TR-IT-0164

音声認識における統計的言語モデルの選択使用の効果
Continuous Speech Recognition Using a Dialog-Conditioned Stochastic Language Model

坂本 博之
Hiroyuki Sakamoto

松永 昭一
Shoichi Matsunaga

1996.3.28

音声認識における統計的言語モデルの選択使用の効果について検討した．ここでは，音声対話システムにおいて「ユーザとシステムとの対話」という状況を想定し，ユーザ発話の音声認識において直前のシステム側の発話内容に基づき，対話の状況に応じてユーザの次発話の予測を行い，予測結果に基づいて統計的言語モデルの選択を行った．用いる統計的言語モデルは，同じくシステム側の発話内容に基づいてテキストデータの分類を行い作成した．この言語モデルの選択使用の効果を音節パープレキシティの比較および文節認識実験によって評価した．この結果，分類した学習テキストから作成された言語モデルには明らかな言語情報の偏りが確認でき，次発話予測による統計的言語モデルの選択使用は音声認識の性能の向上を期待できることが分かった．

1 はじめに

人間のコミュニケーション手段の一つに音声による対話がある．音声認識の分野においても「音声対話システムにおけるユーザとシステムとの対話」という状況を想定し，さまざまな研究が進められている [1][2]．また対話は，同一タスクの中でもさまざまな発話内容 (場面) の推移が行われ，場面毎に異る言語情報が含まれていると考えられる．これまでに，音声対話システムにおいて対話の話題ならびにシステム側の質問の内容・型などの情報を利用して，ユーザの次発話に対し構文・単語予測を行う方法が提案されている [3]．しかし音声対話システムでより自由な発話を扱おうとすると，さまざまな言い回しが発生してしまい，構文の記述が非常に複雑になり学習が困難であるという問題がある．そこで本稿では，音声対話系を想定した大量のテキストデータから比較的容易に学習の可能な統計的言語モデル (N-gram) を使用した方法を検討する．この N-gram は，評価データと同じタスクの大量のテキストデータを用いて作成することで音声認識において高い認識性能を達成することが知られており [4][5][6]，一般には全テキストデータから 1 個の N-gram を学習し使用される．本稿では「音声対話システムにおけるユーザとシステムとの対話」という想定の基で，直前のシステム側の発話内容を基にユーザの次発話を予測するという形式を用いて分類されたテキストから統計的言語モデルを学習し，認識時には直前のシステム側の発話内容を基にこれらのモデルの中から選択を行い使用する方法について提案する．これは対話の状況においては単語や言いまわしの偏りがあると考えられ，これを考慮した言語モデルが音声認識に有効ではないかと考えたためである．具体的には，評価データと同一タスクの対話テキストデータを用いて，1 会話 (会話開始から，対話のやりとり，会話終了まで) の中からユーザ次発話の予測可能な数種類の場面を設定し，各場面毎に分類したテキストから音節 N-gram を作成する．そして，これらの N-gram の中から評価データのシステム側の発話を基にした予測に従い N-gram を選択して使用した．本稿では言語モデルを選択使用した場合と従来の単一のモデルを使用した場合の効果についてパープレキシティおよび音声認識実験を通して検証する．

2 次発話予測に基づく場面設定

提案法は，場面に応じて作成した統計的言語モデルをユーザの次発話の予測に基づいて選択使用するものである．この提案法を想定した音声対話システムの概念図を図 0.1 に示す [7]．こ

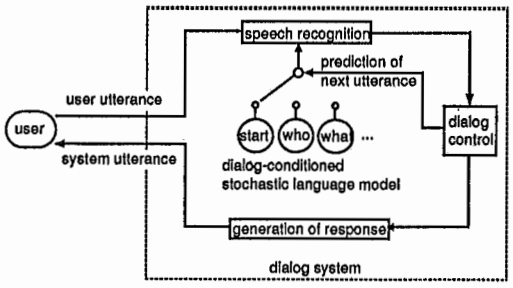

図 0.1: 統計的言語モデルを選択使用する音声対話システム

のシステムでは，"dialog control" 部でシステムの応答生成とユーザの次発話の予測を行い，"speech recognition" 部で予測に基づいて統計的言語モデルを選択し認識を行い，この認識結果を基にさらに次の新たなシステムの応答生成を行うことを想定している．

まず，ユーザの次発話を予測し得る場面を以下の様に設定する．ここでは，国際会議予約というタスクにおける参加者と事務局の対話テキストデータを基に事務局の発話内容から参加者の次発話の予測の可能な場面を調べた．（ユーザを参加者，事務局をシステムと仮定．）この結果から参加者側の全発話 (約 4.7×10^4 文節) を学習テキストとして，対話単位 (発話開始から発話権移動まで) で以下に示す場面を決め，テキストデータの分類を行った．

- 対話開始場面 (start):
  参加者が挨拶や用件導入を発話することを予測
- 事務局が名前を尋ねた場面 (who):
  参加者が名前について発話することを予測
- 事務局が用件を尋ねた場面 (what):
  参加者が用件について発話することを予測
- 事務局が可否を尋ねた場面 (yes-no):
  参加者が「はい」「いいえ」を発話することを予測
- 事務局が日時を尋ねた場面 (when):
  参加者が日時について発話することを予測
- 事務局が場所を尋ねた場面 (where):
  参加者が場所について発話することを予測

また，次発話予測のできない場面 (other) として上記の分類に属さない参加者の発話についても一つの分類として扱うこととした．以下で行う評価では，サンプル数が少ない分類については分類を行わず other に含めて扱うこととした．

3 統計的言語モデル

言語モデルには，統計的言語モデルである N-gram を使用した．N-gram の学習は，分類を行ったそれぞれのテキストデータ (107 音節 + 無音で記述) を用いて音節の unigram, bigram, trigram を計算し，削除補間法 (deleted interpolation[8]) により平滑化を行った音節 trigram($\tilde{P}(w_n|$ を学習し使用することとする (式 (0.1))．

$$\begin{aligned}\tilde{P}(w_n|w_{n-2},w_{n-1}) &= \lambda_0 P_0 + \lambda_1 P(w_n) \\ &+\lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n|w_{n-2},w_{n-1})\end{aligned} \tag{0.1}$$

ここで，w は音節を表し，$P(w_n|w_{n-2},w_{n-1})$ は音節連鎖 w_{n-2},w_{n-1} の後に w_n の生じる確率，λ は $\sum_{i=0}^{3}\lambda_i = 1$ となる重みである．

3.1 統計的言語モデルの作成

我々は，比較のため全学習テキストに対して以下 4 種類の分類を行ったテキストを使ってそれぞれ統計的言語モデルを作成した．

(a) 場面毎に分類したテキスト
次発話予測に基づいて設定した場面毎に学習テキストを分類し，それぞれの場面分類テキストから統計的言語モデルを作成．

(b) (a) と同じ数の文節を場面毎に任意に選択したテキスト
上記モデル (a) との比較のため，場面毎に分類されたテキストと同文節数を全学習テキストから任意に選び，それぞれの場面毎に統計的言語モデルを作成.

(c) 全学習テキスト
全学習テキストにより単一の統計的言語モデルを作成. このモデルを使用する方法が，従来は最も一般的である.

(d) 場面毎に，場面分類したテキストとその残りのテキスト
場面分類されたテキストから作成する場面毎の言語モデルは，全学習テキストの一部のみしか使用しておらずデータ量も極めて少ない. そこで，より信頼性の高い頑健な言語モデルの作成を目指し，場面毎に作成した言語モデル $\tilde{P}$ と，$\tilde{P}$ の学習に使用したデータ以外の残りのテキストからさらに言語モデル $\tilde{P}_{rest}$ を作成し，式 (0.2) で表す統計的言語モデル P^* を作成 [9].

$$P^* = \mu\tilde{P} + (1-\mu)\tilde{P}_{rest} \tag{0.2}$$

ここで μ は重みを表し，場面分類したテキストを用いて削除補間法により求める.

これら 4 種類の統計的言語モデルについて，比較検討を行う.

3.2 統計的言語モデルのパープレキシティの比較

ここで作成された統計的言語モデル (音節 trigram) を対象に言語情報の偏りを確認するため，パープレキシティの比較を行った. これらの統計的言語モデルを用いて評価用テキスト (321 文節) のパープレキシティを計算した結果を表 0.1に示す. 表中 () 内は文節数， start,who,$\cdots$ は (a) の場面分類したテキストから作成された音節 trigram， random は (b) の任意選択テキストから作成された音節 trigram， start*,who*,$\cdots$ は (d) の式 (0.2) により作成した音節 trigram， all は (c) の全学習テキストから作成した音節 trigram を表す.

表 0.1から， 321 文節の評価テキストで (b) のパープレキシティが 8.43 であるのに対して，(a) のパープレキシティは 6.27 である. この結果から， (a) の場面に応じて分類したテキストは言語情報の偏りを捉えていることがわかる. また (a) のパープレキシティは，一般的に用いられる (c) の全学習テキストから学習された単一モデルのパープレキシティ (6.90) よりも小さい. さらに， (d) の式 (0.2) により作成した音節 trigram P^* はパープレキシティが他のモデルに比べて低下していることから，より信頼性の高いモデルであると考えられる.

同様に旅行案内のタスク (学習テキスト約 7.1×10^3 文節) においてもパープレキシティを計算した. この結果を表 0.2に示す.

表 0.2から，他のタスクにおいても場面に応じて分類したテキストの言語情報に偏りがあることが確認できる. しかし， (a) のパープレキシティ (12.2) は， (c) の単一モデルのパープレキシティ (10.8) よりも大きい. これは，学習に使ったテキストデータの量が不足していたためと考えられる. 但し， (d) の式 (0.2) により作成した音節 trigram P^* のパープレキシティは (c) のモデルのパープレキシティよりもさらに小さく，データ量が不足する場合に有効なモデルと考えられる.

3.3 予測次発話の直前のシステム発話の利用の検討

次に我々は，予測するユーザ次発話は直前のシステム発話の影響を受け，同じ単語や類似した言い回しなどが出現する可能性が高いと考えた. 例えば，システム発話が「お名前をお願

表 0.1: (a) のテキストから作成した音節 trigram と他の分類の音節 trigram によるパープレキシティ (国際会議予約)

学習テキスト		評価テキスト	perplexity
start	(4110)	start	5.3
random	(4110)	(74)	5.9
start*			4.7
who	(460)	who	9.1
random	(460)	(61)	33
who*			10
what	(4377)	what	7.9
random	(4377)	(26)	8.2
what*			7.0
yes-no	(5121)	yes-no	7.6
random	(5121)	(54)	8.2
yes-no*			6.9
other	(32515)	other	5.04
random	(32515)	(106)	5.36
other*			5.14
all	(46583)		6.90
average	P	all	6.27
	random	(321)	8.43
	P*		6.02

いします」の時，ユーザ次発話は「名前は鈴木です」の様に「名前」というシステム発話内に含まれる同じ単語を持った発話をする場合がある．そこで，予測するユーザ発話の直前のシステム発話の利用を考える．旅行案内のタスク（客をユーザ，受付をシステムと仮定）に対して対話単位 (発話開始から発話権移動まで) で，客の発話に対して (a) の分類を施したテキストを用意し，このテキストにさらに評価対象である客発話の直前の受付側の 1 対話を追加したテキストを使って音節 trigram を作成する．しかしここでは，客発話の分類テキストに対して受付側の 1 対話のテキストは非常に小さいためこのテキストは多重に追加することにした．この時のパープレキシティを表 0.3に示す．表中 [ ] 内は，パープレキシティが最も小さくなった受付側テキストの追加回数を表す．

表 0.3では，when の分類に関してはパープレキシティの低下が見られるが，全体にパープレキシティの低下は非常に小さい．これは，利用するシステム発話のテキストが 1 対話でありデータ量も小さく，N-gram モデルへの影響が非常に少ないためと考えられる．この直前の発話の利用は，さらに効果的な利用方法について検討する必要がある．

4 統計的言語モデルの選択による文節認識実験

実験には，逐次状態分割法 (SSS) により作成された HMnet と音素コンテキスト依存 LR パーザを統合した SSS-LR 連続音声認識システム [10] を使用した．ここで使用する HMnet には移動ベクトル場平滑化方式 (VFS) に基づく話者適応 [11] を施したものを，統計的言語モデルには前述の 4 種類の音節 trigram をそれぞれ使用し比較する．

実験条件を表 0.4に示す．

表 0.2: (a) のテキストから作成した音節 trigram と他の分類の音節 trigram によるパープレキシティ (旅行案内)

学習テキスト		評価テキスト	perplexity
start	(306)	start	9.6
random	(306)	(67)	18
start*			6.9
who	(424)	who	14
random	(424)	(53)	40
who*			14
when	(333)	when	12
random	(333)	(32)	40
when*			11
where	(366)	where	19
random	(366)	(26)	27
where*			14
all	(7091)		10.8
average	$\tilde{P}$	all	12.2
	random	(1019)	15.5
	P^*		10.3

SSS-LR のスコア計算は以下の式 (0.3) により行う.

$$P_s = (1-\theta) * log(P_h) + \theta * log(P_l) \tag{0.3}$$

ここで, P_s は SSS-LR の認識スコア, P_h は HMnet の音響尤度, P_l は音節 trigram の確率, θ は音節 trigram に対する重みである. θ は, ここでは実験的に求めた値を用いる.

4.1 文脈自由文法を用いた文節認識

国際会議予約の文脈自由文法 (語彙数 : 1321, ルール数 : 2813) を用いた文節認識に, 前述の 4 種類の分類により作成した統計的言語モデルをそれぞれ使用した場合の結果を表 0.5に示す. 表中 () 内は文節数, (a) は場面分類したテキストから作成された音節 trigram, (b) は任

表 0.3: 予測次発話の直前のシステム発話利用時のパープレキシティ (旅行案内)

評価テキスト	学習テキスト	
start	start	start+[2]
	9.55	9.54
who	who	who+[1]
	13.7	13.6
when	when	when+[4]
	12.1	11.3
where	where	where+[2]
	19.4	19.3

表 0.4: 文節認識の実験条件

分析条件	
サンプリング周波数	12kHz
フレーム周期	5ms
ハミング窓	20ms
分析	log power + 16 次 LPC-Cep + Δlog power + 16 次 ΔLPC-Cep
学習条件	
話者	男性 1 名 (MHT)
学習データ	重要語 2620 単語
音響モデル	コンテキスト依存型音素 HMM (1 混合 600 状態の HMnet)
話者適応条件	
話者	男性 1 名 (MTK)
適応データ	25 音素バランス単語
話者適応	移動ベクトル場平滑化方式 (VFS)
認識条件	
話者	男性 1 名 (MTK)
認識データ	国際会議予約の 321 の文節発声
使用 N-gram	音節 trigram
N-gram 学習データ	参加者発話 4.7×10^4 文節 (シンボル数:107 音節 + 無音)
ビーム幅	200

意に選択したテキストから作成された音節 trigram, (d) は式 (0.2) により作成した音節 trigram, (c) は全学習テキストから作成した音節 trigram を表し, top 1 は認識結果上位 1 位の文節認識率, top 5 は上位 5 位までの文節認識率を示す. また, #other は予測不可能分類であるため (c) のモデルを使用している.

文節認識率は, (a) のモデルで 82.2%, (b) のモデルで 80.4%, (c) のモデルで 81.0% , (d) のモデルで 81.3% でありその差は極めて小さい. 統計的言語モデルを用いない ($\theta = 0$ に相当) 場合の文節認識率が 76.0% (top5: 86.0%) であることから, 提案する選択的な言語モデルの使用による飛躍的な認識率の上昇は望めないものの, 本方法が有効であることが分かった. 一方, (a) のモデルが (d) のモデルよりも高い認識率であるが, 今後さらに学習データおよび評価データを増やして (d) のモデルを評価する必要があると思われる.

4.2 文法を用いない文節認識 (音節タイプライタ)

次に, 文脈自由文法の制約なしに自由な音節連鎖を許す音節タイプライタを駆動して, 同様に前述の 4 種類の分類により作成した統計的言語モデルをそれぞれ使用して評価を行った. これは 4.1の実験では文法や単語辞書の影響があるのでこれを除いた N-gram モデルの効果のみを観察するために行うものである. この結果を表 0.6に示す.

文節認識率は, (a) のモデルで 58.6%, (b) のモデルで 49.5%, (c) のモデルで 56.4% , (d) のモデルで 55.5% である. この実験においても, 提案する統計的言語モデルの選択使用は他のモデルと比較して最も高い認識率を達成した.

表 0.5: 文脈自由文法を使った文節認識率 (%) (θ=0.5)

順位	評価文節	統計的言語モデル		
		(a)	(b)	(d)
top 1	start	74	70	73
top 5	(74)	85	82	84
top 1	who	85	79	80
top 5	(61)	89	87	85
top 1	what	73	73	73
top 5	(26)	81	77	77
top 1	yes/no	87	89	89
top 5	(54)	91	91	91
top 1	#other	85.9		
top 5	(106)	92.5		
top 1	average	82.2	80.4	81.3
top 5	(321)	88.8	87.5	87.5

順位	評価文節	(c)
top 1	all	81.0
top 5	(321)	87.2

5　むすび

本稿では，音声認識における統計的言語モデルの選択使用の効果について検討した．「音声対話システムにおけるユーザとシステムとの対話」という状況を想定することにより，直前のシステム側の発話内容を基にユーザの次発話を予測し分類するという形式を用いて統計的言語モデルを作成し，このモデルを選択使用する方法を用いた．そして，その効果をパープレキシティの値およびSSS-LR連続音声認識システムで実際に文節認識実験を行い認識率により評価した．この結果，分類した学習テキストから作成された言語モデルには言語情報の偏りが確認でき，さらにこの言語モデルを場面に応じて用いることで認識率の向上を見ることができた．このことから，次発話予測による統計的言語モデルの選択使用は音声認識の性能向上に効果があることが期待できる．しかし今回認識実験に用いた方法では，言語モデルが音節の trigram であるため今後データ量を増やして単語 N-gram を用いるなどしてモデルを強化する必要がある．また，3.1(d) で述べた分類し作成した言語モデルに対して残りのテキストを使い補間を施す方法，他の分類の言語モデルすべてを使用してそれぞれに対する重みを変える方法などを検討し，より柔軟で効率の良い言語モデルの作成を目指す．

謝辞

研究の機会を与えて頂いたATR音声翻訳通信研究所 山崎泰弘社長，匂坂芳典室長に感謝いたします．また，熱心な御討論と有益な御助言を頂いた研究室の方々に感謝いたします．

表 0.6: 音節タイプライタによる文節認識率 (%) (θ=0.8)

順位	評価文節	統計的言語モデル		
		(a)	(b)	(d)
top 1	start	55	51	58
top 5	(74)	70	62	74
top 1	who	52	18	49
top 5	(61)	57	23	57
top 1	what	38	31	38
top 5	(26)	50	42	54
top 1	yes/no	63	57	50
top 5	(54)	74	67	78
top 1	#other	67.0		
top 5	(106)	79.3		
top 1	average	58.6	49.5	56.4
top 5	(321)	69.8	59.5	71.7

順位	評価文節	(c)
top 1	all	55.5
top 5	(321)	69.8

参考文献

[1] V. W. Zue, J. Glass, D. Goddeau, D. Goodine, H. Leung, M. McCandless, M. Phillips, J. Polifroni, S. Seneff and D. Whitney, "Recent Progress on the MIT VOYAGER Spoken Language System," Proc. ICSLP-90, 29.6.1, pp.1317-1320, 1990.

[2] 白井克彦, "音声対話のモデル化とその機械処理に関する総合的研究", 総合研究 A 0230510, 研究成果報告書, 1992.

[3] 森屋裕治, 阿部野尚, 山本幹雄, 中川聖一, "対話予測を利用した観光案内対話システム", 信学技報, SP92-121, pp.43-50, 1993.

[4] Kai-Fu Lee, "Large-Vocabulary Speaker Independent Continuous Speech Recognition: The SPHINX System," 15213 CMU-CS-88-148, 1988.

[5] A. Averbuch, et al., "An IBM PC Based Large-Vocabulary Isolated-Utterance Speech Recognizer," Proc. ICASSP'86, Vol.1, 2.4.1 pp.53-56, 1986.

[6] T. Araki, J. Murakami and S. Ikehara, "Post-processing in Japanese Speech Recognition Using 2nd-order Markov Model of Syllables," NTT REVIEW, vol.1, no.3, pp.96-104, 1989.

[7] H. Sakamoto and S. Matsunaga, "Continuous Speech Recognition Using a Dialog-Conditioned Stochastic Language Model", Proc. ICSLP94, S16-9.1, pp.811-814, 1994.

[8] F.Jelinek, "The development of an experimental discrete dictation recognizer," Proc. IEEE, vol. 73, pp.1616-1624, 1985.

[9] 松永昭一, 山田智一, 鹿野清宏, "音節連鎖統計情報のタスク適応化", 第 42 情処全大, 6D-5, pp.114-115, 1991.

[10] 永井明人, 鷹見淳一, 嵯峨山茂樹, "逐次状態分割法 (SSS) と音素コンテキスト依存 LR パーザを統合した SSS-LR 連続音声認識システム", 信学技報, SP92-33, 1992.

[11] 大倉計美, 杉山雅英, 嵯峨山茂樹, "混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式", 信学技報, SP92-16, pp.23-28, 1992.