対話システムにおける音声認識の改善を目指した

バックオフ係数の検討

浦谷 則好，小早川 健
ＮＨＫ放送技術研究所

１．はじめに

音声認識技術は利用者が特殊な装置を利用する必要がないなどの手軽さから、いろいろな状況で利用されるようになっている。例えば、ＮＨＫでは誰でも簡単にデジタル放送を楽しめるようにするためのＴＶエージェントとの対話に音声認識を利用している[1]。また、ニュース番組に対するリアルタイムの字幕付与のためにも音声認識を利用している[2]。音声認識の言語モデルとしてはＮグラムモデルが使われ、データスパースネスに対処するために確率をスムージングする処理がなされる。このスムージング手法の中で主に利用されているのが Katz smoothing[3]という手法である。この手法では学習データ中に出現しない未知単語列に配分される頻度は、学習データ中に 1 回だけ出現した単語列の頻度と等しいとみなされ、そこから得られる重み（バックオフ係数）を使って未知単語列の出現確率が計算される。未知語列に対する配慮としては妥当に思われるが、その値の根拠は希薄である。システムを音声で制御する場合、システム側が受容できる発話の種類はかなり制限されていると思われるので、Katz smoothing で得られるバックオフ係数は過剰なものである可能性が高い[4]。また、原稿が存在するニュースや講演で、発話者が原稿にそれほど逸脱しない発話をすることが予想される場合にも、この仮定で得られるバックオフ係数は大きすぎる恐れがある。また、そこでバックオフ係数の最適化を目指して実験を行い、原稿に近い発話が予想される場合にはバックオフ係数を１／１００程度にした方がよいことを確認したので、それについて報告する。

２．実験

音声は 2001 年 6 月 1 日から 14 日までにＮＨＫのニュースで放送されたアナウンサーおよび記者の音声である。音響モデルは不特定話者用のモデル（男声モデル１種、女声モデル 1 種の自動切替）である。雑音が多い、発話が不明瞭などの理由で音声認識率が極端に悪いもの（単語認識率で 80%以下）や、あいさつなどの短い会話は大部分除外した。(あいさつなどは一部、テスト用に残した。）発話の数などの諸元は表 1 の通りである。1 発話（1 文章）当たり平均 25 単語である。

表１　対象としたニュース

ニュース項目	発話数	単語数
９２	８５３	２１，２２５

２．１．　実験１

上述のニュース音声の書き起こしから、122 語の挿入（「え」とか「あのー」というフィラー）と言い直しを認定した。この 122 語を除いたものを推定原稿として音声認識実験を行った。また、書き起こしそのものを原稿とみなした実験も行った。バックオフ係数を 1/10,1/32,1/100,1/316,1/1000
（10 の-1.0,-1.5,-2.0,-2.5,-3.0 乗）と変えて音声認識率の変化を求めた。なお、バックオフ係数以外は変化させていないので、累積接続確率は 1.0 となっていない。原稿の perplexity は書き起こしそのもの、推定原稿それぞれ 7.40,7.43 であった。実験結果を表２に示す。表中で左の数字が Correctnes(正解数／単語数)で、右の数字は Accuracy((正解数-挿入単語数）／単語数)である。表２を見ると完全にクローズドな場合（書き起こし原稿をそのまま使う場合）でもバックオフ係数を小さくするほどよいというわけではない

ことが分かる。バックオフ係数を従来の 1/100 程度にするのがよいと推定される。

表２　実験１の音声認識率（%）

	推定原稿		書き起こし	
	Corr.	Acc.	Corr.	Acc.
元(1)	97.23	96.89	97.64	97.27
1/10	98.20	98.01	98.70	98.54
1/32	98.30	98.13	98.83	98.68
1/100	98.32	98.12	98.88	98.72
1/316	98.12	97.84	98.80	98.61
1/1000	97.94	97.56	98.74	98.51

２．２．　実験２

２．１．の実験では原稿と実際の発話の相違はわずかに 0.58%しかない。「原稿に近い発話が予想される場合」という条件だとしてもその差が小さすぎる。そこで、ロバストネスを確認するために、書き起こし原稿にランダムに単語の 5%を削除、置換、挿入したものを原稿として同様の実験を行った。挿入、置換に使われる単語は元の原稿に含まれていないものから選んだ。結果を表３に示す。「混合」は 1.67%ずつの削除、置換、挿入を与えたものである。また、表で「排除」としているものは、元の原稿の単語から特定の単語（高頻度の機能語的なものと頻度 2 以下のものを除いた中頻度語から選択）をランダムに除いた実験である。それゆえ、「排除」の場合の最大認識率は 95%である。なお、実験は各モード４回ずつ実施した。表中の数値は４回の平均の Accuracy と Test data perplexity である。未知語率は当然、挿入が 0、排除が 5%であり、削除、置換、混合はそれぞれ、

表３　実験２の音声認識率（%）

	元(1)		1/100	
	Acc.	Per.	Acc.	Per.
削除	94.36	11.35	95.37	16.81
置換	94.27	12.13	95.76	18.09
挿入	95.81	10.69	97.45	14.12
混合	94.76	11.45	96.08	16.52
排除	88.34	9.76	89.63	9.94

0.38%, 0.37%, 0.29%であった。なお、Training data perplexity の平均は削除、置換、挿入、混合、排除、それぞれ、8.76, 9.33, 9.11, 9.03, 7.88 であった。

３．むすび

Katz smoothing によるバックオフ係数の算出手法に疑問を感じて、バックオフ係数の値を変化させて音声認識実験を行った。その結果、実際の発声が原稿とあまり違わないかぎりバックオフ係数を 1/100 にした方が良い結果を得られることを確認した。こういうケースは対話による機器操作ばかりでなく、ニュースや講演などでも考えうるものであり、有用な知見が得られたものと考える。なお、まだ実験途中であるが原稿を 10%変化させた場合も、バックオフ係数を 1/100 にした方が元のものより良い結果が得られている。

謝辞

本研究は当研究所の今井享主研をはじめとする音声認識研究チームから音響モデルと音声データの提供を受けて実施したものである。多大な協力に深謝する。

参考文献

[1]Jun Goto et al.:A Spoken Dialogue Interface for TV Operations Basedon Data Collected by Using WOZ Method, IEICE Trans. on Information and Systems, E87-D(6):1397-1404,2004

[2]安藤彰男外：音声認識を利用した放送用ニュース字幕制作システム，電子情報通信学会論文誌，J84-DII(6):877-887,2001

[3]Katz, S.M.:Estimation of probabilities from sparse data for the language model component of a speech recognizer, IEEE Trans. on Acoustics, Speech, and Signal Processing, ASSP-35(3):400-401,1987

[4]Kneser, R. et al.:Improved Backing-off for M-gram Language Modeling, IEEE International Conf. on Acoustics, Speech, and Signal Processing, 1:181-184,1995