論文 / 著書情報
Article / Book Information

論題(和文)	ニュース音声認識における言語モデルの改良
Title(English)	
著者(和文)	桜井直之, 古井 貞熙, 大附克年
Authors(English)	SADAOKI FURUI
出典(和文)	日本音響学会1999年春季講演論文集, Vol. , No. 2-1-3, pp. 57-58
Citation(English)	, Vol. , No. 2-1-3, pp. 57-58
発行日 / Pub. date	1999, 3

2—1—3 ニュース音声認識における言語モデルの改良 *

◎桜井 直之　古井 貞熙 (東工大)　大附 克年 (NTT ヒューマンインタフェース研究所)

1. はじめに

大語彙連続音声認識における日本語特有の問題として同一の漢字表記に対する読みの多様性が挙げられる．それらの読みは実際には偏った頻度で出現し，また前後の単語によって変化する．従来の方法では一つの漢字表記に対し可能性のある全ての読みを登録しておき，それらの読みが等確率で出現するとして認識をおこなっていたが，実際には頻度の低い読みによって認識性能が劣化するという問題があった．そこで本稿では漢字表記を認識単位とするのではなく，同じ漢字表記でも読みの異なるものは別の認識単位として言語モデルを生成するという手法を提案する [1]．また，実際のニュース音声の認識の際に問題となる「え」「えー」などの間投詞による，認識性能の劣化に対応するための言語モデルについても検討する．

2. 読みを考慮した言語モデル

2.1 読みの多様性

日本語の特徴として形態素の読みの多様性が挙げられる．例えば「人」という形態素には「ヒト」「ジン」「ニン」などの読み方があり，これらは前後の形態素によって決定される．従来の音声認識では漢字表記が同じである形態素は一つの認識単位として扱い，認識時に全ての可能性のある読みが等しい確率で出現するとして認識を行なっていた．しかし，このような方法では実際は出現確率の低い読みに対しても高い確率が与えられてしまい，認識誤りの原因となることがあった．また，読みの出現確率に応じて，認識時の確率を変えるという手法も提案されている [2] が，前後の形態素間のつながりを考慮していないという問題点があった．そこで本研究では，形態素解析時に付与された読みを利用した言語モデルの構築に関する検討を行なった．

2.2 読み依存言語モデル

漢字表記の形態素を一つの認識単位とするのではなく，漢字表記と読みを合わせた形態素を一つの認識単位として，言語モデルを作成した．つまり，表 1 のように漢字表記が同じでも読みが異なるものは別の認識単位として扱うことにする．このような，読みに依存した言語モデルを作成するためには形態素解析時にできるだけ正しい読みを自動的に付与することが不可欠である．本研究では高精度 (形態素単位で 99.6%) で正しく読みを付与することができる形態素解析システム JTAG[3] を利用している．また読みごとに別の形態素とした場合，形態素の種類は約 2%増加する．

表 1. 従来法と読み依存言語モデルの比較

従来法		読み依存 LM	
認識単位	発音辞書	認識単位	発音辞書
人	h i t o	人; ヒト	h i t o
	j i N	人; ジン	j i N
	n i N	人; ニン	n i N

3. 間投詞を考慮した言語モデル

放送ニュース音声には，「え」「えー」などの間投詞が文頭，文中に含まれることがある．放送ニュース音声を書き起こしたテキストを分析するとそれらは文頭や読点 (，) の後に多く出現することがわかり，その頻度は表 2 のようになっている．これらの間投詞は言語モデル学習用のニュース原稿テキストには含まれていないので，誤認識の原因になりやすい．そこで，従来は言語モデルの学習時に削除していた読点を削除せずに言語モデルを作成した．さらに文頭及び読点の発音として発音辞書には無音，「え」，「えー」の 3 種類を等確率で与えた．これにより言語モデルで読点が存在する位置で間投詞が発声された場合に読点として認識することが可能になり，前後の単語の認識性能の劣化を抑えることができる．

表 2. 書き起こしテキスト中の「え」「えー」の頻度

	文頭	読点の後
「え」	12.1%	3.4%
「えー」	2.1%	1.0%

4. 認識実験

4.1 言語モデル

言語モデルの学習に用いたデータは放送ニュース原稿テキスト 5 年分 (1992 年 7 月から 1996 年 5 月)，約 50 万文である．形態素解析システム JTAG を用いて形態素に分解し，その形態素を単語として単語 bigram,trigram を学習した．単語出現頻度上位 2 万語を認識語彙とした．読みごとに別の形態素とした場合も同様に認識語彙は出現頻度上位 2 万語に揃えた．観測されなかった n-gram は Katz の back-off 平滑化を適用した．なお言語モデルの作成には CMU/Cambridge Toolkit [4] を用いた．

4.2 音響モデル

今回の実験で用いた音響モデルは，tree-based clustering によって状態共有化を行なった不特定話者文脈依存音素 HMM [5] である．音響特徴量としては 16 次の LPC ケプストラムと正規化対数パワー，及びそれらの一次微分の計 34 次元を用いた．学習用音声

* An Improvement of Language Modeling for Automatic Transcription of Japanese Broadcast-News Speech
By Naoyuki Sakurai and Sadaoki Furui (Tokyo Institute of Technology) Katsutoshi Ohtsuki (NTT Human Interface Laboratories)

データは,ATR 音声データベース B セット，日本音響学会連続音声データベース,および同模擬対話データベースから，男性 53 名による 13270 発話，女性 56 名による 13367 発話を用いた．学習は男女別々に行ない，性別依存モデルを作成した．モデルの総状態数は男性が 2106, 女性が 2083 である．各状態のガウス分布の混合数はすべて 4 である．

4.3 評価用データ

1996 年 7 月に実際に放送されたニュース音声から，スタジオで収録されたクリーンな発話 (clean) とそれ以外の背景に雑音や音楽がのっている発話や記者レポートなどの発話 (noisy) をそれぞれ男女 (male, female)50 文ずつ抽出した．(以下 m/c,m/n,f/c,f/n と略す.) 各評価セットには 5〜6 名の話者の音声が含まれている．

4.4 認識実験結果

まず,従来の読みを等確率に扱う言語モデル (LM1) と読み依存言語モデル (LM2) をそれぞれ用いて認識実験を行なった．各言語モデルの学習テキストおよび各評価セットに対する未知語率を表 3 に示す．語彙を 20k に固定しているので LM2 のカバー率が若干低下するが，それほど大きな低下は起こらないことがわかる．

認識実験の結果を表 4 に示す．全ての評価セットにおいて読みを考慮した言語モデルを用いることで単語正解精度が向上していることが分かる．bigram モデルにおいて平均 4.0%,trigram モデルでは平均 4.5%誤りが減少している．

次に間投詞対策として，形態素に読点を含めた言語モデル (ベースとなっているのは読み依存言語モデル LM2) による認識実験を行なった．文頭記号及び読点の発音として無音のみを与えたもの (LM3-1)，無音，「え」,「えー」の 3 種類を与えたもの (LM3-2) についての実験結果を表 5 に示す．言語モデルに読点を加えるだけでも単語正解精度は若干向上するが，文頭及び読点の発音として「え」「えー」を加えることで，さらに認識性能が向上することがわかる．LM2 に比べて LM3-2 では bigram モデルにおいて平均 6.3%，trigram モデルでは平均 7.8%誤りが減少している．

表 3. 20k 語彙に対する未知語率 [%]

言語モデル	学習用テキスト	評価セット m/c	m/n	f/c	f/n
LM1	2.27	0.81	2.88	1.21	3.49
LM2	2.39	0.86	3.02	1.26	3.68

4.5 間投詞の影響分析

上記の実験に関し，間投詞の影響についてもう少し詳しく分析した．200 文の評価セット中には間投詞「え」「えー」が 33 回出現する．LM2 で認識実験を行なうと，間投詞の前後の単語の単語正解精度は 67.3%となるのに対し，LM3-2 を用いることで 75.5 % になる．この値は評価セット全体の平均単語正解精度 (75.4%) にほぼ等しい．このことからも LM3-2 の間投詞対策の効果が確認できた．

表 4. 読みを考慮した言語モデルの評価 (単語正解精度 [%])

言語モデル		評価セット m/c	m/n	f/c	f/n
bigram	LM1	79.1	59.7	81.7	54.8
	LM2	79.6	60.9	82.7	57.3
trigram	LM1	82.4	62.8	85.7	58.8
	LM2	83.2	64.1	86.4	60.7

表 5. 間投詞対策言語モデルの評価 (単語正解精度 [%])

言語モデル		評価セット m/c	m/n	f/c	f/n
bigram	LM2	79.6	60.9	82.7	57.3
	LM3-1	80.1	62.2	83.1	57.7
	LM3-2	82.0	62.8	83.9	58.3
trigram	LM2	83.2	64.1	86.4	60.7
	LM3-1	83.3	65.4	86.3	60.9
	LM3-2	85.8	66.9	87.1	61.9

5. まとめ

日本語の読みの多様性に対処するために，同じ漢字表記でも読みが異なる単語を別の単語として扱う読み依存言語モデルを提案した．これをニュース音声の認識に適用したところ未知語率はそれほど大きくならずに，単語正解精度は従来の言語モデルに比べて向上した．

さらに放送ニュース中の間投詞によって認識性能が低下するのを抑えるために，読点を残した言語モデルを作成し，文頭と読点の発音として「え」「えー」を加える手法を提案した．その結果，読み依存モデルと合わせると誤り率を約 12%削減することができた．

謝辞

ニュース原稿及び音声データを提供して頂いた NHK 放送技術研究所に感謝します．形態素解析ツールを提供して頂いた NTT ヒューマンインタフェース研究所情報通信研究所知的通信処理研究部の渕武志研究主任に感謝します．日頃討論頂く東工大の研究室の方々に感謝します．

参考文献

[1] 大附, 他, “ニュース音声認識のための言語モデルと音響モデルの検討”, 信学技報,SP98-108,1998

[2] 高木, 他,“ニュース音声を対象とした言語モデルと話題抽出の検討”, 信学技報,SP98-33,1998

[3] 渕, 他,“保守性を考慮した日本語形態素解析システム”, 情報処理学会, 自然言語処理研究会,NL97-4,pp.59-66,1997

[4] http://svr-www.eng.cam.ac.uk/~prc14/toolkit.html

[5] 大附, 他, “ニュース音声を対象とした大語彙連続音声認識と話題抽出”, 信学技報,SP97-27,1997