日本語話し言葉コーパスを用いた講演音声認識

篠崎隆宏[†] 古井貞熙[†]

本論文では日本語話し言葉プロジェクトに関連して進めている，講演音声認識に関する種々の分析結果について述べる．日本語話し言葉コーパスに基づき学習したモデルと，従来のモデルの性能の比較を行い，話し言葉コーパスから作成した言語モデルや音響モデルが講演音声の認識において有用であることを確認した．話し言葉を対象とした音声認識における，話者間での単語正解精度分布の構造の解析を行い，個人差の要因として発話速度，未知語率および言い直し頻度の影響が比較的大きいことを明らかにした．また，教師なし話者適応化は単語正解精度の向上に効果的に働くが，適応化を行った後も発話速度の影響は減少しないことを示した．

Presentation Transcription Using a Japanese Spontaneous Speech Corpus

Takahiro Shinozaki[†] and Sadaoki Furui[†]

This paper reports various investigations on recognizing spontaneous presentation speech in connection with the “Spontaneous Speech” national project. Experimental results show that acoustic and linguistic models based on the Japanese spontaneous speech corpus are far more effective than the conventional models for recognizing the presentation speech. Individual differences in the speech recognition performances are analyzed. A restricted set of the speaker attributes comprising the speaking rate, the out of vocabulary rate and the repair rate is found to be most significant to yield individual differences in the word accuracy. It is shown that unsupervised MLLR speaker adaptation works well for improving the word accuracy but does not compensate for the effect of the speaking rate.

1. はじめに

現在の音声認識技術は，書き言葉の読み上げ音声などは高い精度で認識ができるが，話し言葉を対象とすると単語正解精度が著しく低下する問題がある．話し言葉を認識するためには，話し言葉に基づいた音声認識システムが必要である．しかし，日本語においては，大語彙連続音声認識を対象とした話し言葉の大規模なコーパスはこれまで存在せず，話し言葉を対象にした研究は，タスクを限定した対話音声などに限られていた[1),2)]．

このような背景から，話し言葉のためのモデルと技術の構築を目指し，文部科学省科学技術振興調整費を財源とする，開放的融合研究推進制度によるプロジェクト「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」が，1999 年から 5 年間の計画で行われている[3)]．プロジェクトの主たる研究グループは，国立国語研究所，通信総合研究所，東京工業大学である．主に以下の 3 点を目標としている．1) 話し言葉コーパス “CSJ” の構築．2) 音響情報，言語情報，パラ言語情報を用いた，音声認識理解や音声要約を目的とした音声のモデル化．3) 話し言葉音声の要約システムのプロトタイプの構築．

話し言葉における発話の自発性に関しては様々な程度が考えられるが，プロジェクトの主な対象は講演音声などのモノローグである．CSJ は完成すればモノローグの話し言葉を集めたコーパスとして，世界で最も大規模なものの 1 つとなる．日本語の話し言葉を対象としたコーパスとしてはこのほかに，IBM による講義音声を対象としたコーパスの構築があげられる[4)]．IBM のコーパスは放送大学による放送音声を対象としており，非公開である．

現在の音声認識の認識性能は，認識に使用するモデルの学習セットと認識対象の性質の違いに影響を受けやすい．話し言葉の音声認識の認識性能が低い理由の 1 つとして，これまで話し言葉を基にしたモデルがなく，書き言葉に基づいたモデルを用いていたことがあ

† 東京工業大学大学院情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

げられる．このため，まず話し言葉に基づくモデルを使用した場合にどの程度まで単語正解精度が改善するか調べることが重要である．文献 4) では放送大学の講義音声を対象として，講義音声から学習したモデルと従来モデルとの比較を行っている．本論文では CSJ を試験的に用い，学会などの講演音声を対象として CSJ に基づくモデルと従来のモデルとの性能の比較を行うとともに，公開を前提としたコーパスである CSJ を利用した話し言葉音声認識のベースラインを明らかにする．

話し言葉では話者の自由度が大きいことから，書き言葉の読み上げ音声に比較して発話スタイルの個人差が大きくなっていると考えられる．そこで音響的特徴の個人差に対応するため，音響モデルの教師なし話者適応化を行い単語正解精度の改善を試みた．さらに話し言葉における発話スタイルの種々の音響的・言語的な個人差が認識性能にどう影響を与えるか，多数の話者を対象とした認識実験を基に要因の分析を行い，個人差の構造を明らかにする．

2. 日本語話し言葉コーパス（CSJ: Corpus of Spontaneous Japanese）

CSJ には約 7 M 語 700 時間の話し言葉が収録される予定である．主な収録対象は学会講演や模擬講演，ニュース解説などのモノローグである．模擬講演はコーパスの収録のために行われた，一般の話者による 10 分程度の日常的な話題のスピーチである．録音された音声に対し，人手により書き起こしが作成される．書き起こしには時間情報やフィラーや言い直しなど，発声の属性を示す付加情報が含まれる．CSJ において「フィラー」とされるのは言葉を探すときなどに場つなぎ的に発声される「あ」「あのー」「えーとー」などや，「ほおー」などの感情表出系感動詞である．「言い直し」とされる発声は，言い直された単語の断片および助詞，助動詞である．書き起こしには，かな漢字混じりの「基本形」書き起こしと，発音に忠実なカタカナ表記の「発音形」書き起こしがある．

3. 音声認識システム

3.1 実験条件

音声は 16 kHz で標本化，16 ビットで量子化を行った．音響パラメータは MFCC12 次元，Δ ケプストラム 12 次元，対数パワーの 1 次差分の計 25 次元で，切り出した疑似的文単位ごとに，平均ケプストラムによる正規化（CMS）を行った．音響モデルの学習と話者適応には HTK2.2[5] を使用した．

形態素解析には，NTT で開発された形態素解析ツール JTAG を使用した．言語モデルは，CMU SLM Tool Kit v2.05[6] を使用して作成した．デコーダは Julius3.1[7] を使用した．

認識実験の際，言語重み，挿入ペナルティは音響モデルと言語モデルの組合せごとに最適化し，テストセット中では共通の値を用いた．

単語正解精度の計算は形態素を単位とし，フィラーを含めて行った．その際，フィラーのうち「あー」と「あ」などは同一のものとした．これは，これらの区別があまり重要ではないと考えられること，書き起こしの際，どちらの表記とするかはっきりしない場合も多いことによる．「えー」と「えーっと」などは別のものとして区別した．

音声認識には，エネルギーを用いて切り出した単位を基に，単語の途中などで切れないように人手により修正した音声単位を用いた．認識単位の切れ目はおよそ 500 ms 以上の無音に対応する．

3.2 言語モデル

以下の言語モデルを作成した．どのモデルも 2 gram と逆向き 3 gram からなり，語彙サイズは 30 k である．形態素とその発音のペアをモデル化の単位として用いた．**SpnL** は CSJ を用いて学習したモデル，**WebL** は**SpnL** との比較用に用いた CSJ を用いないモデルである．語彙はモデルごとに，学習データ中の出現頻度が上位のものとした．3 gram のカットオフは 1 とした．

SpnL： CSJ においてすでに書き起こしが得られている部分を試験的に使用して作成した言語モデルである．モデルの学習にはかな漢字まじり文として書き起こされた「基本形」講演書き起こしテキストを用いた．学習セットとして使用したのは，610 講演であり，内訳は多い順に模擬講演が 336，音響学会が 139，言語処理学会が 63 講演，その他 72 講演となっている．形態素数にすると約 1.5 M のサイズがある．

フィラーに関しては他の単語と言語的な性質が異なることが予想されることから透過単語としてモデル化する方法や 1 つのクラスにして取り扱う方法などを検討したが，良い結果が得られなかったことから，単に通常の単語としてモデル化した．CSJ の書き起こし作業ではフィラーに関して比較的ゆるい規則を用いている．このためフィラーの種類は様々であるが，コーパス中で複数回出現するものはおよそ 150 種であり，さらにその中で主要な 20 種ほどがフィラーの出現延べ数の 95%以

上を占める．言語モデルには学習セット中で複数回出現するフィラーはすべて含まれている．言い直しは Ngram で有効にモデル化することが難しいことから，学習テキストから取り除きモデル化は行わなかった．

WebL： 話し言葉コーパスを使用せずに講演音声の認識を目的とした言語モデルを作る方法として，World Wide Web 上で公開されている講演書き起こしテキストを利用することが考えられる．**WebL** は Web から収集した講演書き起こしから学習したモデルである．収集した講演のテーマ数は 43，総形態素数は約 2 M である．話題は社会問題や回顧録など一般的なものである．このようにして集めたテキストは，読みやすさの観点から書き起こす際にフィラーや言い直しなどは取り除かれ，編集されている．そこでフィラーに関しては，統計的特徴に基づき文頭および句点の前後に確率的に補うことにより，補正を行った．加えたフィラーは 22 種類である[3]．

3.3 音響モデル

以下の 2 つの音響モデルを用意した．どちらも状態共有 triphone モデルで状態数は 2000，混合数は 16 である．使用した音素は 43 種類である．**SpnA** は話し言葉に基づくモデル，**RdA** は書き言葉の読み上げ音声に基づくモデルである．

SpnA： **SpnL** の学習に使用した CSJ 学習セットのうち，女性話者を除いた男性話者による 338 講演，約 59 時間を用いて学習を行った，性別依存・不特定話者モデルである．講演の内訳は音響学会が 122，模擬講演が 116，言語処理学会が 52 講演，その他 48 講演である．音声はヘッドセットマイクロホンにより録音されている．

RdA： IPA による「日本語ディクテーション基本ソフトウェア 99 年度版」[7] に収録されている音響モデルである．読み上げ音声から学習されたモデルとして，日本の音声認識の研究において一般的に用いられているモデルである．約 40 時間の複数の男性話者による読み上げ音声から作られている．音声はヘッドセットマイクロホンにより録音されている．

4. CSJ に基づくモデルと従来モデルの比較

本章では CSJ を使用し学習したモデルと CSJ を利用しないで学習したモデルの，講演音声認識における性能の比較を行う．同時に，CSJ を使用したシステムにおける認識性能のベースラインを明らかにする．

表 1 テストセット（10 名）の概要
Table 1 Recognition test set of presentations (10 males).

Presentation ID	Short name	Conference name	Length [min]
A01M0035	A22	日本音響学会	28
A01M0007	A23	日本音響学会	30
A01M0074	A97	日本音響学会	12
A02M0117	J01	国語学会	57
–	K05	国語研究所	42
A03M0100	N07	言語処理学会	15
A05M0031	P25	音声学会	27
A06M0134	S05	社会言語科学会	23
–	Y01	融合研究会	14
–	Y05	融合研究会	15

4.1 認識対象

話し言葉コーパス中の講演のうち，男性による 10 講演をテストセットとした．概要を**表 1** に示す．すでに公開されている CSJ コーパスに含まれている講演については，その ID を示してある．すべて音響モデル，言語モデルの学習セットに登場しない話者による講演であり，話者オープンである．

4.2 パープレキシティおよび未知語率の比較

言語モデル**WebL** および**SpnL** のテストセットパープレキシティを**図 1** に，未知語率を**図 2** に示す．パープレキシティの計算には未知語の予測は含めていない．**WebL** の 3 gram パープレキシティの平均は 450 と大きな値となった．また 2 gram と 3 gram を比較すると，2 gram の方がかえって低い値となった．これらのことは**WebL** の学習に使われたテキストは書き言葉として編集されていることから，表 1 に示したテストセット講演との言語的な特性の差異が大きいためと考えられる．CSJ から作成された**SpnL** のパープレキシティの平均は 200 で，**WebL** の半分以下である．また 2 gram と比較して 3 gram の方がパープレキシティが低く，3 gram の効果が出ていることが分かる．

未知語率に関しても**WebL** は 10 講演の平均が 4.8% と高い値となった．未知語となった単語には話し言葉特有の表現も含まれるものの，多くは専門用語であった．**WebL** の学習セットは Web から集めた一般的な話題の講演であり，学会講演などからなるテストセット講演に対して語彙が不足するためである．**SpnL** の未知語率の平均は 1.5%と**WebL** の 3 分の 1 以下となった．

話し言葉からなる学会などの講演に対して，**SpnL** が**WebL** と比較して大きく優れていることが分かる．

4.3 単語正解精度の比較

言語モデルに**WebL** または**SpnL**，音響モデルに**RdA** または**SpnA** を用いた場合の 4 通りの単語正

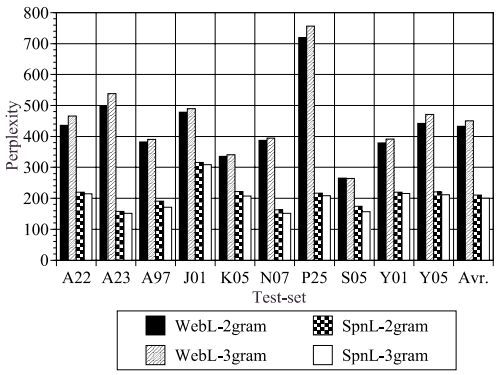

図 1 各言語モデルのパープレキシティ
Fig. 1 Test-set perplexity for each language model.

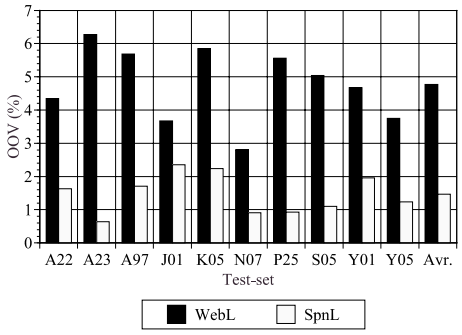

図 2 各言語モデルの未知語率
Fig. 2 OOV rate for each language model.

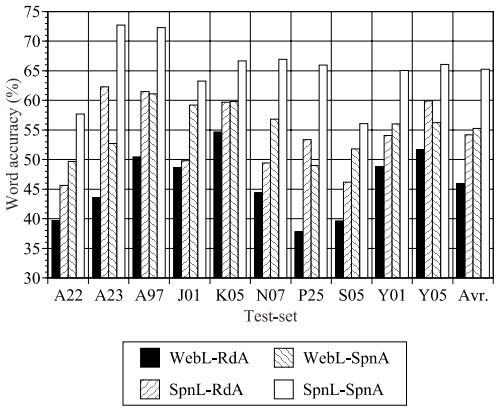

図 3 各音響モデル・言語モデルを用いたときの単語正解精度
Fig. 3 Word accuracy for each combination of models.

解精度を**図 3** に示す.

音響モデルが**SpnA** のとき，言語モデル**WebL** と**SpnL** を比較すると単語正解精度の平均はそれぞれ 55.2%，65.3%で，**SpnL** を用いた場合の方が単語正解精度が平均で 10%高い値となった．この結果は 4.2 節で示した，パープレキシティや未知語率で見た言語モデルの性能比較の結果と対応している．

言語モデルに**SpnL** を用いたとき，音響モデル**RdA** と**SpnA** を比較すると単語正解精度の平均はそれぞれ 54.2%，65.3%で，**SpnA** を用いた場合の方が単語正解精度が平均で 11%高い値となった．この違いは，それぞれのモデルの学習に使用された読み上げ音声と話し言葉音声では triphone の分布や音響的特徴の点で違いがあり，話し言葉の認識のためには話し言葉音声から学習した音響モデルが必要になるためと考えられる．

言語モデル，音響モデルどちらに関しても，CSJ から作成したモデルの方が講演音声認識に対して明らかに優れていることが分かる．言語モデルに**WebL**，音響モデルに**RdA** を用いた場合の単語誤り率は 54.1%，**SpnL** と**SpnA** を用いた場合の単語誤り率は 34.7%であり，単語誤り率の相対的な削減率は 35.7%である．

5. 教師なし話者適応

本章では個人差による音響特徴量の変化に対応する目的で，音響モデルの教師なし話者適応化を行う．テストセットは表 1 に示す 10 講演で，前章で使用したものと同じである．

5.1 適応化方法

適応化の手順を**図 4** に示す．適応化は，1）不特定話者モデル SI-HMM からスタートし，認識結果を基に MLLR を行い話者適応化モデル SA-HMM を作成する，2）得られた SA-HMM を基に，さらに話者適応化を繰り返す，ことにより行った．MLLR による適応化の学習データを与える単位に関しては，各適応化段階においてそれぞれの講演全体をまとめて用いる．

適応化の基にした不特定話者音響モデルは**SpnA** である．言語モデルには**SpnL** を使用した．MLLR は HMM 中の全正規分布をあらかじめ 64 の葉を持つ 2 分木の葉に対応させることで分類しておき，学習時のデータ量に応じて，適応化に使用するクラスを決定する方法を用いた．正規分布の分類は centroid-splitting により行い，平均値のみを適応化した．信頼尺度は用いなかった．この実験では話者適応化の結果，音響モデルは講演ごとに異なったものとなるが，各適応化段階において，言語重みや挿入ペナルティは 10 講演共通とした．

5.2 単語正解精度

図 5 に話者適応化の効果を示す．図で “SI” は不特定話者音響モデル “SpnA” を用いたベースラインを示す．MLLR による適応化は 3 回まで行った．図で “SA1”，“SA2” および “SA3” はそれぞれ適応化を 1，2 および 3 回行った場合である．

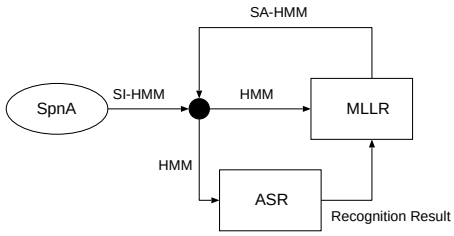

図 4 教師なし話者適応化の手順
Fig. 4 Procedure of MLLR unsupervised adaptaion.

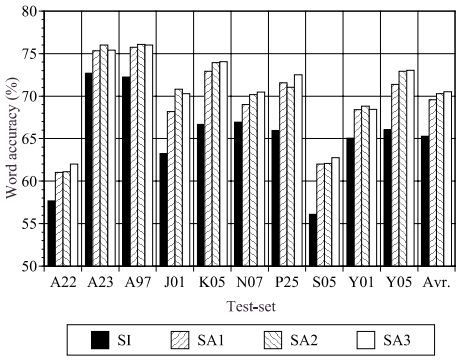

図 5 教師なし話者適応化と単語正解精度
Fig. 5 Effectiveness of unsupervised adaptation.

はじめの MLLR 適応化の効果が大きく，単語正解精度は絶対値で 2〜6%向上した．2 回目の適応化では 1 回目と比較して平均 0.7%程度改善した．3 回の MLLR 適応化で単語誤り率は不特定話者モデルの場合に比べ相対的に 15%削減された．その結果，平均単語正解精度は 70.5%となった．なお，言語重みと挿入ペナルティを話者ごとに最適化した場合の単語正解精度は，MLLR 適応化を 3 回行ったモデルを使用した場合で 71.0%であった．

6. 話し言葉における話者間単語正解精度分布の構造

本章では話し言葉における発話スタイルの様々な個人差と認識性能の関わりを明らかにする．

6.1 認識タスク

話者間の個人差の解析のため，男性 51 名の話者による講演音声をテストセットとした．テストセット中の話者はすべて異なる話者であり，また学習セットには含まれていない．分析には各講演の初めの 10 分間を用いた．**表 2** にテストセットの概要を示す．

6.2 話者属性

音声認識に関わる話者（講演）の属性として以下の 7 種類を分析の対象とした．

Acc 講演の単語正解精度（%値）．

表 2 テストセット（51 名）の概要
Table 2 Recognition test set of presentations (51 males).

Presentation	No.presentations
人工知能学会	32
日本音響学会	12
その他	7

AL 講演の平均フレーム音響ゆう度．

SR 講演の平均発話速度（1 秒あたりの音素数）．

PP テストセットパープレキシティ．

OR 未知語率（%値）．

FR フィラー頻度（正解文の単語数に対するフィラーの%値）．

RR 言い直し頻度（正解文の単語数に対する言い直しの%値）．

実験に用いた言語モデルは CSJ を学習に用いた**SpnL** である．音響モデルは不特定話者モデル**SpnA**，および**SpnA** を基に MLLR による教師なし話者適応化を行ったモデルを使用した．発話速度と音響ゆう度は正解音素列の強制アライメントの結果から，無音にマッチした部分を除いて求めた．正解音素列にはフィラー部分も含まれている．アライメントされたフレームのうち平均して約 8.8%はフィラー部分である．パープレキシティは言語モデルに 3 gram を用いて求めた．未知語の予測は含めていない．単語がフィラーや言い直しかどうかの判定には，CSJ の書き起こしに含まれているタグ情報を用いた．未知語率，パープレキシティの計算では，正解文として言い直しを除いたものを用いた．発話速度は**SpnA** を用いて求めた．以下では，不特定話者音響モデルを SI，教師なし話者適応化音響モデルを SA と表記する．

6.3 平均および標準偏差

テストセット 51 名の話者における各属性の平均と標準偏差を**表 3** に示す．単語正解精度の平均は音響モデルに SI を用いた場合で 64.1%，SA を用いた場合で 68.6%であった．標準偏差は SI を用いた場合で 7.4%，SA を用いた場合で 7.5%であり，単語正解精度が話者により大きくばらつくことが分かる．

なお，これらの属性値を認識単位ごとに求めた場合，話者内においてもかなり変動する．特にパープレキシティや未知語率，言い直し頻度の話者内での標準偏差は表 3 に示す話者間の標準偏差の数倍となる．各講演の初めから第 n 文までの属性の平均値と n の関係を調べたところ，10 分間のデータを用いることで，すべての属性で話者間のおおまかな順位が決まる程度には値が収束する様子が観察できた．ただし，パープレキシティ，未知語率，言い直し頻度の収束はやや不十分で

表 3 各属性の平均と標準偏差
Table 3 Mean and standard deviation for each attribute.

	Acc(SI)	Acc(SA)	AL(SI)	AL(SA)	SR	PP	OR	FR	RR
Mean	64.1	68.6	−55.5	−53.1	15.0	227	2.09	8.60	1.54
Standard deviation	7.4	7.5	2.3	2.2	1.3	63	1.18	3.67	0.73

表 4 相関係数行列；下三角行列は相関係数，上三角行列は有意確率を示す．5%水準において有意となる相関係数は，太字で表記した
Table 4 Correlation coefficient matrix; the lower triangular matrix shows the correlation coefficients and the upper triangular matrix shows the *p*-values, that is, the significance levels. Bold face indicates a significant value with the significant level of 5%.

	Acc(SI)	Acc(SA)	AL(SI)	AL(SA)	SR	PP	OR	FR	RR
Acc(SI)		–	4.7%	–	0.2%	0.4%	0.0%	0.6%	3.4%
Acc(SA)	–		–	2.2%	0.1%	1.9%	0.0%	0.5%	2.5%
AL(SI)	**0.28**	–		–	0.0%	55.8%	12.0%	6.8%	53.9%
AL(SA)	–	**0.32**	–		0.0%	56.3%	8.3%	6.7%	33.4%
SR	**−0.42**	**−0.47**	**−0.54**	**−0.62**		92.0%	1.7%	0.0%	20.2%
PP	**−0.40**	**−0.33**	−0.08	−0.08	−0.01		0.0%	20.0%	69.4%
OR	**−0.54**	**−0.51**	−0.22	−0.25	**0.33**	**0.52**		0.3%	66.5%
FR	**0.38**	**0.38**	0.26	0.26	**−0.50**	−0.18	**−0.41**		33.8%
RR	**−0.30**	**−0.31**	−0.09	−0.14	0.18	0.06	−0.06	0.14	

あった．また使用した各講演 10 分間の中では，文の出現の順番と属性の間に目立った相関は見られなかった．

6.4 相関分析

各属性間の相関を示すため，**表 4** に相関行列を示す．表において下三角行列は相関係数，上三角行列は有意確率を示す．太字で表記された相関係数は，5%水準において有意であることを示している．

6.4.1 音響ゆう度と発話速度

音響モデルに SI を用いた場合，音響ゆう度と発話速度の相関係数は −0.54 である．**図 6** に音響ゆう度と発話速度の散布図を示す．図では 2 乗誤差を最小にするようにフィットさせた直線を重ねて示してある．発話速度の速い話者で音響ゆう度が低下する傾向が観察される．他方，発話速度が非常に遅い場合であっても，音響ゆう度の低下は見られない．赤池情報量基準（AIC）[8] においても音響ゆう度を予測する 1 次と 2 次の発話速度を用いた回帰式では，1 次のモデルが選択された．講演を単位として見た場合，発話速度が増加すると音響ゆう度が減少する直線的な関係があるといえる．発話速度の増加とともにゆう度が下がる原因としては，調音結合の増加による音響的特徴の不鮮明化などが考えられる．教師なし話者適応を行った場合，音響ゆう度は全体的に上がるものの，発話速度との負の相関関係は残ることが図 6 より分かる．

6.4.2 パープレキシティと言語的属性

パープレキシティと未知語率の相関係数は 0.52 である．**図 7** にパープレキシティと未知語率の散布図を示す．未知語率の高い講演ではパープレキシティも高い傾向があることが分かる．

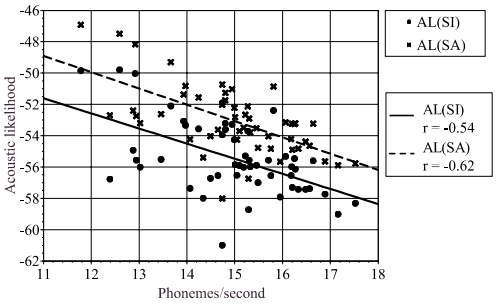

図 6 音響ゆう度と発話速度
Fig. 6 Acoustic likelihood vs. speaking rate.

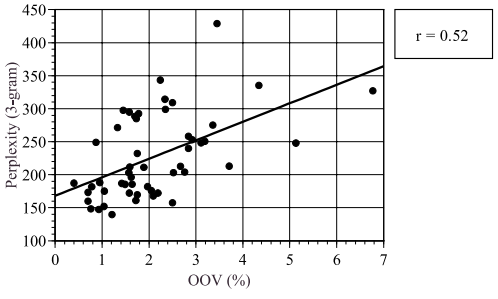

図 7 パープレキシティと未知語率
Fig. 7 OOV vs. word perplexity.

フィラー頻度とパープレキシティの相関係数は −0.18 であり，ほぼ相関はないといえる．言い直し頻度とパープレキシティの相関係数は 0.06 である．パープレキシティは言い直しを除いた正解文を用いて計算していることから，この結果は言い直しを除いた部分の言語

的な難しさはもともとの言い直し頻度とは関係ないことを示している．

6.4.3 単語正解精度と諸属性

発話速度と単語正解精度（SI）の相関係数は −0.42 である．**図 8** に発話速度と単語正解精度の散布図を示す．発話速度と単語正解精度の関係は単調であり，図 6 における発話速度と音響ゆう度の場合と同様に発話速度が非常に遅い場合でも単語正解精度が下がらない様子が観察される．AIC においても単語正解精度を予測する 1 次と 2 次の発話速度を用いた回帰式では，1 次のモデルが選択された．発話速度と単語正解精度（SA）の相関係数は −0.47 であり，教師なし適応化を行っても相関係数は減少しないことが分かる．

音響モデルに SI を用いた場合の音響ゆう度と単語正解精度との相関係数は 0.28 であり 5%水準で有意であるが，発話速度を制御した場合の単語正解精度と音響ゆう度の偏相関係数は 0.07 と，小さな値となった．音響ゆう度を制御した単語正解精度と発話速度の偏相関係数は −0.33 であり，5%水準で有意である．また，単語正解精度を制御した音響ゆう度と発話速度の偏相関係数は −0.48 であり，1%水準で有意である．このことは，音響ゆう度と単語正解精度の相関は発話速度を介したみかけの相関であることを示している．発話速度の増加は，音響ゆう度と単語正解精度をそれぞれ独立に低下させているといえる．音響モデルに SA を用いた場合の音響ゆう度と単語正解精度との相関係数は 0.32 である．偏相関係数を用いた分析に関しても SI の場合と同様の結果となった．

言い直し頻度と単語正解精度（SI）の相関係数は −0.30 である．**図 9** に言い直し頻度と単語正解精度（SI）の散布図を示す．

フィラー頻度と単語正解精度（SI）の間には相関係数 0.38 の正の相関が見られる．しかし，発話速度を制御したフィラー頻度と単語正解精度（SI）の偏相関係数は 0.22 であることから，この相関はみかけの相関であることが分かる．フィラー頻度を制御した発話速度と単語正解精度の偏相関係数は −0.29 で，5%水準で有意である．単語正解精度を制御したフィラー頻度と発話速度の偏相関係数は −0.40 であり，1%水準で有意である．

図 10 に未知語率と単語正解精度（SI）の散布図を示す．未知語率と単語正解精度の相関係数は −0.54 である．

パープレキシティと単語正解精度の間には −0.40 の相関係数があるが，これもみかけの相関である．未知語率を制御したパープレキシティと単語正解精度の偏相関係数は −0.16 である．パープレキシティを制御した未知語率と単語正解精度の偏相関係数は −0.43，単語正解精度を制御したパープレキシティと未知語率の偏相関係数は 0.39 である．なお，単語などのより短い単位で認識率との関係を調べる際には，パープレキシティそのものよりもパープレキシティの対数を用いた方がより認識率との線型性が良い．しかし講演単位での分析では，単語正解精度との相関，偏相関係数はどちらを用いてもあまり差はなかった．

以上の相関関係，みかけの相関関係を**図 11** に示す．

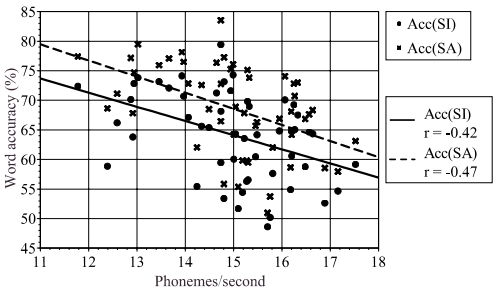

図 8 発話速度と単語正解精度
Fig. 8 Speaking rate vs. word accuracy.

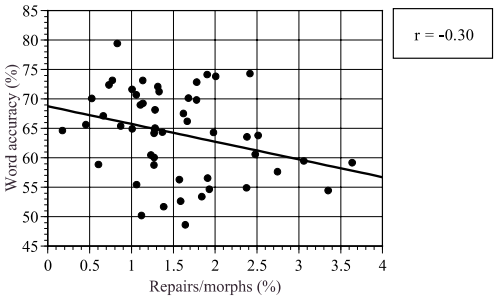

図 9 言い直し頻度と単語正解精度（SI）
Fig. 9 Repair frequency vs. word accuracy (SI).

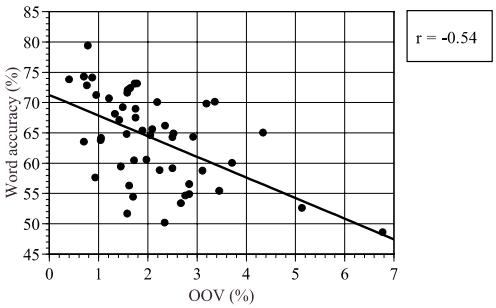

図 10 未知語率と単語正解精度
Fig. 10 OOV rate vs. word accuracy (SI).

6.5 重回帰分析

音響モデルに SI および SA を用いた場合の単語正解精度を諸属性から予測する重回帰式を式 (1) および

(2) に示す.

$$Acc_{SI} = 0.12AL_{SI} - 0.88SR - 0.020PP - 2.2OR + 0.32FR - 3.0RR + 95 \quad (1)$$

$$Acc_{SA} = 0.024AL_{SA} - 1.3SR - 0.014PP - 2.1OR + 0.32FR - 3.2RR + 99 \quad (2)$$

式 (1) において言い直し頻度の係数は −3.0 である．このことは 1%の言い直し頻度の増加が 3.0%の単語正解精度の低下に相当することを示している．同様に，未知語率の係数は −2.2 であり，1%の未知語率の増加は 2.2%の単語正解精度の低下に相当する．これらは言い直しや未知語による 1 つの認識誤りが，言語的なつながりにより 2 次的な誤りをひき起こすためと考えられる．回帰式の決定係数は式 (1) の場合で 0.48，式 (2) の場合で 0.47 である．このことは話者間の単語正解精度分布の分散の約半分が，回帰式により説明されることを示している．回帰式 (1) と (2) を比較すると，音響モデルの教師なし話者適応を行う前と後で，定数項が増加していること，発話速度の係数の大きさが減少しないことなどが分かる．

説明変量の影響力の大きさを見るため，各変量を平均と分散で正規化した後，回帰分析を行った，標準偏回帰係数，有意確率および 95%信頼区間を**表 5** に示す．係数の値が比較的小さくなった音響ゆう度，パープレキシティおよびフィラー頻度は 6.4 節において偏相関係数を用いた分析で示したように，単語正解精度との直接の相関が小さい変量である．話者適応を行った場合は発話速度，未知語率，言い直し頻度が比較的大きな回帰係数を示し，話者適応を行わない場合は未知語率と言い直し頻度が比較的大きな係数を示した．ただし，回帰式の決定係数が 1/2 程度とあまり大きくなく，本研究で対象としていない説明要因の影響が大きい可能性がある．新たな要因を取り入れることで係数の重要性が変わることも考えられる．

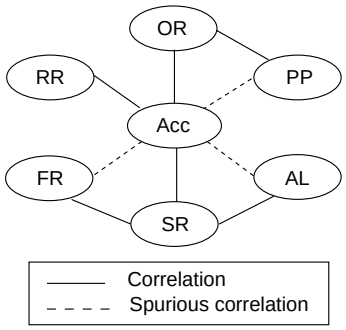

図 11 相関関係とみかけの相関関係
Fig. 11 Correlation and spurious correlation.

6.6 主要属性の分析

重回帰モデルにおいて単語正解精度を予測するうえで重要な説明変量を特定するため，変数減少法を用いた分析を行った．変数減少法ではまずすべての説明変量を含む回帰式を求める．回帰式中で一番大きな有意確率を持つ説明変量を 1 つ取り除き，残った説明変量を用いて回帰式を再計算する．変数を取り除く作業を回帰式中のすべての説明変量の有意確率が 0.10 以下になるまで繰り返す．この操作において回帰式に残った説明変量は，音響モデルに SI を用いた場合も SA を用いた場合も同じであり，発話速度，未知語率，言い直し頻度であった．これらの変数はいずれも相関分析の結果として図 11 に示した単語正解精度と直接の相関を有する属性であり，音響モデルに SI を用いた場合の発話速度を除けば，表 5 において比較的大きな係数を持つ変数に一致している．

同定された 3 属性のみを用いた重回帰式の決定係数は，音響モデルに SI および SA を用いた場合とも 0.44 であり，6 種類すべての説明変量を用いた場合とほぼ同じであった．単語正解精度の個人差の主たる要因は発話速度，未知語率および言い直し頻度であるといえる．

表 5 単語正解精度に対する標準偏回帰分析．標準偏回帰係数（Coeff），有意確率（P），95%信頼区間（95% CI）を示す
Table 5 Results of standardized regression analysis for word accuracy, showing standardized regression coefficient (Coeff), p-value and 95% confidence interval (95% CI).

	Coeff(SI)	P	95% CI		Coeff(SA)	P	95% CI
AL(SI)	0.04	76.7%	(−0.22, 0.30)	AL(SA)	0.01	96.1%	(−0.28, 0.29)
SR(SI)	−0.16	32.8%	(−0.47, 0.16)	SR(SI)	−0.23	19.0%	(−0.57, 0.12)
PP	−0.17	20.1%	(−0.44, 0.10)	PP	−0.11	40.4%	(−0.39, 0.16)
OR	−0.34	2.2%	(−0.63,−0.05)	OR	−0.32	3.2%	(−0.62,−0.03)
FR	0.16	24.9%	(−0.11, 0.43)	FR	0.16	26.0%	(−0.12, 0.44)
RR	−0.30	1.4%	(−0.53,−0.06)	RR	−0.31	1.3%	(−0.54,−0.07)

7. おわりに

本論文では，自由発話された講演音声の認識実験について報告した．

日本語話し言葉コーパス CSJ を試験的に使用して学習した言語モデルや音響モデルと従来のモデルの比較を行い，CSJ から学習したモデルが学会講演などの話し言葉に対してパープレキシティ，未知語率および単語正解精度の点できわめて優れていることを示した．CSJ に基づく言語モデルと音響モデルを用いた場合の，講演音声の単語正解精度は 10 講演の平均で 65.3%であった．

MLLR を用いた教師なし話者適応では 3 回適応化を繰り返すことで，不特定話者音響モデルを使用した場合と比較して平均で 15%単語誤り率が減少し，単語正解精度は 70.5%となった．繰返し適応化においては 1 回目の適応化の効果が大きかった．

発話スタイルの個人差の様々な要素が単語正解精度にどのように影響を与えているか，51 名の話者を用いた認識実験を基に解析を行った．種々の話者（講演）の属性のうちで，未知語率および言い直し頻度が単語正解精度に与える影響が大きいことを示した．また発話速度に関しては話者適応化を行わない場合の回帰係数は小さくなったが，相関・偏相関係数を用いた分析や変数減少法を用いた分析では単語正解精度に与える影響が比較的大きい変数であることを示した．逆に，正解文の音響ゆう度やテストセットパープレキシティと，単語正解精度との直接の相関は小さいことを示した．MLLR を用いた教師なし話者適応は単語正解精度の向上に効果的に働くものの，適応化を行っても発話速度の影響は減少しないことを示した．発話速度に対応するためには別途の対処法が必要である．種々の話者属性を考慮した重回帰式により，単語正解精度の分散の約半分が説明された．また，単語正解精度の説明において効果の大きい変量として同定された 3 属性のみを用いた重回帰式においても，ほぼ同じ説明力が得られた．ただし，本論文において対象とした話者属性以外に認識率の個人差へ大きな影響を与える要因が存在する可能性もある．文献 9) では読み上げ音声と摸擬対話音声で音韻間距離に違いがあり，音響モデルの比較において音響ゆう度よりも音韻間距離の違いの方が音節認識率への影響が大きいことが示されている．各話者に適応化させた音響モデルにおける音韻間距離をその話者の特性と考え，話者間の単語正解精度に対する説明変数として用いることも考えられる．

なお，発話スタイルと認識性能の関係としては個人差以外に，認識に必要となる語彙サイズや，システムが受理すべき文集合の特性なども考えられる．文献 10) では 1109 語までの孤立単語の認識において，語彙サイズと認識性能の関係を理論的実験的に分析し，複合 2 項分布を用いることで語彙サイズがシステムの性能へ及ぼす影響を効果的にモデル化できることが示されている．また文献 11) では 500 語彙程度までのシステムを対象としてパープレキシティと文認識率の関係の分析が行われている．文献 10) や文献 11) では単語誤り率は語彙数の対数や平方根にほぼ比例することが示されている．ただしこれはシステムが対象とすべき語彙があらかじめ限定されている場合の結果である．話し言葉ではシステムに入力されるべき語彙集合をあらかじめ定めることはできず，単純に語彙数を制限することは未知語率の増加をまねく結果となる．様々な単語に対応するためには語彙サイズを大きくせざるをえず，このことも話し言葉の認識を難しくしている 1 つの要因といえる．

今後の課題としてはさらに広い範囲の発話や，人による講演の主観評価とデコーダによる認識性能の関係へと分析対象を広げること，本論文において同定された単語正解精度に対する影響の大きい属性への対処法を開発することがあげられる．

謝辞 プロジェクトの推進研究者各位に感謝する．

参考文献

1) 村上仁一，嵯峨山茂樹：自由発話音声における音響的な特徴の検討，信学論，Vol.J78–D–II, No.12, pp.1741–1749 (1995).
2) 松井知子，内藤正樹，ハラルドシンガー，中村篤，匂坂芳典：地域や年齢的な広がりを考慮した大規模な日本語音声データベース，1999 秋季音学講論集，pp.169–170 (1999).
3) Furui, S., Maekawa, K., Isahara, H., Shinozaki, T. and Ohdaira, T.: Toward the realization of spontaneous speech recognition, *Proc. ICSLP*, Vol.3, pp.518–521 (2000).
4) 西村雅史，伊東伸泰：講義コーパスを用いた自由発話の大語彙連続音声認識，信学論，Vol.J83–D–II, No.11, pp.2473–2480 (2000).
5) Entropic Ltd: *The HTK Book* (*for HTK Version 2.2*) (1999).
6) Clarkson, P. and Rosenfeld, R.: Statistical language modeling using the CMU-Cambridge toolkit, *Proc. Eurospeech*, Vol.5, pp.2707–2710 (1997).
7) Kawahara, T., Lee, A., Kobayashi, T., Takeda, K., Minematsu, N., Sagayama, S., Itou, K., Ito, A., Yamamoto, M., Yamada,

A., Utsuro, T. and Shikano, K.: Free software toolkit for Japanese large vocabulary continuous speech recognition, *Proc. ICSLP*, Vol.4, pp.476–479 (2000).
8) Akaike, H.: Information theory and an extension of the maximum likelihood principle, *Proc. ISIT*, pp.267–281 (1973).
9) 山本一公，中川聖一：発話スタイルによる話速・音韻間距離・ゆう度の違いと音声認識性能の関係，信学論，Vol.J83–D–II, No.11, pp.2438–2447 (2000).
10) Rosenberg, A.E.: A probabilistic model for the performance of word recognizers, *AT&T Bell Lab*, Vol.63, No.1, pp.1245–1277 (1984).
11) 中川聖一，大黒嘉久，村瀬 功：連続音声認識システムの評価法—タスクの複雑性と文認識率との関係，信学論，Vol.J73–D–II, No.5, pp.683–693 (1990).

(平成 13 年 11 月 15 日受付)
(平成 14 年 4 月 16 日採録)

篠崎 隆宏（正会員）

平成 11 年東京工業大学工学部情報工学科卒業．平成 13 年同大大学院博士前期課程修了．現在，同大学院博士後期課程在学中．音声認識に関する研究に従事．日本音響学会会員．

古井 貞熙（正会員）

昭和 43 年東京大学工学部計数工学科卒業．昭和 45 年同大大学院修士課程修了，NTT 電気通信研究所入社．以後，音声認識，話者認識，音声知覚等の研究に従事．ベル研究所客員研究員，NTT 基礎研究所第四研究室長，NTT ヒューマンインタフェース研究所音声情報研究部長，同古井特別研究室長を経て，平成 9 年東京工業大学大学院情報理工学研究科計算工学専攻教授．工学博士．科学技術庁，IEEE，電子情報通信学会，日本音響学会等より論文賞等受賞．著書「音響・音声工学」,「音声情報処理」等．IEEE および米国音響学会 Fellow．International Speech Communication Association（ISCA）会長．日本音響学会会長．