論文 / 著書情報
Article / Book Information

論題(和文)	話し言葉音声認識のための音響・言語モデル
Title	Acoustic and Linguistic Modeling for Spontaneous Speech Recognition
著者(和文)	篠崎隆宏, 堀智織, 古井貞熙
Author	Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	話し言葉の科学と工学ワークショップ予稿集, Vol. , No. , pp. 101-108
Journal/Book name	, Vol. , No. , pp. 101-108
発行日 / Issue date	2001, 3

話し言葉の科学と工学ワークショップ講演予稿集
東京工業大学、2001年2月28日～3月1日、pp. 101-108

話し言葉音声認識のための音響・言語モデル

篠崎 隆宏，堀 智織，古井 貞熙

東京工業大学大学院情報理工学研究科計算工学専攻
〒152-8552 東京都目黒区大岡山 2-12-1
Tel／Fax：03-5734-3480
Email: {staka, chiori, furui}@furui.cs.titech.ac.jp

あらまし　日本語話し言葉プロジェクトに関連して，講演音声を対象として進めている音声認識の研究状況を報告する．男性10名の話者による，のべ約4時間半にわたる種々の講演音声について，認識実験を行った．その結果，話し言葉コーパスから作成した音素モデルや言語モデルが，従来の主として書き言葉コーパスから作成したモデルと比較して，極めて有用であることが確認された．現在利用可能な話し言葉データに対して適切なモデルの精密さについても検討を行った．認識性能に個人差が大きく，発話速度，フィラー数，言い直し数などに関連していることを確認した．また，音響モデルの教師なし話者適応化が有効であることを確認した．しかし，話し言葉の音声認識性能はまだ低く，今後解決しなければならない研究課題が多い．

キーワード　話し言葉音声認識，話し言葉プロジェクト，講演，発話速度，教師なし話者適応

Acoustic and Linguistic Modeling for Spontaneous Speech Recognition

Takahiro Shinozaki, Chiori Hori and Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
Tel/Fax：03-5734-3480
Email: {staka, chiori, furui}@furui.cs.titech.ac.jp

Abstract　This paper reports various investigations on recognizing spontaneous presentation speech in connection with the "Spontaneous Speech" national project started in 1999. Various presentation speech uttered by 10 male speakers having approximately 4 hours and a half in total has been recognized. Experimental results show that acoustic and linguistic modeling based on actual spontaneous speech corpora is far more effective than the conventional modeling mainly using read speech. We examined the balance between precision and robustness of the models. The recognition accuracy has a wide speaker-to-speaker variability according to the speaking rate, the number of fillers, the number of repairs, and so forth. It was also confirmed that unsupervised speaker adaptation of acoustic models was effective to improve the recognition accuracy. However, the recognition accuracy for spontaneous speech is still rather low, and there exists a large number of research issues

Key words　spontaneous speech recognition, national project, presentation, speaking rate, unsupervised speaker adaptation

1. はじめに

自由発話された音声を十分な精度で認識することは音声認識の用途を広げていく上で重要である．従来の書き言葉を対象としたコーパスに基づくモデルでは，話し言葉に対応できず低い認識率しか得られていない．本論文では，まず話し言葉コーパスを利用したモデルと，従来の書き言葉を用いたモデルの比較検討[1]について述べる．後半では話し言葉コーパスを利用したモデルに関して，音響モデルの状態数や言語モデルの語彙サイズ，認識率の個人差，教師なし話者適応などについて検討する．

2. 使用コーパス

本研究では以下の二つのコーパスを用いている．

- **話し言葉コーパス**

話し言葉コーパスは構築中であり，現在も毎月利用可能なデータが増えている．本論文の内容は2000年12月の時点で利用可能なデータに基づいている．書き起こしデータには未チェックのものとチェック済みのものがあるが，両方混ぜて使用した．テストセットを除いた全講演数は610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．形態素にすると約1.5Mである．またそのうちで男性話者による講演は338講演である．

- **Web コーパス**

World Wide Web上で公開されている講演書き起こしテキストを本研究室で収集した．総形態素数は約2Mである．書き起こす際にフィラーや言い直しなどは取り除かれ文章として編集されている．話題は一般的なものである．

3. 実験条件と認識タスク

3.1 実験条件

音声は16kHzで標本化，16bitで量子化した．音響パラメータはMFCC12次元，Δケプストラム12次元，対数エネルギーの1次差分の25次元で，切り出した発話区間ごとに平均ケプストラムによる正規化(CMS)を行った．入力音声の切り出しは書き起こしに含まれるラベルに基づき，500ミリ秒の無音を基準とした．音響モデルの学習，話者適応にはHTK2.2を使用した．言語モデルの作成にはCMU SLM Tool Kit v2.05を使用した．使用した言語モデルは全てbackoff N-gramでGood-Turing discountingを使用している．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．デコーダにはJulius3.1を使用した．

3.2 認識タスク

話し言葉コーパス中の10講演を，認識対象(テストセット)として用いた．全て男性話者である．概要を表1に示す．正式データ名は表第1列の通りであるが，以下では簡単のため第2列に示す表記を用いる．表中はじめの4講演は音響学会，または音声学会の講演を収録したもので音声に関連した話題である．各講演の講演時間を表第4列に示す．

認識実験では，言語重み，挿入ペナルティは特にことわった場合を除いて講演毎に適した値を用いている．

表1 テストセットの概要

データ名称	略称	学会／研究会	講演時間
AS99SEP022	A22	日本音響学会	28分
AS99SEP023	A23	日本音響学会	30分
AS99SEP097	A97	日本音響学会	12分
PS99SEP025	P25	音声学会	27分
JL99OCT001	J01	国語学会	57分
KK99DEC005	K05	国語研究所	42分
NL00MAR007	N07	言語処理学会	15分
SG00MAR005	S05	社会言語科学会	23分
YG99JUN001	Y01	融合研究会	14分
YG99MAY005	Y05	融合研究会	15分

3.3 テストセットの特徴

テストセット中の各講演の発話速度を図1に示す．発話速度は，発話していない時間を除いた実際の発話時間を基に計算した．A22, P25, S05は発話速度が速い．

各講演のフィラーと言い直しの頻度を図2に示す．A22, P25, S05はフィラーが多い．A22, S05は言い直しも多い．全体としてフィラーと言い直しではフィ

がっている．音声の話題を扱った講演に対し，音声の教科書を加えるタスク適応が未知語率に関して有効であることがわかる．**Spon** の未知語率は全体に **WebSp** よりも低い．

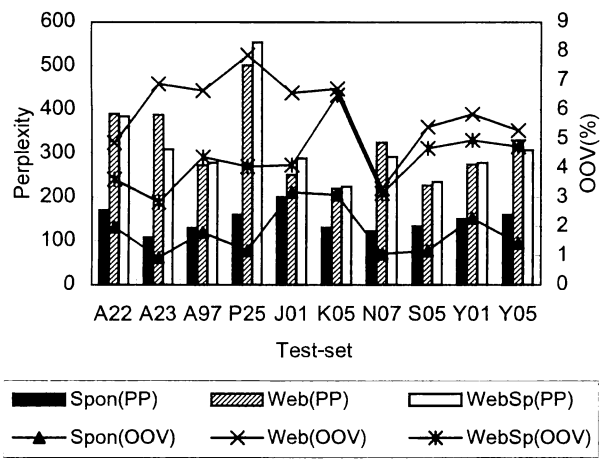

図 3 パープレキティと未知語率

5.3 言語モデルと認識率

言語モデルと認識率の関係を図 4 に示す．音響モデルは **TS2k** である．言語モデルに **Spon** を使用すると **Web** や **WebSp** に比較して高い認識率が得られ，話し言葉から作成したモデルが有効であることが分かる．**WebSp** は **Web** より高い認識率となり，話題適応が有効であることが分かる．

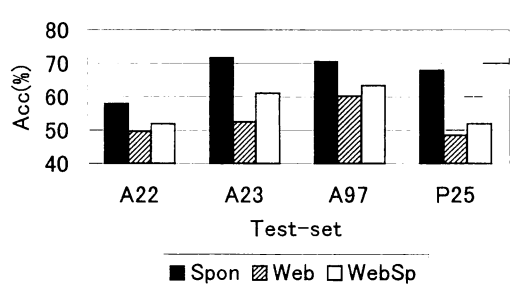

図 4 言語モデルと認識率

5.4 音響モデルと認識率

音響モデルと認識率の関係を図 5 に示す．言語モデルは **Spon** を使用した．従来の読み上げ音声に基づく **IPA2k** を使用した場合と比べ，話し言葉から作成した **TS2k**, **PTM2k** を使用すると高い認識率が得られ，話し言葉から作成したモデルが有効であることが分かる．

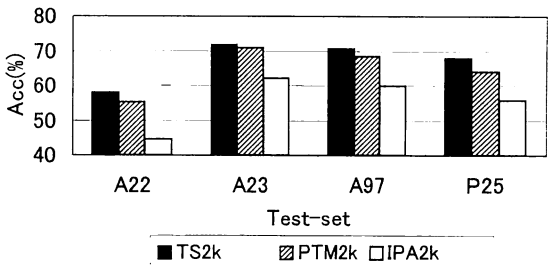

図 5 音響モデルと認識率

6. 話し言葉コーパスを利用したモデルの検討

本章以降では話し言葉コーパスを利用したモデルについて各種の検討を行う．

6.1 使用モデル

使用した言語モデルの概要を表 4 に示す．学習セットは全て **SponT** である．表第 1 列がモデル名を表している．モデル名中の *cxy* はバイグラムのカットオフが *x*，トライグラムのカットオフが *y* であることを示している．第 3 列は左から順に 2-gram, 3-gram のカットオフ数である．第 4 列は N-gram の種類を示している．例えば **v30k-c01** は 2-gram と逆向き 3-gram のセットであり，逆向き 3-gram のカットオフを 1 としたモデルである． また，**v30k-c011** は逆向き 4-gram であり，逆向き 3-gram と逆向き 4-gram のカットオフを 1 としたモデルである．

表 4 言語モデルの概要

言語モデル	語彙数	カットオフ (2-gram, 3-gram,...)	N-gram
v10k	10k	0, 0	2,rev3
v20k	20k	0, 0	2,rev3
v30k	30k	0, 0	2,rev3
v30k-c01	30k	0, 1	2,rev3
v30k-c02	30k	0, 2	2,rev3
v30k-c11	30k	1 ,1	2,rev3
v30k-c011	30k	0, 1, 1	rev4
v30k-c0111	30k	0, 1, 1, 1	rev5

音響モデルの概要を表 5 に示す． **TS1.5k**, **TS2k**,

ラーの頻度の方が高い．

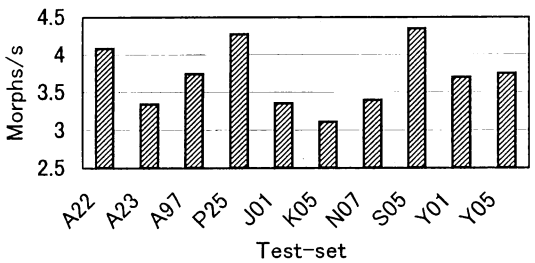

図 1　発話速度

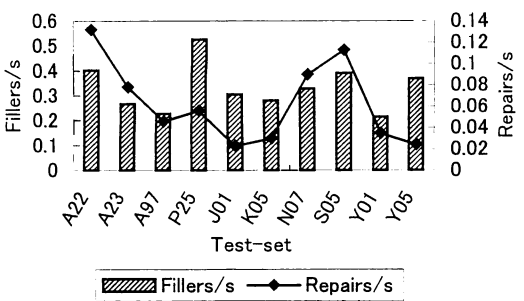

図 2　発話中のフィラーと言い直しの頻度

4. 学習セット

4.1 言語モデル用データ

言語モデルの作成に使用した学習データは以下の 3 セットである．

SponT　：話し言葉コーパス中の講演書き起こし 610 講演．テストセット中の講演話者と同一話者による講演は除外してある．読点に関しては 200ms 以上の無音を境界の基準としている転記基本単位の間に補った．

WebT　：Web コーパスの全テキスト．間投詞の補正を行った．

WebSpT　：**SponT** に教科書「音声情報処理」（総形態素数：63k）を足し合わせた．混ぜる際，重みつけはしていない．

4.2 音響モデル用データ

音響モデルの学習データは以下の 3 種類である．

SponS　：話し言葉コーパス中で男性による 338 講演．テストセット中の講演話者と同一話者による講演は除外してある．

SponfS　：SponS 中発話速度が上位 6 割の講演．

IPA　　：読み上げ音声約 40 時間．（話し言葉音声との比較用に，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」に含まれる音響モデルを使用した．）

5. 話し言葉・書き言葉に基づくモデルの比較

話し言葉コーパスのデータから作成した音響モデル・言語モデルと，書き言葉スタイルのデータから作成したモデルの比較を行った．

5.1 言語モデルと音響モデル

使用した言語モデルの概要を表 2 に示す．各言語モデルは 2-gram と逆向き 3-gram で構成されている．語彙数は全て 20k である．

表 2　言語モデルの概要

言語モデル	学習セット	学習形態素総数	語彙数
Spon	SponT	1.5 M	20 k
Web	WebT	2 M	20 k
WebSp	WebSpT	2+0.06 M	20 k

使用した音響モデルの概要を表 3 に示す．**PTM2k** は 2k 状態のトライフォンを元にした PTM(phonetic tied-mixture)[2]，その他は状態共有モデルである．

表 3 音響モデルの概要

モデル名	状態数	混合数	学習セット	データ量
PTM2k	129(2k)	64	SponS	59時間
TS2k	2k	16	SponS	59時間
IPA2k	2k	16	IPA	40時間

5.2 パープレキシティと未知語率

テストセット中の各講演について，図 3 に逆向き 3-gram のテストセットパープレキシティと未知語率を示す．話し言葉から作成した **Spon** では他のモデルに比べ低いパープレキシティが得られた．**Web** は書き起こす際に文章として編集されていること，話の内容が一般的なものであることなどからパープレキシティ，未知語率とも高くなっている．**WebSp** では図中左側の 4 講演で **Web** に比べ特に未知語率が下

TS3k は状態数がそれぞれ 1.5k, 2k, 3k の 3 状態 left-to-right 型のモデルである．**TSf2k** は **SponfS** から作成し，データ量が 36 時間である以外は **TS2k** と同様である．**TSt3k** は状態 1 から 3 への遷移を加えた以外は **TS3k** と同様である．これは遷移の際，状態のスキップを許すことにより，発話継続時間の短い音素に対応する目的で作成した[3]．混合数は全て 16 である．

表 5　音響モデルの概要

モデル名	状態数	学習セット	データ量
TS1.5k	1.5k	SponS	59時間
TS2k	2k	SponS	59時間
TS3k	3k	SponS	59時間
TSf2k	2k	SponfS	36時間
TSt3k	3k	SponS	59時間

6.2　音響モデルの状態数と認識率

音響モデルの状態数と認識率の関係を図 6 に示す．言語モデルは **v30k** である．どの状態数がよいかは講演により異なる．なお，言語重み，挿入ペナルティを各講演で共通にした場合は，**TS1.5k**, **TS2k**, **TS3k** の認識率の平均はそれぞれ 63.4%, 63.7%, 64.0%となり，**TS3k** で最もよくなった．学習データ量が 59 時間程度と比較的多いことから状態数の多いモデルが有効であることが分かる．

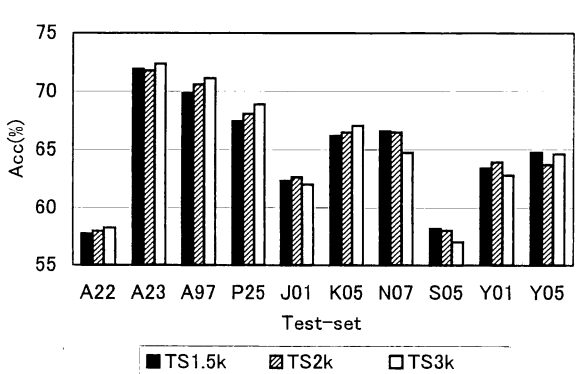

図 6　音響モデルの状態数と認識率

6.3　言語モデルの特性

6.3.1 未知語率

語彙数を 10k, 20k, 30k としたときの未知語率を図 7 に示す．J01 と K05 で未知語率が高い．**v30k** での未知語率は 1%程度である．なお，**v20k** は **Spon** と同一である．

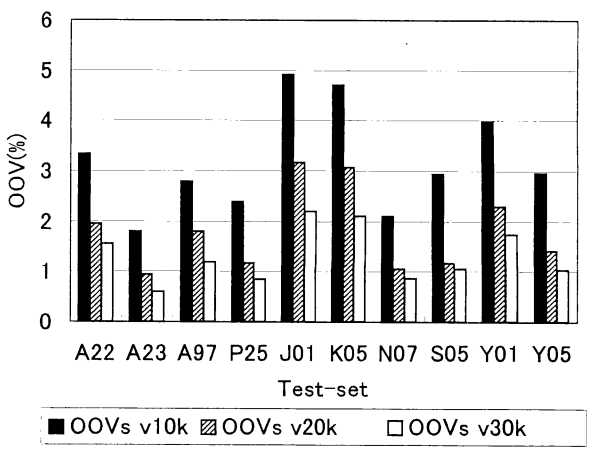

図 7 語彙数と未知語率

6.3.2 パープレキシティ

カットオフ方法の違う 4 種類の逆向き 3-gram モデルのパープレキシティを図 8 に示す．語彙数は全て 30k である．3-gram のカットオフを 1 としたモデルのパープレキシティが全体として一番低くなった．3-gram のカットオフを 2 とした場合，1 とした場合と比較し僅かに値が増えた．語彙数 10k, 20k の場合も図 8 同様の傾向が得られている．

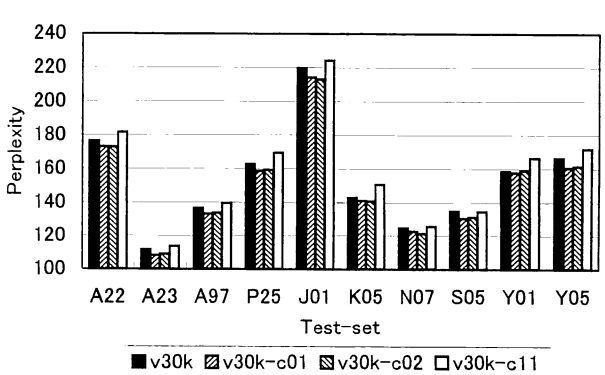

図 8 カットオフとパープレキシティ

N-gram のコンテキストの長さとパープレキシティの関係を図 9 に示す．バイグラムに関しては順向き，3-gram 以上に関しては逆向きモデルの値を示す．また，3-gram 以上に関してカットオフは 1 である．

2-gram と 3-gram を比較すると 3-gram の方がどの講演に対しても低い値となっており，トライグラムの効果がはっきり出ていることが分かる．3-gram と 4-gram を比較すると同程度であった．4-gram と 5-gram を比較すると，5-gram では値が大きくなった．**v30k-c011** で 4-gram のヒット率は 11%程度，**v30k-c0111** で 5-gram のヒット率は 3%程度であった．

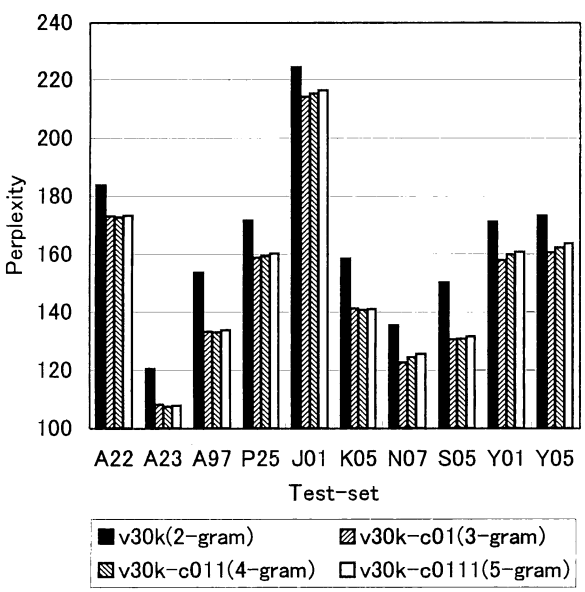

図 9　コンテキスト長とパープレキシティ

6.3.3　語彙数と認識率

語彙数と認識率の関係を図 10 に示す．音響モデルは **TS3k** を使用した．語彙数 30k のモデルを使用したときの認識率が全体として最もよいが，**v20k-c01** との差は最大で 0.5%程度であった．

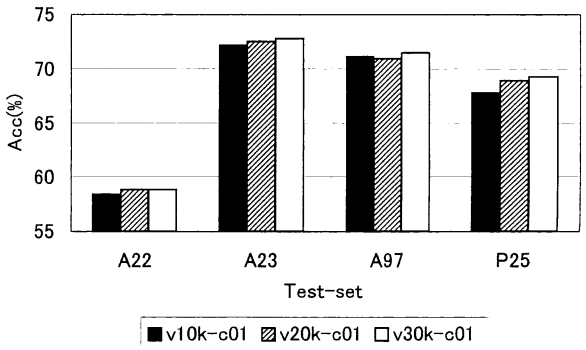

図 10　語彙数と認識率

6.4　発話の個人差

図 6 などに示すように A22 の認識率は 60%未満であるのに対し，A23 は 70%を超えるなど認識率に個人差が大きいことが分かる．本節では個人差について検討を行う．

6.4.1　発話の個人差と認識率

発話速度と認識率の関係を図 11 に示す．音響モデルは **TS3k**，言語モデルは **v30k-c01** である．図で 1 つの点は 1 つの講演に対応する．またこれらの点を近似した直線を重ねて示す．発話速度の速い講演の認識率が低いことが分かる．

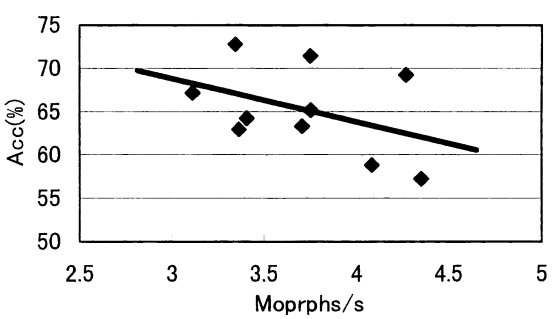

図 11　発話速度と認識率

フィラー頻度と認識率の関係を図 12 に，言い直し頻度と認識率の関係を図 13 に示す．フィラー頻度，言い直し頻度が多い講演で，認識率が低いことが分かる．

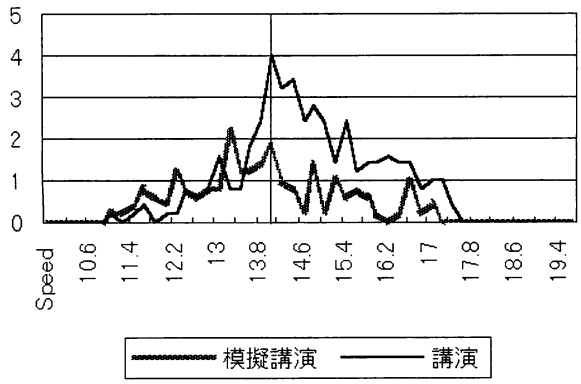

図 16　発話速度の分布

6.5　話者適応

音響モデルの教師なし話者適応による認識率の向上の効果について調べた．テストセットには A22, A23, A97, P25 の 4 講演を用いた．話者適応は以下の手順で行った．

1. 不特定話者モデルを用い，講演全体の認識を行う．4 講演の平均の認識率を最もよくする，4 講演共通の言語重み，挿入ペナルティを求める．

2. 1 で求めたパラメータに対応する認識結果を正解文として用い，それぞれの講演において話者適応を行う．

話者適応には MLLR を用いた．モデル中の全正規分布を予め 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 MLLR に使用するクラスタを決定する方法を用いた．

また上記話者適応で得られたそれぞれの講演に対するモデルを同様の手順で再度話者適応化したモデルも作成した．

認識結果を図 17 に示す．認識に用いた不特定話者音響モデルは **TS3k**，言語モデルは **v30k-c01** である．図で，上記手順で 1 回話者適応を行ったものが **mllr**，2 回行ったものが **mllr-i2** である．**TS3k** と **mllr** の認識率の差は 3%から 4%程度，**mllr** と **mllr-i2** との認識率の差は 1%程度であった．言語重み，挿入ペナルティを各講演で共通にした場合の 4 講演の平均認識率は **mllr-i2** では 71%となり，**TS3k** に比べ 13%の単語誤り率削減となった．

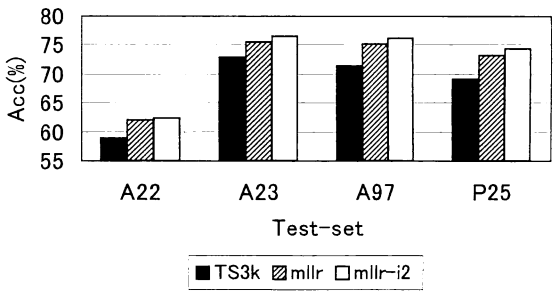

図 17　教師なし話者適応と認識率

7. まとめ

話し言葉コーパスを利用した言語モデルや音響モデルと従来の書き言葉スタイルのモデルの比較を行い，話し言葉コーパスを利用したモデルが講演音声の認識に有効であることを示した．2000 年 12 月時点で利用可能なデータ量に適したモデルの精密さの検討を行った．発話速度やフィラー頻度などが認識率に関係することを示した．教師なし話者適応を行うことで 4 講演の平均で 71%の認識率となった．

なお，本論文では認識率を求める際は講演ごとの言語重みや挿入ペナルティを使用したが，テストセットで共通とした場合，講演によっては認識率の値が本論文の値よりも 1%程度下がることがある．

今後の課題としては認識率をより向上させる必要がある．そのために発話速度やフィラー，言い直しへの効果的な対処法の検討，学習データをより増やしたモデルの作成などが挙げられる．また単に文字列を出力するのではなく発話の意味的まとまりも同時に抽出することも実用上重要となる．

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] 篠崎隆宏，斎藤洋平，堀智織，古井貞熙：“話し言葉音声の認識を目指して”，信学技報，SP2000-96 (2000-12)

[2] A.Lee, T.Kawahara, K.Takeda, and K.Shikano. “A new phonetic tied-mixture model for efficient decoding”. In Proc. ICASSP, pp. 1269-1272 (2000-6)

[3] 本間真一，小林彰夫，佐藤庄衛，今井亨，安藤彰男：“ニュース解説を対象にした音声認識の検討”信学技報，SP2000-99 (2000-12)

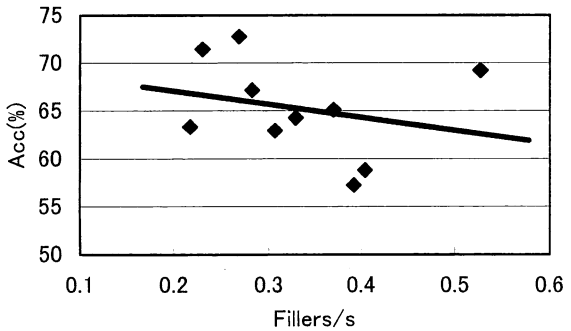

図 12　フィラー頻度と認識率

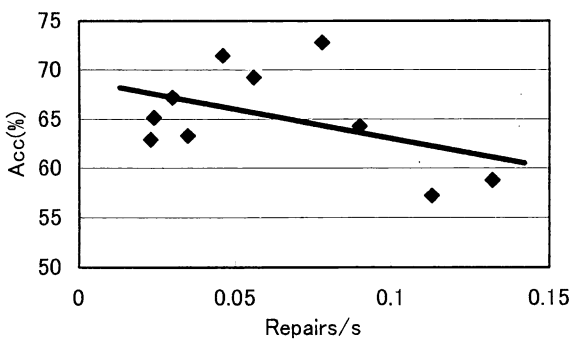

図 13　言い直し頻度と認識率

6.4.2　発話速度への対応の検討

発話速度の速い講演で低い認識率となることから，対応の検討を行った．

TSf2k は発話速度が上位 6 割に入る講演を選んだ学習セット **SponfS** から作成した音響モデルである．音響モデルに **TS2k** と **TSf2k** を使用した場合の認識率を図 14 に示す．言語モデルは **v20k** である．発話速度の速い講演のうち A22 で認識率が向上したものの，P25 では逆に下がった．また発話速度が速くない A23 で認識率が僅かながら上がっているなど，期待した効果は得られなかった．

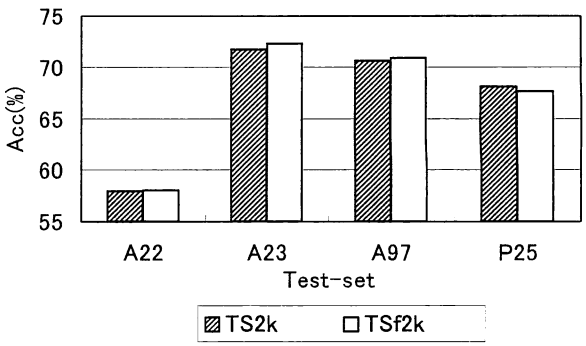

図 14　速い発話速度のモデル

発話速度へのもう一つの対応として音響モデル **TSt3k** を作成した．音響モデルに **TS3k** と **TSt3k** を使用した場合の認識率を図 15 に示す．言語モデルには **v30k-c01** を用いた．A22 や P25 で認識率が下がるなど期待した効果は得られなかった．

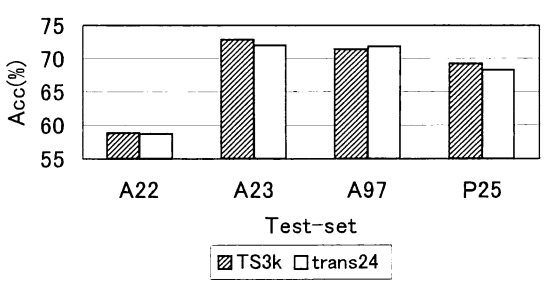

図 15　スキップを許すモデルと認識率

TSf2k は学習データ量が **TS2k** に比べ減っているにもかかわらず，全体的に認識率の低下が見られない．その理由の一つとして発話速度の上位 6 割には講演が多く残り模擬講演が残りにくいことから，講演音声からなるテストセットの認識に際し認識率の低下につながらなかったと考えられる．図 16 に学習セット中の通常の講演と模擬講演の発話速度の分布を示す．横軸は音素数で見た発話速度，縦軸は階級幅を 0.2 としたときの頻度である．図中の縦棒より左側が発話速度上位 6 割となる．