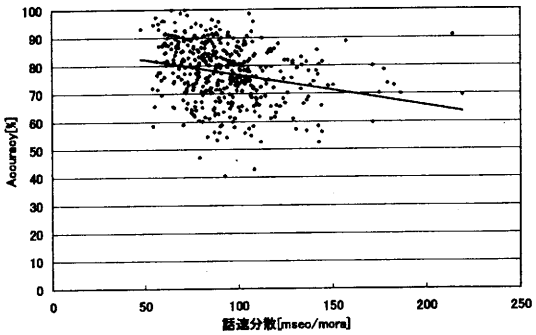



図 10　話者ごとの認識率と話速の分散 [msec/mora]

## 5. 回帰分析

図 11 では単語正解率とパープレキシティ，SNR，1 文の平均継続時間との相関係数が示されている。もっとも高い相関係数を示すのは単語正解率とパープレキシティとの相関で特に ASR セッションでは 0.69 である。単語正解率と SNR との関係は接話マイクより遠隔マイクのほうが相関が高い。どのセッションでも相関係数にして 0.1 程度の差が見られる。また，1 文の平均継続時間はあまり認識率に相関がないことがわかる。

## 6. 結　　論

本研究では，対話発声がどのようにして対話モードの違いで変化するかを調査した。異なるモードでの大規模なドライバ発話は語彙サイズ，文のパープレキシティ，SNR や話速によって特徴付けられることがわかった。もっとも明確で重要な結果は ASR システムに対して人間は大きな声で話し，さらにあまり複雑な文を話さないということである。声の大きさは他のセッションと比べて平均 2dB 以上大きい。パープレキシティも他のセッションに比べて小さくなっている。また話速に関しても ASR システムに対しては，他のセッションに比べ 1 モーラに対する平均継続時間が長いことから，ゆっ

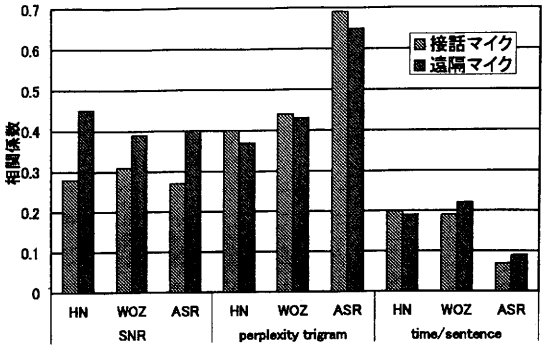



図 11　SNR，パープレキシティ，1 文当たりの平均継続時間と認識率との相関係数

くり話していることがわかる。以上の結果から，人間に対しての対話と，機械に対しての対話で種々の特徴量が変化していることがわかる，すなわち人間は機械を意識すると話し方に変化があらわれるのである。認識率との回帰分析もまた重要な結果を示している。それは認識率を改善するためには SNR と文の複雑さの改善が求められることである。話速に関しては今回の研究では認識率との相関は得られなかった。

この研究の議論は主に話者ごとの平均値基準でなされている。それは話者による特徴の違いがかなり大きいことを示している。これからはさらに細かい基準での調査が必要である。また一般的に認識率に影響を与えるという話速に関してさらなる調査を要する。また，今回の結果は音響モデルに関しても言語モデルに関してもクローズなデータで算出されたものである。今後は，少ないデータで多くのデータを近似できるようなテストセットを構築し，オープンなデータでその性能を評価する。その上で話者適応などの既存の手法を行い，認識性能を悪化させる要因についてさらに詳細に調査する。

## 文　　献

[1] N.Kawaguchi,et.al.,"Construction of speech corpus in moving car environment", Proceedings of International Conference on Spoken Language Processing,pp.1281-1284,2000.

[2] A.Lee,et.al.,"Continuous Speech Recognition Consortium — an Open Repository for CSR Tools and Models —", Proc International Conference on Language Resources and Evaluation (LREC2002), pp.1438-1441, 2002.

[3] M.Kondo,K.Takeda,F.Itakura,"Predicting the Degradation of Speech Recognition Performance from Sub-band Dynamic Ranges", IPSJ Journal,pp.2242-2248,2002.

[4] T.Shinozaki,S.Furui,"Analysis on Individual Differences In Automatic Transcription Of Spontaneous Presentations", Proc. ICASSP2002, Orlando, U.S.A., vol.1, pp.729-732,2002.