1994 ARPA Human Language Technology Workshop 参加報告

古井　貞熙
NTTヒューマンインタフェース研究所
〒180 武蔵野市緑町3-9-11

1994年3月8日～11日に、米国New Jersey州PrincetonのMerrill Lynch Conference Centerで開かれた、ARPA Human Language Technology (HLT) Workshopの出席報告である。ARPA HLT Programが進めている音声認識のタスクは、Wall Street Journalタスク（ディクテーション）とATISタスク（計算機システムとの対話）で、従来から変わっていないが、着実に語彙数やデータ量が増えている。今回の目新しい点は、CSRの評価に関して、多様な項目からなる"Hub and Spoke"パラダイムが決められ、robustnessが明確に表に出てきたことである。ATISに関しては、MADCOWの活動としてATIS-3 corpusが集められ、タスクとして難しさが増している。ベンチマークテストでは、CSRに関しては、米国外のLIMSI（フランス）とCambridge大（イギリス）がトップを占めるという皮肉な結果になった。ATISに関しては、米国のBBNとCMUがトップを占めた。

Report of the 1994 ARPA Human Language Technology Workshop

Sadaoki Furui
NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan

This paper is a report of the 1994 ARPA Human Language Technology (HLT) Workshop, which was held at the Merrill Lynch Conference Center in Princeton, New Jersey, from March 8 to 11, 1994. The ARPA HLT Program included two speech recognition tasks: the Wall Street Journal task (dictation) and the ATIS task (dialog with a computer system). Although these tasks have not been changed over the last few years, the sizes of the vocabulary and the database have been increased each year. One of the topics this year was the establishment of the "Hub and Spoke" paradigm for CSR evaluation, which consists of various issues mainly focusing on the robustness of the technology. Concerning the ATIS task, the ATIS-3 corpus was collected as an activity of MADCOW, which was more difficult than previous corpora. In the benchmark tests, the two best results of CSR were from volunteers, LIMSI (France) and Cambridge University (UK), and the two best results of ATIS were from contractors, BBN and CMU.

1. ARPA Human Language Technology Workshop

1994年3月8日～11日に、米国 New Jersey州 PrincetonのMerrill Lynch Conference Centerで、ARPA Human Language Technology (HLT) Workshopが開かれた。ARPA HLT Programの参加機関の代表に、主な研究機関からの招待者を加えて、200名以上の参加者があった。音声関係では、日本からATRの匂坂氏と私が参加する機会を得たので、会議での発表や議論の中から、最近の動向のポイントと思われる部分をまとめてみた。本ワークショップに先立って、同じ会場で2日間にわたって、Spoken Language Systemsのワークショップが開かれたが、内容的にはHLTワークショップとの重複が多いので省略する。昨年までの動向については、文献[1, 2]を参照頂きたい。なお今回のProceedingsも、例年通りMorgan Kaufmann Publishersから出版される予定である[3]。

2. 全体的な動向

ARPA Human Language Technology Programには、次のプロジェクトが含まれている。

- Robust speech recognition
- Domain-specific speech understanding
- Domain-specific text understanding
- Machine translation of text
- TIPSTER: Text understanding
- DIMUND: Document understanding

プログラム・マネジャーであるGeorge Doddingtonによるoverviewの内、音声言語処理に関する主な内容は以下の通りである。図1に示すように、音声言語処理をNational Information Highwayの出入り口を提供するものと位置づけている。現在のミッションは、基幹技術の開発、技術のデモ、役に立つ場への技術移転、の3点である。具体的には、技術の研究開発のための技術目標の設定と研究開発基盤の整備（データベース、評価法、システムの評価）、技術移転のサポートなどを行なっている。現状認識としては、計算機パワーの向上が音声言語処理の可能性を増していること、C&Cの情報化社会が進展していること、ハイテクへの要請が高まっていることを上げ、総じて順風であると述べていた。さらに、計算機パワーの進歩により、高性能低コストのソフトが実行できる時代になったことを強調していた。

評価の重要性が強く認識されているのは前からであるが、その主な目的は、R&Dの方向づけ、技術的進歩の測定とR&Dへのフィードバックなどである。具体的に行なっている内容は、客観尺度の提供、評価ツールの提供、評価用データの提供、学習用データの提供などである。

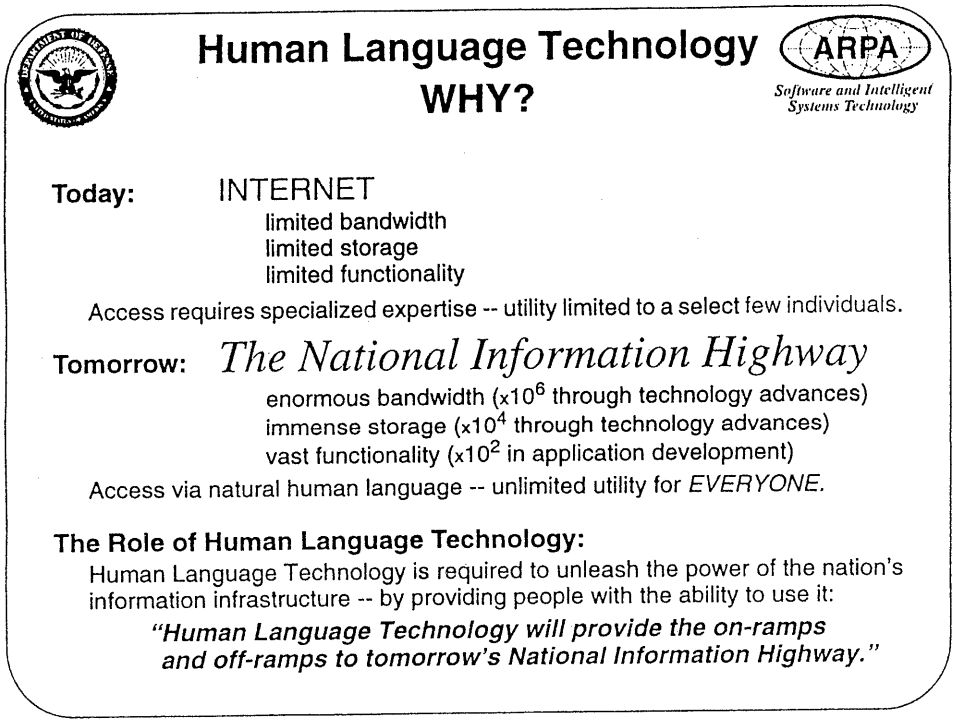

図1．ARPAのHLT(Human Language Technology)の役割

図２～３には、音声認識技術の進歩が認識誤り率の変化で示されている。この１年の数値的進歩は、Wall Street Journalタスク（5K語）とATISタスクに共通して小さい。

今後の方向としては、CSR（Wall Street Journalのディクテーション）に関しては無限語彙と電話音声、ATISに関しては電話音声と対話を目標にしている。評価に関しては、単なるマクロな認識率でなく、中身が理解できるような方法"SemEval"（Semantic Evaluation）や、タスクに独立な方法を目指している。成果の移転先としては、ATISの実用化と音声翻訳が考えられている。

3.CSRの評価のためのHub and Spokeパラダイム[4]

3.1 概要

BBNのFrancis Kubalaを中心とするCSR Corpus Coordinating Committee(CCCC)で、CSR（直接的にはNovember 1993 test data）を多面的に評価するHub and Spokeパラダイムが決められた。"Hub"テストは、すべてのsiteに共通で必須なテスト法で、site間の比較が目的である。一方"Spoke"テストは、各siteが選択して使用することができる。問題の難しさを正規化するため、Hubテストの結果と比較できるようになっている。学習とテストのデータにミスマッチがある場合など、主にrobustnessに関するテスト内容になっており、システム内でのアルゴリズムの比較ができる。

話者の変動への対処法は、次の３つのカテゴリーに分けられている。

- Static SI：不特定話者（適応なし）
- Unsupervised incremental adaptation：認識結果を適応に用いる
- Supervised incremental adaptation：認識後に正しいtranscriptionが与えられる

最も典型的なテストは、10話者（男女半々）各20-40文（語彙数: 64K語）の音声を、Sennheiser HMD-410マイクロホンで収録し、Static SIで評価するという方法である。

3.2 1993 Hub

(1) Read WSJ Baseline (H1)

Primary H1 (H1-P0): 64K語、語彙外なし、任意の言語モデル、Unsupervised incremental adaptation可

Contrastive test (H1-C1): 指定言語モデル（20K語3-gram、35M語からなる３年間のWSJのテキストから作成）、Static SI、音響学習データは37.2K発声（62時間）（284話者、各100-150発声、又は37話者、各600又は1200発声）、語彙外あり

H1-C2:20K語bigram、他の条件はH1-C1と同じ

(2) 5K-Word Read WSJ Baseline (H2): 5K語以外はわずか

H2-P0、H2-C1（5K語bigram）の性格は上と同様（縮小版、84または12話者の7.2K発声による音響学習）

3.3 1993 Spokes

S1：言語モデルの適応化（WSJ内、supervised/unsupervised）

S2：言語モデルの適応化（San Jose Mercury vs. WSJ）

S3：話者適応化（non-native speakers）

S4：話者適応化（incremental adaptation、supervised/unsupervised）

S5：マイクロホン適応化（伝送路も含む、unsupervised、各発声毎）

S6：マイクロホン適応化（既知の２種：電話と卓上マイク、unsupervised、各発声毎）

S7：周囲雑音（S6と同じ２種のマイク、雑音下）

S8：周囲雑音（S7と同様、異なる雑音）

S9：大語彙CSRの応用（記者による記事のspontaneous dictation）

S1, S2, S9以外は5K語の読み上げ音声

S5-S8は種々の２次マイクロホン使用、他はSennheiserマイクロホン

4. ATIS-3 Corpus[5]

MADCOW(Multi-Site ATIS Data COllection Working group)の活動として、ATIS-3 corpusが集められている。46の米国とカナダの都市と、23,457のflightsが含まれている。表１に示すように、BBN, CMU, MIT, NISTおよびSRIで集められた8297の学習用発声と、3211のテスト用発声からなる。2906の学習用発声はannotation（注釈付け）されている。ほとんどの発声の書き下しと解釈は、実際のATISシステム自身によって行なわれている。MITだけが書き下しにwizardを用いているが、MITでも解釈には自然言語システムを用いている。annotationされているものの内、文脈独立（A）、文脈依存（D）、評価不能（X）の発声の内訳を表２に示す。

評価法に関して、昨年提案されたend-to-end evaluationの他、SemEval（前出）が検討されている。前者には

(1) タスク終了までの所要時間

(2) システムの答えの適切さとユーザの解に関しての、人によるlogfileを用いた評価

(3) ユーザの満足度

などを含んでいる。

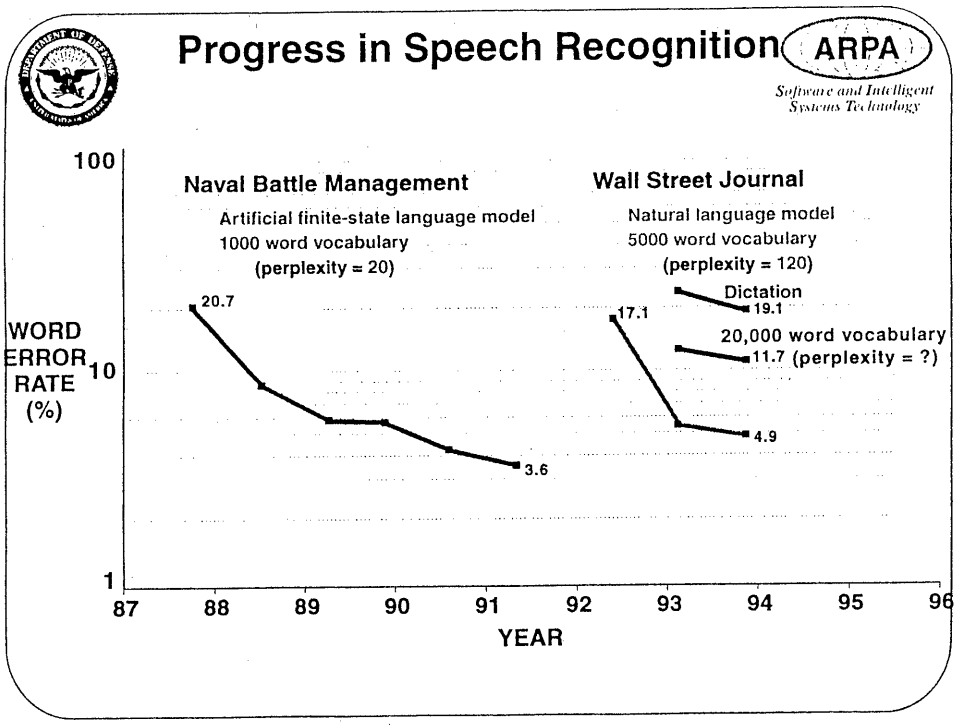

図2．ARPAの旧タスクとWSJタスクの音声認識の進歩

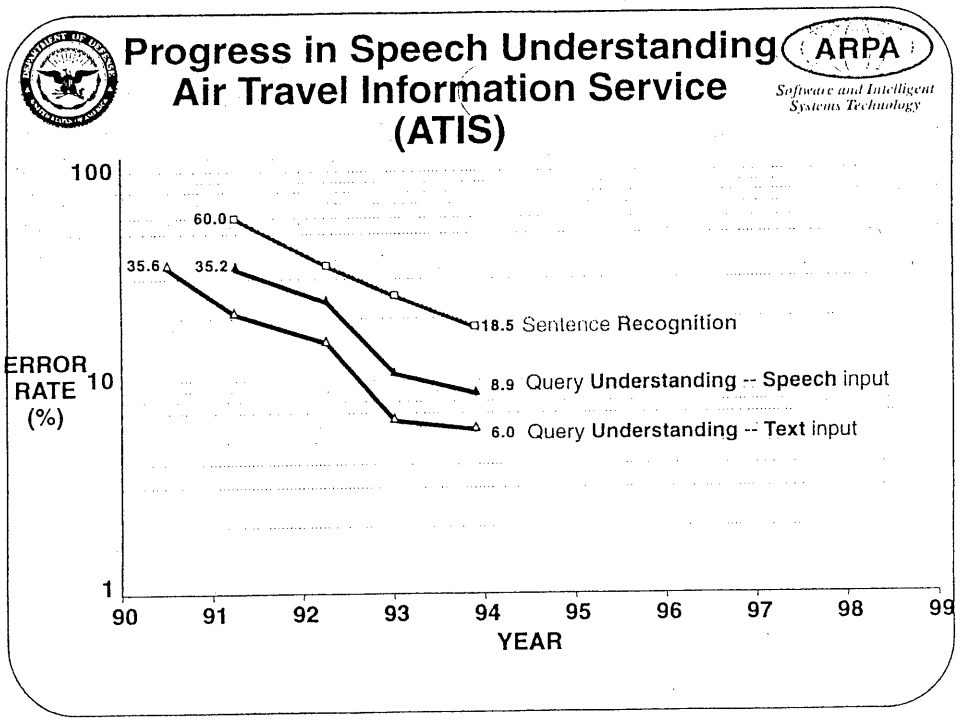

図3．ARPAのATISタスクの音声認識の進歩

表1．ATIS-3 corpusの内訳

Training Pool (including 1993 Test Data)				Test			SemEval Dry Run			Total		
Site	#Spkr	# Sess	#Utts	#Spkr	#Sess	#Utts	#Spkr	#Sess	#Utts	#Spkr	#Sess	#Utts
BBN	14	55	1101	9	37	389				23	92	1490
CMU	15	177	1462	8	70	387				23	247	1849
MIT	30	146	954	25	120	418				55	266	1372
NIST	49	253	2510	22	179	201	12	67	500	71	432	3211
SRI	17	62	2270			418						2688
Total	125	693	8297			1813	12	67	500			10610

表2．ATIS-3 corpusのClass A, D, Xへの分類

Training Data					December 93 Evaluation Test Data					Total
Site	Class A	Class D	Class X	Total	Site	Class A	Class D	Class X	Total	
BBN	282	182	193	657	BBN	89	57	53	199	856
CMU	417	197	103	717	CMU	113	50	36	199	916
MIT	391	239	121	751	MIT	82	50	32	164	915
SRI	329	351	106	786	SRI	80	86	34	200	986
NIST	0	0	0	0	NIST	84	82	82	203	203
Total	1419	969	523	2911		448	325	192	965	3876

5．音声言語システムのベンチマークテストの結果[6]

5.1 WSJ-CSR

WSJ-CSRに関しては、3章で述べたHub and Spokeパラダイムに従って、ベンチマークテストが行なわれた。その結果、Hub1テストでは、フランスのCNRS-LIMSIが最も低い単語誤り率を達成したが、英国のCambridge大のHTKを使ったシステムの結果との統計的有意差はない。この両者は共に連続型HMMを用いており、状態を結びにしたりしているが、半連続HMMを用いる場合よりも結びが少ない。

表3と表4は、それぞれHub1とHub2テスト全体の結果をまとめて示したものである。Spokeの結果は省略するが、robustnessに関しては技術的にまだかなり問題がある。

BBNでは、言語モデルを20K語から40K語に増やして実験を行ない、テスト文中の未知語が2.5%から0.2%に減った上、元の20K語の認識率はあまり低下しなかったと報告している。

5.2 ATIS

ATISのSPRECの結果を表5に、NLの結果を表6に、SLSの結果を表7に示す。昨年は重みつき誤りでNLとSLSの評価が行なわれ、"wrong answer"に"no answer"の2倍の重みをかけて加算していたが、今年は重みなしになった。これにより、"no answer"の使用が減った。Class A+DのSPRECでは、BBNとCMUが最も低い単語誤り率を示し、この両者の結果には統計的有意差はない。

なお、今年は全発声の46%がClass A、34%がClass D（計80%が""answerable"=Class A+D）で、昨年に比べるとClass Dが25%から大幅に増えている。このためと、今年のClass Dには難しい質問が多数含まれているため、昨年に比べてClass Dの誤り率が大幅に増加している。

6．興味を引いた論文

(1) Cambridge大学[7]

HMMに基づく（一般的に統計的手法による）音声認識において重要な点の一つは、モデルの複雑さと学習データ量とのバランスをとることである。このため、本論文ではtriphoneモデルのHMMの種々の状態を、音素決定木を用いたクラスタリングによって結びにする方法を提案している。これにより、従来のデータに基づく方法と同様の認識精度が得られ、しかも学習データに含まれないtriphoneのマッピングができるようになった。（CMUの方法と類似）

表3．Hub 1テストの結果

Nov 93 Hub and Spoke CSR Evaluation
Hub 1: 64K Read WSJ Baseline

GOAL: improve basic SI performance on clean data.
DATA: 10 speakers * 20 utts = 200 utts 64K-word read WSJ data, Sennheiser mic.

Primary and Contrast Conditions

P0 (opt) any grammar or acoustic training, session boundaries and utterance order given as side information.

C1 (req) Static SI test with standard 20K trigram open-vocab grammar and choice of either short-term or long-term speakers

C2 (opt) Static SI test with standard 20K bigram open-vocab grammar and choice of either short-term or long-term speakers

SIDE INFO: Session boundaries and utterance order are known for H1-P0 only.

	Primary P0	Contrast C1	Contrast C2
System	Word Err. (%)	Word Err. (%)	Word Err. (%)
bbn1	12.2	14.2	
bu1		15.7	
bu2		14.3	
bu3		14.5	
cmu1		13.6	
cmu2	13.9		
cu-htk1		12.7	14.4
dragon1		19.0	
limsi1		11.7	15.2
mit-ll1	16.8	18.6	
philips2		14.8	17.2
sri1		14.4	16.5

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
bbn1	P0:C1	13.9%	P0	same	P0	P0
mit-ll1	P0:C1	9.8%	P0	P0	P0	P0

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
cu-htk1	C1:C2	11.7%	C1	C1	C1	same
limsi1	C1:C2	22.7%	C1	C1	C1	C1
philips2	C1:C2	14.0%	C1	C1	C1	same
sri1	C1:C2	13.0%	C1	C1	C1	C1

表4．Hub 2テストの結果

Nov 93 Hub and Spoke CSR Evaluation
Hub 2: 5K Read WSJ Baseline

GOAL: improve basic SI performance on clean data.
DATA: 10 speakers * 20 utts = 200 utts 5K-word read WSJ data, Sennheiser mic.

Primary and Contrast Conditions

P0 (opt) any grammar or acoustic training, session boundaries and utterance order given as side information.

C1 (req) Static SI test with standard 5K bigram closed-vocab grammar and choice of either short-term or long-term speakers from WSJ0 (7.2K utts).

SIDE INFO: session boundaries and utterance order are known for H2-P0 only.

	Primary P0	Contrast C1
System	Word Err. (%)	Word Err. (%)
bu1	6.7	11.6
bu2	5.4	10.3
bu3	5.8	10.8
cu-con1		13.5
cu-htk2	4.9	8.7
cu-htk3		12.5
icsi1		17.7
limsi2	5.2	9.3
philips1	9.2	12.3
philips2	6.4	

COMPARISONS AND SIGNIFICANCE TESTS

	Test Comp.	% Change W.E.	Significance Tests: MAPSSWE	Sign	Wilcoxon	McN
bu1	P0:C1	42.4%	P0	P0	P0	P0
bu2	P0:C1	47.4%	P0	P0	P0	P0
bu3	P0:C1	46.6%	P0	P0	P0	P0
cu-htk2	P0:C1	43.4%	P0	P0	P0	P0
limsi2	P0:C1	43.7%	P0	P0	P0	P0
philips1	P0:C1	25.5%	P0	P0	P0	P0

(2) SRI[8]

Cambridge大学がHMMの状態を結びにしているのに対して、SRIではmixtureを複数のモデル間で結びにする方法を提案している。

(3) CMU[9]

CMUではR. M. Sternを中心に、cepstral compensationを基本とするrobust音声認識の研究を続けている。これまでの方法には、CDCN、SDCN、FCDCN、MFCDCNなどがある。今回提案しているのは、PDCN (phoneme-dependent cepstral normalization) と称する方法で、MFCDCNと似ているが、補正ベクトルが音素仮説に依存して変わる点が異なる。

(4) N-best対1-pass

ARPAのタスクのdecoderとして、N-bestを用いているBBNやCMUと、1-passを用いているSRI、MIT Lincoln Lab[10]、Cambridge大[11]などがある。N-bestには種々のメリットがあるが、欠点として一旦候補から落ちてしまうと回復できない問題がある。このためBBNの今回の論文[12]では、高度な知識をできるだけまとめてpassの数を減らすことを試みたが、誤りはほとんど減らなかった。このことから、（少なくとも現在のタスクでは）N-bestにはまだ意味があると結論づけている。

7. デモセッション

このARPAのワークショップでは、デモセッションが一つのhighlightになっている。今年は、以下の約10件のデモが行なわれた。いずれも、ワークステーション（HP735, IRIS Indigo, Sun SPARC10など）についているAD変換器のみを使った、アクセラレータ・ボードを使わないソフトウェアのみのシステムである。

(1) CMU
- 文章（文字）の読み方の訓練機
- ATISデモ

(2) BBN
- データベース検索 (Air Force Resource Management)
- 40K語WSJ
- データ入力システム

(3) SRI
- 英語→スウェーデン語自動翻訳

(4) MIT
- GALAXYシステム（実on-lineデータサービス、天気、地理案内、航空券など、多言語インタフェース）

(5) その他
- Microsoftの"Whisper"のビデオデモ
- New Mexico State大の情報検索システム
- Unisysの情報検索システム

8. むすび

ARPAの音声言語処理プロジェクトの最近の動向について、3月に開かれたワークショップを元に報告した。

文　献

[1] "Proceedings Human Language Technology Workshop", Morgan Kaufmann Publishers (1994)

[2] 古井貞熙：" DARPAの音声言語処理プロジェクトの現状—第5回ワークショップ（1992年2月）報告—"、信学技報、SP92-35 (1992)

[3] 古井貞熙：" 米国における音声理解システムの動向—ARPAシステムの現状と今後—"、音声入出力方式専門委員会資料、日本電子工業振興協会(1993)

[4] F. Kubala et al.: "The Hub and Spoke paradigm for CSR evaluation", Proc. ARPA Human Language Technology Workshop, pp.40-44 (1994)

[5] D. A. Dahl et al.: "Expanding the scope of the ATIS task: The ATIS-3 corpus", Proc. ARPA Human Language Technology Workshop, pp.45-50 (1994)

[6] D. S. Pallet et al.: "1993 Benchmark tests for the ARPA spoken language program", Proc. ARPA Human Language Technology Workshop, pp.51-73 (1994)

[7] S. J. Young et al.: "Tree-based state tying for high accuarcy acoustic modelling", Proc. ARPA Human Language Technology Workshop, pp.286-291 (1994)

[8] V. Digalakis and Hy Murveit: "High-accuracy large-vocabulary speech recognition using mixture tying and consistency modeling", Proc. ARPA Human Language Technology Workshop, pp.292-297 (1994)

[9] F.-H. Liu et al.: "Signal processing for robust speech recognition", Proc. ARPA Human Language Technology Workshop, pp.309-314 (1994)

[10] D. B. Paul: "The Lincoln large-vocabulary stack-decoder based HMM CSR", Proc. ARPA Human Language Technology Workshop, pp.374-379 (1994)

[11] J. J. Odell et al.: "A one pass decoder design for large vocabulary recognition", Proc. ARPA Human Language Technology Workshop, pp.380-385 (1994)

[12] L. Nguyen et al.: "Is N-best dead?", Proc. ARPA Human Language Technology Workshop, pp.386-389 (1994)

表5．ATISタスクのSPRECの結果

Class A+D+X Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-adx	10.2	91.5	6.2	2.3	1.7	46.3	964
bbn3-adx	4.1	97.0	2.4	0.7	1.0	20.9	964
cmu2-adx	4.1	96.8	2.4	0.9	0.9	22.1	964
crim3-adx	8.3	94.5	4.6	0.9	2.8	36.8	964
mit_lcs2-adx	5.6	95.2	3.3	1.5	0.8	28.4	964
sri3-adx	5.4	95.5	3.0	1.5	0.9	27.5	964
sri4-adx	5.2	95.7	2.8	1.4	0.9	26.0	964
unisys2-adx	5.2	96.1	3.0	0.9	1.3	26.3	964
unisys3-adx	4.9	96.3	2.9	0.9	1.2	23.5	964

Class A+D Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-a_d	9.0	92.5	5.4	2.1	1.5	42.2	773
bbn3-a_d	3.3	97.5	2.0	0.5	0.8	18.0	773
cmu2-a_d	3.3	97.3	2.0	0.7	0.6	19.7	773
crim3-a_d	6.6	95.7	3.6	0.7	2.3	31.3	773
mit_lcs2-a_d	4.5	96.1	2.6	1.2	0.7	23.3	773
sri3-a_d	4.6	96.1	2.5	1.4	0.7	23.3	773
sri4-a_d	4.5	96.3	2.3	1.3	0.8	22.3	773
unisys2-a_d	4.0	97.0	2.3	0.7	1.0	21.9	773
unisys3-a_d	3.9	97.1	2.3	0.6	0.9	19.7	773

Class A Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-a	8.6	93.1	5.0	1.9	1.7	44.4	448
bbn3-a	3.0	97.9	1.7	0.4	0.9	18.5	448
cmu2-a	3.0	97.5	1.8	0.7	0.6	19.9	448
crim3-a	6.3	96.1	3.2	0.7	2.4	31.9	448
mit_lcs2-a	4.3	96.4	2.4	1.1	0.8	24.1	448
sri3-a	4.0	96.6	2.0	1.3	0.7	22.1	448
sri4-a	3.9	96.8	1.9	1.3	0.7	21.0	448
unisys2-a	3.6	97.4	1.9	0.7	1.0	20.8	448
unisys3-a	3.5	97.4	1.9	0.6	1.0	19.6	448

Class D Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-d	9.6	91.4	6.1	2.4	1.0	39.1	325
bbn3-d	4.0	96.8	2.6	0.6	0.8	17.2	325
cmu2-d	3.9	96.8	2.5	0.7	0.7	19.4	325
crim3-d	7.2	94.9	4.3	0.7	2.1	30.5	325
mit_lcs2-d	4.9	95.6	3.1	1.4	0.5	22.2	325
sri3-d	5.7	95.2	3.4	1.5	0.9	24.9	325
sri4-d	5.5	95.4	3.2	1.4	0.9	24.0	325
unisys2-d	4.9	96.2	3.2	0.6	1.1	23.4	325
unisys3-d	4.4	96.4	2.9	0.6	0.9	19.7	325

Class X Subset

	W. Err	Corr	Sub	Del	Ins	U. Err	# Utt.
att2-x	15.5	87.3	9.7	3.0	2.8	62.8	191
bbn3-x	7.2	94.7	3.9	1.4	1.9	32.5	191
cmu2-x	7.3	94.7	3.8	1.5	2.0	31.9	191
crim3-x	15.0	89.7	8.7	1.6	4.7	59.2	191
mit_lcs2-x	10.0	91.5	5.9	2.5	1.6	49.2	191
sri3-x	8.7	92.8	5.2	1.9	1.6	44.5	191
sri4-x	8.0	93.3	4.9	1.8	1.3	41.4	191
unisys2-x	10.1	92.3	5.9	1.8	2.4	44.5	191
unisys3-x	9.3	93.0	5.2	1.8	2.4	39.3	191

表6．ATISタスクのNLの結果

	Class A+D 773 Utts.	Class A 448 Utts.	Class D 352 Utts.
system	UW Err.	UW Err.	UW Err.
att1	10.2	7.4	14.2
bbn1	14.7	9.6	21.8
bbn2	22.4	16.1	31.1
cmu1	9.3	6.0	13.8
crim1	36.4	21.7	56.6
crim2	20.8	14.7	29.2
mit_lcs1	12.5	10.0	16.0
sri1	21.9	14.3	32.3
sri5 **	18.2	10.5	28.9
unisys1	43.1	28.6	63.1

表7．ATISタスクのSLSの結果

	Class A+D 773 Utts	Class A 448 Utts.	Class D 352 Utts.
system	UW Err.	UW Err.	UW Err.
att1	24.6	22.1	28.0
bbn1	17.5	13.8	22.5
cmu1	13.2	8.9	19.1
crim1	43.3	28.6	63.7
crim2	28.2	23.7	34.5
mit_lcs1	14.2	11.8	17.5
sri1	24.8	16.5	36.3
sri2	25.4	18.5	34.8
sri5 **	20.7	14.1	29.8
sri6 **	21.2	13.8	31.4
unisys1	46.8	33.5	65.2