音声言語情報処理 34－23
(2000. 12. 22)

講演音声認識のための音響・言語モデルの検討

加藤 一臣　南條 浩輝　河原 達也

京都大学大学院 情報学研究科 知能情報学専攻
〒606-8501 京都府京都市左京区吉田本町
e-mail: {kazuomi,nanjo,kawahara}@kuis.kyoto-u.ac.jp

あらまし　融合研究プロジェクトにおいて構築が進められている講演音声と書き起こしテキストのデータベースを用いて、講演音声の認識のための音響・言語モデルを作成した。男性4名による音声・言語関係の学会講演を対象として評価を行った。これらのデータには話し言葉に固有の表現や発声の怠けなどが頻出する。利用できる講演の種類と量が多様であるため、音響・言語モデルそれぞれの学習にどのようなデータの組み合わせが適当であるかの検討を行った。音響モデルの学習には、学会講演というスタイルに一致させることの効果を確認し、言語モデルにおいても同様の傾向を確認した。現時点では平均59.8%の単語認識精度を得ることができた。

キーワード　音声認識、講演、話し言葉、統計的音響モデル、統計的言語モデル

Acoustic and Language Models for Lecture Speech Recognition

Kazuomi Kato　Hiroaki Nanjo　Tatsuya Kawahara

Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
e-mail: {kazuomi,nanjo,kawahara}@kuis.kyoto-u.ac.jp

Abstract　The acoustic and language models for automatic lecture speech recognition are addressed. We have constructed both acoustic and language models with the lecture corpus that is being developed under the priority research program of the science and technology agency. In order to investigate the best data-set combination, we evaluated sets of models with conference lectures by four male experienced speakers. We confirmed significance of matching the training data-set style of the acoustic model to the conference lecture style. A similar tendency was also found in the language model training. So far, we have got a word accuracy of 59.8%.

key words　automatic speech recognition, lecture speech, spontaneous speech, statistical acoustic model, statistical language model

1 はじめに

ディクテーションシステムに代表される「書き言葉」に対する音声認識は、大語彙連続発声でも9割程度達成されており[1]、成功を収めている。これらは、発話が辞書の表記どおりの明瞭なものであるという前提に基づくものである。一方、人間どうしの会話のような「話し言葉」に対する音声認識は、十分な成果が得られていない。

これまでの「話し言葉」を対象とした音声認識の研究は、主として旅行案内や情報検索といった対話音声を対象とした研究であった。これらはドメイン限定で語彙数も少なく、話者は計算機に認識されることを意識している。そのため、タスクとしては比較的容易であるが、汎用性はほとんどない。任意語彙の「話し言葉」の自動書き起こしが可能であればドメインを限る必要はなく、汎用的なシステムの実現が可能となる。この技術は話題同定や音声理解、講演録や会議の議事録作成や同時通訳など様々な用途へ応用ができる。

「話し言葉」においても純粋な日常会話はくだけすぎであり、また学習コーパスの収集自体が困難である。一方、講演音声は「書き言葉と話し言葉の中間」と考えられ[2]、発話はある程度整っているが自然である。講演者は認識器を意識しておらず、話し言葉の特徴を多く有すると考えられる。さらに、講演音声は、その自動書き起こしに対する要望も大きい。また、開放的融合研究『話し言葉工学』プロジェクト[3]において、大規模な講演音声と書き起こしのコーパスが収録されており、本格的な話し言葉のモデルの研究が可能になってきた。本研究では同コーパスを試験的に用いて、講演音声を認識するための音響・言語モデルを作成し、予備的な評価を行った。

2 講演音声データベース

融合研究コーパスは様々な学会講演と模擬的な講演から構成される。まだ収集、書き起こしの途上であり、現在は音声・言語関連の学会発表が中心となっているが、その内訳は以下の通りである。

- 日本音響学会春季&秋季研究発表会 (AS)
- 電子情報通信学会音声研究会 (SP)
- 言語処理学会年次大会 (NL)
- 国語学会 (JL)
- 音声学会全国大会 (PS)
- 国立国語研究所内で行われた種々の研究会 (KK)
- 融合研究会 (YG)
- 模擬講演 (IG,ST)…テーマ自由

これらには、音声データと人手による書き起こしテキストが存在し、書き起こしについては (1) 一通りの書き起こししかないもの (trans1) と (2) 表記も含めてチェック済の書き起こしがあるもの (trans4) の2種類存在する。音響モデルの学習には trans1 も含め、言語モデルの学習には trans4 のみを使用した。

表 1: 本研究で用いた融合研究コーパス（2000年10月）

講演種	音響モデル用	言語モデル用
AS	102	50
SP	11	0
NL	45	7
JL	9	6
PS	17	9
KK	6	7
YG	5	5
IG	78	67
ST	26	35
合計	299	186

数字は講演数: ただし音響モデルは男性のみ

表 2: テストセットの概要

	講演時間		言い淀みの割合	
	時間	単語総数	間投詞	言い直し
AS99SEP022	28分	6305語	9.0%	2.9%
AS99SEP023	30分	4391語	7.5%	2.2%
AS99SEP097	13分	2508語	5.7%	1.1%
PS99SEP025	27分	5372語	11.9%	1.2%

2000年10月時点で音響モデルと言語モデルの学習に利用したデータを表1に示す。言語モデル用には男女両方の講演を含めたが、女性のデータは少ないので音響モデル用には男性のみを用いた。

次にテストセットを表2に示す。話者は4名の男性である。いずれも講演に熟練した話者であり、原稿を用いずに話している。

これらは音響・言語モデルの学習データから除かれ、テスト話者による他の講演も音響モデルの学習から除かれており、完全にオープンである。発表時間と発話総単語数の関係から、およその発話速度を推定することができる。同一の学会の発表にも関わらず、AS99SEP022とAS99SEP023では大きく発話速度が異なる。

3 音響モデル

3.1 音響モデルの作成

音響モデルは混合連続分布HMM（対角共分散）に基づいている。

音声は各講演においてヘッドセットマイクで収録されている。16kHz,16bitでサンプリングされたものをフレーム長25ms（ハミング窓）、フレーム周期10msで分析した。各フレームで12次元のメル周波数ケプストラム係数(MFCC)、その一次差分 (ΔMFCC) とパワーの一次差分 (ΔPower) を計算し、計25次元の特徴量ベクトルを入力として用いた。

音素は43種類でIPAモデル[1]と同様である。各音素は3状態left-to-right(飛び越し遷移なし)HMMでモデル

化した。

また、音素環境を考慮したモデル (triphone 及び PTM) も作成した。その際、すべての音素の 3 つ組に対して十分な学習データはないので、決定木に基づくクラスタリングを行い、状態数 1000, 2000, 3000 のモデルを学習した。PTM(phonetic tied-mixture) モデルは、各状態毎に中心音素が同じであればガウス分布集合を共有し、音素環境に依存してその重みを変えるモデルである [4]。通常の状態共有 triphone は 16 混合で、PTM は 64 混合である。

3.1.1 学習に適したデータの選択と音素ラベルの作成

音素ラベルの作成は、人手による書き起こしに基づいて行った。書き起こしの例を表 3 に示す。"0001 00:00:061-00:05:288 L:" の部分は、"時間ラベル_始まり時間-終わり時間_チャネル:" を示し、この下に当該音声に対応する書き起こしが記述される。この区間の音声を切り出し、対応する音素ラベルを "&" の右側にあるカナ表記から作成する。その際、学習に適さない音声データを除くため、以下に従ってデータを分類する。

1. 非音声データ
   0002 00:01:462-00:01:493 L:＜雑音＞
   言語音以外の場合は "L:" の後に音のラベルがある
   → 音声ではないので学習データから除く
2. 学習対象音声データ
   1. 以外のデータで、発音のカナ表記がカナだけのもの、及び以下のラベルのみが含まれているものは学習に用いる
   言い直し (D), 言い淀み (F), メタ的引用 (M) 発音の転化・なまけ (W), 漢字表記不可 (R), 口語[1] (S)
3. 学習対象外音声データ
   1.,2. 以外、すなわち発音のカナ表記に以下が含まれているもの
   ささやき声 (L), 外国語 (O), 不明 (?),
   不明瞭＜FV＞, 長すぎる母音＜H＞
   ＜笑＞, ＜息＞, ＜咳＞, ＜泣＞
   → 曖昧であったり日本語でなかったりするので学習データから除く

本研究では、2. の学習対象データのみを用いて音響モデルの学習を行う。

3.1.2 学習セット

テストセットは表 2 に示す通り、主に日本音響学会における講演 (AS) である。そこで学習セットとして
(set-1a) AS のみ
(set-2a) AS+他の学会講演 (JL+KK+NL+PS+SP+YG)
(set-3a) AS+他の学会講演+模擬講演 (IG+ST)
の 3 種類を選んだ（表 4)。

[1] 口語に頻出するくだけた表現、すなわち発音の転訛に関わる表現

表 3: 書き起こし例

0001 00:00:061-00:05:288 L:		
(F えー) 発表内容	&	(F エー) ハッピョーナイヨー
(F あー) まず	&	(F アー) マズ
0002 00:01:462-00:01:493 L:＜雑音＞		
0004 00:05:550-00:08:368 L:		
その後	&	ソノゴ
従来の	&	ジューライノ
...		
0006 00:08:767-00:09:320 L:		
それから	&	ソレカラ

表 4: 音響モデルの学習セット

	使用講演（講演数）	データ量
set-1a	AS(102)	13.2 時間
set-2a	set-1a + SP(11) + NL(45)+ JL(9) + PS(17) + KK(6) + YG(5)	35.3 時間
set-3a	set-2a + IG(78) + ST(26)	47.3 時間
IPA	JNAS コーパス	約 40 時間

set-1a はテストセットと講演内容がほぼ一致しているが、データ量は少ない。set-2a は学会講演というスタイルにおいて一致しており、学習量も比較的多い。set-3a はやや発話スタイルが異なる模擬講演を加えているが、学習量は多くなっている。

これらにより、学習データ量は増えるが学習データとテストデータの一致度が低くなることの効果を見ることができる。

3.2 音響モデルの評価

それぞれの学習セットで種々の音響モデルを作成し、認識実験を行った。言語モデルは、融合研究コーパス（全講演）と Web 講演録全てから学習し、予稿テキストによる話題適応は行っていない（4 章参照)。比較のため、読上げ音声で学習された IPA モデル [1] でも評価を行った。デコーダは Julius-3.1[5] である。

表 5, 6 にテストセットに対する単語正解精度を示す。set-1a の triphone や PTM の性能から、set-1a は学習量が不足していることが分かる。また、PTM はパラメータ数が少ない分、少ない学習量でもよい性能を示している。さらに、講演 13 時間 (set-1a) で学習したモデルでも読上げ音声 40 時間で学習したモデルより性能がよく、認識タスクと学習コーパスのスタイルを一致させる効果も確認できた。

set-2a で学習したモデルは set-1a に比べ全体的に性能がよく、他の講演を加えてデータ量を増やすことは有効であった。また、set-2a のそれぞれの triphone を比較すると、状態数が多い triphone の学習には、依然データが不足していることがわかる。

また、set-2a と set-3a から学習したモデルには差がほ

表 5: AS99SEP022 に対する単語正解精度 (%)

モデル	学習セット			
	set-1a	set-2a	set-3a	IPA
monophone 129x32	43.5	44.9	44.3	-
monophone 129x64	45.7	45.9	46.2	-
triphone 1000x16	51.3	54.6	54.2	41.7
triphone 2000x16	50.4	55.6	54.1	42.0
triphone 3000x16	49.5	54.6	54.3	41.9
PTM 129x64(s1000)	51.1	53.4	52.9	-
PTM 129x64(s2000)	51.2	53.9	53.4	-
PTM 129x64(s3000)	51.5	53.7	53.8	41.9

表 6: AS99SEP023 に対する単語正解精度 (%)

モデル	学習セット			
	set-1a	set-2a	set-3a	IPA
monophone 129x32	55.4	56.8	56.4	-
monophone 129x64	57.3	58.3	57.3	-
triphone 1000x16	54.7	59.9	60.4	51.5
triphone 2000x16	51.7	59.3	58.8	49.9
triphone 3000x16	48.6	55.0	55.3	48.6
PTM 129x64(s1000)	58.9	62.7	62.7	-
PTM 129x64(s2000)	60.1	62.9	62.2	-
PTM 129x64(s3000)	59.1	62.2	61.0	54.5

とんどない。この理由には、スタイルが異なる模擬講演を加えても効果がない、もしくは既に十分な学習量が得られたということが考えられるが、triphone の性能をみる限り、十分な学習量が得られているとは考えられない。従って、模擬講演を加える効果は少ないといえる。これは、[3] でも分析が行われているが、発声の状況が異なるための心理的影響に起因すると考えられる。模擬講演は比較的カジュアルな内容をゆっくり語りかけるようなものが多く、学会講演とはかなり異なる。

4 言語モデル

4.1 言語モデルの学習

言語モデル（単語 N-gram モデル）の学習には、表 1 で示した合計 186 講演のチェック済の書き起こしテキスト (trans4) を使用した。表 3 に示す書き起こしの"&" の左側に書かれる見出し語に形態素解析を施し、この形態素を単語として学習を行った。形態素解析には Chasen ver2.02 を用いた。書き起こしテキストに記される言い淀み (F) は、全て学習テキストに含めるが、言い直し (D) は出現の特徴が容易にモデル化できないので削除した。口語表現 (S) や書き起こしに自信がない部分 (?) など、

表 8: 言語モデルの学習に用いるテキストセット

	融合研究コーパス			Web
	set-1b AS のみ	set-2b AS+ 他学会	set-3b 全講演	講演録
講演数	49 種類	84 種類	186 種類	81 種類
単語総数	104401	271348	466838	1696875
語彙数	4563	10276	17176	37592

特殊なラベルをもつものは、文を理解するのに必要な単語のみ使用した。

これらの学習テキストは、日本音響学会等の学会での口頭発表を書き起こしたものや関心あるテーマに基づく模擬講演からなり、音響モデルと同様に以下の 3 つのセットに分類し、それぞれのモデルを作成した。

(set-1b) 音響学会の講演 (AS) のみ

(set-2b) AS+他の講演 (JL+KK+NL+PS+YG)

(set-3b) AS+他の講演+模擬講演 (IG+ST)

少量であるがテストセットと内容がほぼ一致する set-1b、分野が近い他の学会講演を加えた set-2b、種々の話題からなる模擬講演をさらに加えた set-3b という構成になっている。

また融合研究コーパスとは別に World Wide Web 上で公開されている講演録を収集しこれらを用いて、より効果的なモデル化を試みた。これらの統計量を表 8 に示す。Web 講演録のテキストサイズは融合研究コーパスの 4 倍近くであり、学習量の強化を見込める。融合研究コーパスには言い淀みである間投詞が忠実に書き起こされ、発声中の無音部分も時刻ラベルに基づいて判定することができるが、言語的な文や意味的まとまりを決定づける句読点が存在しない。一方、Web 講演録は編集によって間投詞などの言い淀みは削除され、整形された文になっているが、句読点は書き起こしに挿入されている。講演音声の意味的まとまりを考える際にも認識結果の自動整形を考える際にも句読点の役割は重要になると考えられる。

音声を認識する際に、適切な長さに区分化する必要がある。本研究では書き起こしテキストに含まれる時間ラベルの情報は用いないで、テストセットの音声を無音により適当な長さにしたものを用いた。したがって、用いる音声は人手で分割したものとは異なり、一つの文単位や意味的なまとまりとは限らない。書き起こしに存在する時間ラベルに基づき、一定の無音が存在する区間を、仮説の始端<s>と終端</s>として言語モデルの学習に用いた。無音区間の長さを変化させ実験を試みた結果、500msec 以上の無音という設定が最適であった。テストセットの講演音声はおおむね 500msec の無音で分割しており、このモデル化は妥当である。

表 7: 種々の言語モデルによるカバレージとパープレキシティ

	融合研究コーパス				Web 講演録	融合研究 +Web
	set-1b	set-2b	set-3b			
	cut-off 0 4563 語	cut-off 1 6272 語	cut-off 1 10349 語	AS+話題独立 7140 語	話題独立 8K (+間投詞)	AS+話題独立 10771 語
AS99SEP022	91.6%, 176.2	93.4%, 126.3	94.4%, 126.5	94.0%, 121.3	91.4%, 225.2	95.0% 143.5
AS99SEP023	94.6%, 120.4	95.3%, 109.4	95.7%, 115.1	95.7%, 115.9	87.7%, 233.3	96.1% 144.7
AS99SEP097	94.1%, 112.9	95.3%, 108.5	95.3%, 122.5	95.1%, 115.5	89.3%, 156.5	95.9% 141.7
PS99SEP025	94.1%, 137.2	95.2%, 143.2	95.7%, 162.7	95.2%, 156.1	86.0%, 302.3	95.9% 205.5

各項目: (カバレージ, パープレキシティ)

4.2 言語モデルの評価

4.2.1 カバレージとパープレキシティ

融合研究コーパスの各セットによる語彙と Web 講演録から抽出した語彙を用いて、テストセットの講演に対するカバレージとパープレキシティを算出した結果を表 7 に示す。テストセットには言い淀み・言い直し等すべて含まれる。CMU_ToolKit ver.2[6] を用いて tri-gram 言語モデルを構築し単語パープレキシティを算出した。

融合研究コーパスの語彙は、各セットにおいて出現頻度上位単語で制限する典型的な手法で構成した。しかし、特に set-3b に含まれる模擬講演 (IG, ST) などには、学会講演に通常現れない単語が多く含まれる。そこでカバレージと学習量を維持しつつ語彙サイズを削減するために、テストセットに最も近い set-1b の AS の語彙を cut-off 0 とし、残りの他の講演について、話題と出現単語の相互情報量 [7] に基づいて、いずれの講演でも出現する単語を抽出した語彙 [AS+話題独立] を構成した。

Web 講演録には、学会の内容とは異なる政治や医療という様々な話題の講演があるが、これらを用いて講演音声に普遍的な言語モデルを構築した [8]。このモデルは、講演の話題と単語の相互情報量に基づき特定の話題に顕著に出現する専門用語を削除することで、すべての講演のベースとなるモデルである。この話題独立の語彙 8000 語にフィラー単語を加えたものを用いた。

表 7 に示す通り、テストセットに最も合致している set-1b の語彙のみによるカバレージは他と比べて低い。これは学習データ量の不足によって語彙サイズが小さいためであり、set-2b のように他の学会講演を含めることで改善された。また模擬講演を含めた set-3b の語彙ではカバレージは若干だけ向上した。

パープレキシティは語彙サイズに依存するために一概に比較することは困難であるが、同一語彙では一貫して話者 PS99SEP025 が高く、認識が困難であることが予測できる。また融合研究コーパスと比較して、Web 講演録の話題独立モデルはパープレキシティが高い。間投詞のモデル化を含めて、忠実な書き起こしテキストでないことの影響であると考えられる。

4.2.2 認識実験

4 名のテストセットに対して、これらの言語モデルを用いて認識実験を行った。音響モデルには、3 章で述べた融合研究コーパス学習セット set-2a の PTM 64mix(s2000) を用いた。デコーダは Julius-3.1 である。

なおショートポーズはコンテキストとして重要な役割を担っていないと考えられるので透過単語として扱い、単語履歴からは除外することで言語モデルの実効性を損なわないようにした。そのため認識率が表 5, 6 の結果と異なる。融合研究コーパスを学習テキストとして用いたモデルでは、ショートポーズを uni-gram 確率で与えた。実験結果を表 9 に示す。

表 2 に示すように、話者により発話中の間投詞や言い直しの割合は大きく異なり、これらが認識に影響を及ぼす。AS99SEP022 は発話速度が速く、間投詞と言い直しの割合も大きい。また PS99SEP025 は間投詞の割合が最も大きく、咳や講演の内容外の発話も多く含まれるため、これらの話者は他と比べて認識が困難である。

表 9 に示す通り、融合研究コーパスを学習テキストとしたモデルのうち、AS 以外の講演の語彙を相互情報量により選定したものは、cut-off 1 のモデルに比べて小さい語彙サイズで同等の認識精度が得られた。Web 講演録による話題独立モデルは、融合研究コーパスの結果を大きく下回った。これは講演の話題に関連する専門用語をカバーしていないことや、間投詞のモデル化が厳密でないことに起因する。融合研究コーパスと Web 講演録の話題独立モデルを混合したモデルでは、認識精度の改善が見られる一方、低下した話者もあった。今回試験的に、融合研究コーパスと Web 話題独立モデルをテキスト混合比 1:1 で併合したが、Web 講演録が 3 倍以上の学習テキストサイズを持つため、双方の混合比を適切にとる必要がある。また表 5, 6 との比較から、ショートポーズを透過単語として扱うことが一定の効果を収めた。

4 名の平均の認識精度が最もよかったのは、set-3b の AS+話題独立語彙を用いた場合で、59.8%であった。

表 9: 種々の言語モデルによる単語認識精度 (%)

	融合研究コーパス set-3b		Web 講演録	融合研究 +Web	融合研究 +Web+予稿
	cut-off 1 10349 語	AS+話題独立 7140 語	話題独立 8K (+間投詞)	AS+話題独立 10771 語	AS+話題独立 10771+α語
AS99SEP022	53.1	53.6	51.7	55.0	55.5
AS99SEP023	66.5	66.3	49.6	65.5	68.5
AS99SEP097	67.2	66.9	60.8	66.1	なし
PS99SEP025	58.4	58.4	46.8	56.6	なし

4.2.3 予稿テキスト利用の効果

講演には、予稿テキストが用意されることが多い。話者も予稿に沿って講演を進めることが多く、予稿はその講演の話題に関連する語を多く含むので、これを使うことでより当該講演に適したモデルが構築できる。

ここでテストセットの 2 名のそれぞれに対して、予稿テキストを用いて話題適応を行った。前述の融合研究コーパスと Web 講演録を併合したものに、さらに予稿を加えて構築したモデルを用いて認識実験を行った。予稿テキストの混合重みは、あらかじめパープレキシティにより最適な値を用いた。結果を表 9 に示す。いずれも認識精度が向上し、特に AS99SEP023 は専門用語と言い回しをカバーすることで大きく認識精度が向上した。

5 まとめ

講演データベースを用いて、講演音声の認識に適した音響・言語モデルの構築を試みた。

その結果、音響モデルにおいては話し言葉の学習コーパスを用いることが有効であり、特に学会講演というスタイルに一致させることの効果を確認した。また言語モデルにおいても同一スタイルのテキストを用いることが有効であり、ショートポーズの透過単語化や予稿テキストの利用の効果が確認された。

今後も学習データ量を増加させてさらに精度のよいモデルを構築するとともに、音響的には発話の速度やなまけに対応可能なモデル化を試み、また言語的には間投詞の精密なモデル化や他のコーパスとの効果的な混合を行う予定である。

謝辞 本研究は、開放的融合研究『話し言葉工学』プロジェクトの一環として行われた。アドバイスを頂きました東京工業大学の古井貞熙教授をはじめとして、ご協力を頂いた関係各位に感謝する。また、講演予稿テキストを提供して頂いた 2 名の先生方に感謝する。

参考文献

[1] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏. 日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価. 情報処理学会研究報告, SLP-31-2, NL-137-7, 2000.

[2] 峯松信明, 片岡嘉孝, 中川聖一. 講演調の話し言葉に対する言語的解析. 情報処理学会研究報告, 95-SLP-8-7, 1995.

[3] 龍宮隆之, 菊池英明, 小磯花絵, 前川喜久雄. 大規模話し言葉コーパスにおける発話スタイルの諸相 −書き起こしテキストの分析から−. 日本音響学会研究発表会講演論文集, 2-Q-9, 秋季 2000.

[4] A.Lee, T.Kawahara, K.Takeda, and K.Shikano. A new phonetic tied-mixture model for efficient decoding. In *Proc. IEEE-ICASSP*, pp. 1269–1272, 2000.

[5] 李晃伸, 河原達也, 堂下修司. 単語トレリスインデックスを用いた段階的探索による 大語彙連続音声認識. 電子情報通信学会論文誌, Vol. J82-DII, No. 1, pp. 1–9, 1999.

[6] P.R. Clarkson and R. Rosenfeld. Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings ESCA Eurospeech*, 1997.

[7] T.Kawahara and S.Doshita. Topic independent language model for key-phrase detection and verification. In *Proc. IEEE Int'l Conf. Acoust. Speech & Signal Process.*, pp. 685–688, 1999.

[8] K.Kato, H.Nanjo, and T.Kawahara. Automatic transcription of lecture speech using topic-independent language modeling. In *Proc. ICSLP*, Vol. 1, pp. 162–165, 2000.