論文 / 著書情報
Article / Book Information

論題(和文)	高次n-gramを用いた大語彙連続音声認識の検討
Title(English)	A Study of Large-Vocabulary Continuous Speech Recognition Using Higher Order n-gram Language Models
著者(和文)	大附克年, 吉田航太郎, 松岡達雄, 古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会 1997年春季講演論文集, Vol. , No. 2-6-2, pp. 47-48
Citation(English)	, Vol. , No. 2-6-2, pp. 47-48
発行日 / Pub. date	1997, 3

2—6—2　高次n-gramを用いた大語彙連続音声認識の検討*

◎大附克年[1]　吉田航太郎[2]　松岡達雄[1]　古井貞熙[1,2]
([1]NTTヒューマンインタフェース研究所　[2]東京工業大学)

1. はじめに

単語系列の確率に基づくn-gram言語モデルは，大量の学習データがあれば比較的容易に推定することができるため，英語をはじめとする各言語の大語彙連続音声認識においてよく用いられている．我々は，新聞記事読み上げ音声に対して単語bigramを適用し，日本語の大語彙連続音声認識に対しても統計的言語モデルが非常に有効であることを示した[1,2]．現在，より精密な言語モデルとして，trigram，4-gramといった高次のn-gramの大語彙連続音声認識への導入を検討している[3]．本稿では，高次n-gram言語モデル，およびそれらを大語彙連続音声認識に用いるためのマルチパス探索アプローチについて述べる．

2. 高次n-gram言語モデル

約5年分の新聞記事データベース[9]を用いて単語n-gram(n=1 ～ 4)の推定を行った．日本語のテキストは単語ごとに分かち書きがされていないため，テキストに対して形態素解析を行い，解析結果の形態素を単語とした．本稿で対象とするタスクは，語彙を学習データ中の出現頻度の高いものから7000語に限定したものである．7000語彙によって学習データ中の90.3%の単語をカバーしている．学習用テキストデータベース中に観測された各言語モデルの種類数，評価セットtrn(言語モデルclosed)およびtst(言語モデルopen)の各100文に対するテストセットパープレキシティを表1に示す．trnセットおよびtstセットともに未知語は含まれていない．ここで，n=2以上のn-gramについては，Katzのback-offスムージング[4]を適用した．また，言語モデルは句読点も含めて学習されているが，パープレキシティを求める際に，『句点-文末』という遷移の確率が非常に高くなり値が偏ってしまうため，文末シンボルをパープレキシティの計算から除いてある．読点はパープレキシティの計算に含まれる．表1をみると，n-gramのnを大きくするにしたがってtrnセットに対するパープレキシティが減少していくのに対して，tstセットに対するパープレキシティの減少は小さくなっており，両セットに対する値の差が大きくなっている．これは，学習データの不足とそれによって起こるtrnセットに対する過学習のためだと考えられる．tstセット100文における各n-gram系列の出現数および学習データで観測されなかったためback-offされるn-gramの数を表2に示す．表2をみると，4-gramでは28.1%がtrigram以下の値によってスムージングされていることがわかる．

表1: n-gram言語モデルの種類数とperplexity

	unigram	bigram	trigram	4-gram
観測された種類数	7k	2.2M	17.6M	43.3M
test set perplexity (trn)	620.5	56.3	24.8	14.6
test set perplexity (tst)	636.4	58.6	31.4	31.2

表2: tstセットに対するn-gram言語モデルの統計量

	unigram	bigram	trigram	4-gram
tstセット中の出現数	2354	2254	2154	2054
異なり数	848	1828	2028	2014
back-offされる数 (割合)	-	10 (0.4%)	147 (6.8%)	577 (28.1%)

3. マルチパス探索アプローチ

大語彙連続音声認識は，入力音声に対して音響モデルと言語モデルの与える事後確率が最大となる単語系列を求めるものである．しかし，性能向上のために精密なモデルを用いると探索空間が非常に複雑で膨大なものになってしまう．そこで，最初に比較的精度の粗いモデルを用いて候補を絞り込み，絞り込んだ候補に対してより精密なモデルを用いて探索を行うことにより，高精度なモデルを用いた効率的な探索を行うことができる[5]．

粗いモデルを用いた単語系列候補の絞り込みには，尤度順で上位N個の候補文を出力するN-best形式と，複数の候補文から単語ネットワークを生成するword lattice形式とが考えられる．N-best形式の場合は，Nを大きくするほど候補に正解が含まれる数が増えるが，探索空間も広がってしまう．したがって，正解が十分含まれ，かつできるだけ小さいNを用いることが望ましい．一方，word lattice形式は，N-bestに比べて非常に多くの候補を残すことができ，Nが大きいN-best候補を圧縮したものと考えることができる．そのため，N-best候補としてNが小さくても十分正解が含まれる場合には，word latticeを用いると探索空間を必要以上に広げてしまうため望ましくない．我々は，粗いモデルによる候補絞り込みには，N-best形式を採用した．

我々は探索を2段階に分けて行った．まず，最初の探索において，単語bigramを用いて入力音声に対する単語系列のN-best候補を出力する(first-pass探索)．次に，単語trigram，単語4-gramなどの高精度なn-gramをN-best候補に対して適用して単語系列の尤度を再評価することにより認識結果を得る(second-pass探索)．音響モデルについても，first-pass

* A Study of Large-Vocabulary Continuous Speech Recognition Using Higher Order n-gram Language Models.
By Katsutoshi Ohtsuki[1], Kotaro Yoshida[2], Tatsuo Matsuoka[1] and Sadaoki Furui[1,2]
([1]NTT Human Interface Laboratories, [2]Tokyo Institute of Technology)

で粗いモデルを用いることにより，高速な処理が実現できるが，今回はfirst-pass，second-passとも高精度な音響モデルを用いて探索を行った．

4. 大語彙連続音声認識実験

高次n-gram言語モデルとマルチパス探索手法を用いて語彙7000語の大語彙連続音声認識実験を行った．評価データは，男性10名による新聞記事読み上げ音声200文(trn, tst各100文)を用いた．

4.1 音響モデル

音響モデルとして文脈依存音素HMMを用いた．音響モデル学習用データ13270発話中における出現頻度の高かったdiphone，triphoneおよび文脈独立音素HMMからなる合計748モデルのセット(CD)を作った．音響特徴量は，16次のLPCケプストラムと正規化対数パワーおよびそれらの1次時間微分の34次元を用いた．学習用データと評価用データの収録系の違いに対処するためcepstral mean normalizationにより正規化したケプストラムを用いたモデルによる実験も行った(+CMN)．また，学習データ中に観測されなかった音素環境に対して近いモデルを割り当てるために，tree-based clustering[6]に基づいて文脈依存音素HMMのセットを設計した(TBC)．各音響モデルによる音素認識実験およびbigram言語モデルを用いた7000語彙の連続音声認識実験の結果を表3に示す．表3よりすべての音素環境を考慮したTBC+CMNが最も高い性能を示しており，このモデルを用いてマルチパス探索の実験を行う．

表3: 音響モデルの性能比較

音響モデル	状態数	音素誤り率 [%]	単語誤り率 [%]
CD	2198	42.8	18.1
CD+CMN		40.2	15.9
TBC+CMN	2106	36.7	13.3

4.2 first-pass探索

first-pass探索には，音響モデルとして前述のtree-based clusteringに基づくモデル，言語モデルとして単語bigramを用いる．trn，tst各セットに対するN-best探索の累積単語誤り率を表4に示す．Nの値が300程度で性能が飽和しているため，first-pass探索の出力として300-bestを用いる．すなわち300-bestの結果がsecond-pass探索の上限値となる．

表4: first-pass(N-best)探索の累積単語誤り率[%]

N	1	10	20	50	100	200	300
trn	15.7	9.3	8.6	7.3	7.0	6.3	5.9
tst	13.3	8.8	8.1	7.1	6.5	6.2	6.0

4.3 second-pass探索

second-pass探索では，first-pass探索と同じ音響モデルと単語trigramおよび4-gram言語モデルを用いる．音響スコアは，first-pass探索で得られたものをそのまま用い，300の候補に対して高次n-gramを適用して認識結果を得る．単語trigramおよび4-gramによるsecond-pass探索の結果(単語誤り率，word error rate: WER)および低次のn-gramに対する誤り削減率(error reduction: ER)を表5に示す．単語trigramを用いることにより，trnセット，tstセットともに単語誤り率が10%以下となっている．また，単語4-gramを適用することにより，trnセットではtrigramの結果に比べてさらに10%程度誤りが削減されているが，tstセットについてはtrigramの場合よりも性能が悪くなってしまっている．これは，表1においてtstセットのパープレキシティがtrigramと4-gramとでほとんど変わらないことからもわかるように，4-gramの推定には学習データが十分ではないためであると考えられる．ARPAのWSJ，NABタスクにおいて4-gramを適用した場合でも，trigramからの誤り削減率は5%未満となっており[7.8]，効果はそれほど大きくない．

表5: マルチパス探索による大語彙連続音声認識性能

評価セット	trigram		4-gram		
	WER [%]	ER[%] from bigram	WER [%]	ER[%] from bigram	ER[%] from trigram
trn	9.7	38.2	8.7	44.8	10.6
tst	9.5	27.7	10.0	24.2	-4.8

5. まとめ

本稿では，マルチパス探索を採用することにより，単語trigramや4-gramなどの高次n-gram言語モデルを大語彙連続音声認識に適用した．second-pass探索に単語trigramを用いることにより，単語bigramの場合に比べ約30%誤りが削減され，単語誤り率が10%未満となった．しかし，単語4-gramを適用した場合にはパープレキシティも下がらず，tstセットにおける性能改善はみられなかった．音声認識に有効な4-gramを推定するためには，非常に大量の学習データを用意するか，削除補間法などの補間を行うことが必要であると考えられる．

謝辞

tree-based clusteringによる音響モデルを作成していただいた山形大学の堀貴明氏に感謝します．形態素解析ツールを提供していただいたNTTヒューマンインターフェース研究所映像処理研究部の田中一男主幹研究員に感謝します．テキストデータの使用を許諾していただいた日本経済新聞社に感謝します．また，日頃御討論いただくNTTヒューマンインターフェース研究所古井特別研究室，早大白井研究室，東工大古井研究室の皆様に感謝します．

参考文献

[1] 大附他，信学技報，SP95-90.

[2] 松岡他，信学論，Vol. J79-D-II, No. 12, pp.2125-2131, 1996.

[3] 吉田他，信学技報，SP96-82.

[4] S. M. Katz, IEEE Trans. vol. ASSP-35, pp.400-401, 1987.

[5] R. Schwartz et al., in Automatic Speech and Speaker Recognition, ed. C.-H. Lee et al., pp.429-456, 1996.

[6] S. J. Young et al., Proc. ARPA HLT Workshop, pp.307-312, 1994.

[7] A. Ljolje et al., Proc. ARPA SLST Workshop, pp. 162-165, 1995.

[8] P. C. Woodland et al., Proc. ICASSP'95, pp. 73-76.

[9] "日本経済新聞CD-ROM版1990年版～1994年版，" 日本経済新聞社，1994-1995.