音声言語情報処理 35－3
（2001．2．2）

高齢者向け音響モデルによる大語彙連続音声認識の評価

馬場 朗* 芳澤 伸一* 山田 実一**
李 晃伸** 鹿野 清宏**

* イメージ情報科学研究所
** 奈良先端科学技術大学院大学 情報科学研究科

あらまし 近年、大語彙連続音声認識を利用したシステムの普及に伴い、さまざまな用途で音声認識が利用されるようになりつつある。音声認識システムの性能に悪影響を与える要因の一つとして、システムを利用するユーザーの音響特性と音響モデルとの間のミスマッチがある。一般に、音響モデルの学習には成人音声を用いるため、高齢者の音響特性との間にミスマッチを生じ、認識率の低下を生じる可能性がある。本論文では、大規模な高齢者音声データベース(200文章×301人)を用いて音響モデルの学習を行い、この高齢者向け音響モデルを大語彙連続音声認識システムにおいて評価した。実験結果では、成人データベース(150文章×260人)から学習したモデルによる認識結果と比較して、3～5%の単語認識率の改善が得られた。

キーワード 高齢者音声認識, 大語彙連続音声認識, 音響モデル

Acoustic Model Evaluation for Elder Speaker Speech Database

Akira BABA* Shinichi YOSHIZAWA*
Miichi YAMADA** Akinobu LEE** Kiyohiro SHIKANO**

* Laboratories of Image Information Science and Technology
** Graduate School of Information Science, Nara Institute of Science and Technology

Abstract Speech recognition technologies have been widely used in various areas due to the recent developments of large vocabulary continuous speech recognition (LVCSR) algorithms. Acoustical difference among speakers is considered to be one of the main reasons for the degradation of speech recognition rates. Especially, the acoustic difference between elder speaker speech database and usual adult speech database should be evaluated and researched for elder speakers to use speech recognition systems. In this paper, we evaluate elder speaker acoustic models in LVCSR, which are trained by the 301-elder-speaker utterance database, where each speaker utters 200 sentences. The elder speaker PTM acoustic model attains 88.9% word recognition rate, which is better than 85.4% word recognition rate by the usual aduld PTM acoustic model.

Key words Recognition of elder speaker, Large Vocabulary Continuous Speech Recognition, Acoustic model

1 はじめに

近年、大語彙連続音声認識を利用したシステムの普及に伴い、さまざまな用途で音声認識が利用されるようになりつつある。音声認識システムの性能に悪影響を与える要因の一つとして、システムを利用するユーザーの音響特性と音響モデルとの間のミスマッチがある。一般に、音響モデルの学習には成人音声を用いるため、高齢者の音響特性との間にミスマッチを生じ、認識率の低下を生じる可能性がある。この問題に対して、音響モデルを高齢者音声により学習する方法が提案されているが、語彙数,データ長も少なく、大語彙連続音声認識に適用できるかどうかの検証はなされていない [1]。本論文では、大規模な高齢者音声データベース (200 文章× 301 人) を用いて音響モデルの学習を行い、この高齢者向け音響モデルを大語彙連続音声認識システムにおいて評価する。

2 高齢者向け音響モデル

2.1 音声データベース

音響モデル作成用として、60 歳から 91 歳までの 301 人の被験者が、それぞれ ATR 音素バランス文 100 文と新聞記事読み上げ文 100 文の読み上げを行なった。音声データは、標本化周波数 16kHz、16bit で量子化されている。音声データ量は時間換算で約 133 時間である。また評価用話者として 100 人の被験者が同様の収録を行なっているが、本稿ではこのうち 46 人の話者を評価用話者とする[1]。書き起こしは人の手作業により行い、発声を可能な限り音としてカタカナ表記してある。各被験者は関西在住であり、収録は録音ブースにおいてヘッドセットマイクとデスクトップマイクを用いて行った。被験者の性別・年齢構成を表 1 に示す。

2.2 音響モデル

ヘッドセットマイクの高齢者音声データベースを用いて 3 種類の音響モデル (高齢者音響モデル) の学習を行なった。各音響モデルは、43 音素からなる 3 状態の left-to-right 型の HMM である。

モノフォンモデルは、各状態につき 16 個の混合ガウス分布をもつ。トライフォンモデルは、2000 状態の状態共有がなされており、各状態での混合ガウス分布数は 16 である。同様に、PTM(Phonetic Tied Mixture) モデル [2] は、2000 状態の状態共有がなされており、共通の中心音素をもつトライフォンの各状態間での出力確率も共有されている。各状態での混合ガウス分布数は 64 である。

各音響モデルには、それぞれ性別非依存 (GI: Gender Independent) モデルと性別依存 (GD: Gender Dependent) モデルがある。表 2 に音響分析時のパラメータを示す。

表 1: 高齢者話者の年齢構成

	学習用話者			評価用話者		
	男性	女性	合計	男性	女性	合計
60-64 歳	47	52	99	6	12	18
65-69 歳	49	46	95	9	7	16
70-74 歳	39	35	74	6	2	8
75-79 歳	11	14	25	2	2	4
80-84 歳	4	2	6	0	0	0
85-89 歳	1	0	1	0	0	0
90-94 歳	0	1	1	0	0	0
合計	151	150	301	23	23	46

表 2: 音響分析パラメータ

分析窓長	25msec, ハミング窓
分析間隔	10msec
特徴パラメータ	MFCC+ΔMFCC+ΔPow(計 25 次元)

3 評価実験

3.1 評価システム

大語彙連続音声認識システム Julius[3] より、新聞記事読み上げ文の認識実験を行う。言語モデルは新聞記事から作成され、その語彙数は 2 万語である。301 人の学習用話者とは異なる 46 人の評価用話者による合計 200 文の新聞記事文[2]の発声により認識実験を行う。

高齢者音響モデルとの比較評価のため、JNAS

[1]この 46 人は、JNAS の 20k システムでの評価データと同じ文を発声している。

[2]音響モデル学習用の話者の読み上げ文にも同一の文章が含まれているが、これらの文章を除いた文章を用いて学習した場合でも認識結果は変わらなかった。

音声データベース [4] を用いて音響モデル (JNAS 音響モデル) の作成を行う。表 3 に JNAS 話者の性別・年齢構成を示しておく。学習用の音声データとして、男女合計 260 人の ATR 音素バランス文 50 文と新聞記事読み上げ文 100 文から、高齢者音響モデル同様の手法で音響モデルの学習を行った。学習に用いた音声データの総量は時間換算で約 69 時間である。また、高齢者音響モデルと JNAS 音響モデルの構造・学習手順は全く同一である。

表 3: JNAS 話者の年齢構成

	学習用話者			評価用話者		
	男性	女性	合計	男性	女性	合計
10-19 歳	1	0	1	0	0	0
20-29 歳	73	68	141	17	13	30
30-39 歳	35	39	74	5	8	13
40-49 歳	10	14	24	1	2	3
50-59 歳	5	5	10	0	0	0
60 歳以上	5	3	8	0	0	0
不明	1	1	2	0	0	0
合計	130	130	260	23	23	46

3.2 高齢者音響モデルによる認識実験

まず、高齢者音響モデルとの比較のため、JNAS 音響モデルでの JNAS 評価用話者の認識実験結果 (単語正解率 (word correct)) を表 4 に示す。表 4 において、各数値は評価用話者の音声の認識結果を単語正解率で表してある。各音響モデルについて、GI モデルと GD モデルでの男女平均, 男性平均, 女性平均が示されている。モデル毎に結果を見ると、モノフォン <PTM< トライフォンの順に認識率が高く、また全モデルで男性より女性の方が認識率が高い。

高齢者音響モデルでの高齢者音声の認識実験結果を表 5 に示す。また、JNAS 音響モデルでの高齢者音声の認識実験結果を表 5 の () の中に示す。表の構成は表 4 と同じである。

表 4, 表 5 より、モノフォンモデル (GID) では評価話者が高齢者になることによって JNAS 音響モデルでは認識率が 8.4%低下したが、高齢者音響モデルで認識することにより 4.0%改善されている。同様に PTM モデル (GID) では 9.0%低下したが、高齢者音響モデルにより 4.1%の改善効果が得られた。トライフォンモデルでは 9.6%低下したが、高齢者音響モデルにより 4.4%の改善効果が得られた。

表 4: JNAS 音響モデルによる JNAS 音声の単語認識率

	モノフォン		PTM		トライフォン	
	GI	GD	GI	GD	GI	GD
平均	83.8	86.1	92.5	93.7	94.2	94.9
男性	81.5	84.2	91.4	92.8	93.8	94.8
女性	86.2	88.0	93.7	94.6	94.7	95.0

これらの結果より、高齢者音響モデルの有効性が確認された。認識率を GD モデルで男女間で比較すると、JNAS 音響モデルでは、女性話者の方がモノフォンで 3.8%、PTM で 1.8%高いが、高齢者音響モデルでは、女性話者の方がモノフォンで 12.1%、PTM で 7.6%高い。この結果より、高齢者話者では特に男性話者の認識率が低いことが分かる。

次に、高齢者データベースと JNAS データベースをあわせたデータベース (561 人) で学習した音響モデル (高齢者+JNAS 音響モデル) の効果を調べるために、同じく大語彙連続音声認識実験を行なった。認識実験結果を表 6 に示す。音響モデルは全て性別非依存で、高齢者音響モデル,JNAS 音響モデルと同様の手法で作成されている。表 6 において、認識結果は高齢者評価話者と JNAS 評価話者それぞれ独立に集計されている。単語正解精度 (word accuracy) で集計した結果を () 内に示す。

表 4, 表 5 と比較すると、モノフォンモデルでの認識率は、JNAS 話者で 2.4%、高齢者話者で 0.3%低下している。同様に PTM モデルでは、JNAS 話者で 0.8%、高齢者話者で 0.6%低下している。トライフォンモデルでは、JNAS 話者で 0.2%向上し、高齢者話者で 0.3%低下している。この結果より、成人向け音響モデル, 高齢向け音響モデルを別々に作成することの必要性が確認できた。

表 5: 高齢者音響モデルによる高齢者音声の単語認識率

	モノフォン		PTM		トライフォン	
	GI	GD	GI	GD	GI	GD
平均	79.5(75.5)	81.8(78.3)	87.1(83.5)	88.9(85.4)	89.0(84.6)	90.0(84.9)
男性	73.5(68.2)	75.7(72.7)	82.6(78.1)	85.1(81.5)	86.0(79.8)	87.1(81.5)
女性	85.4(82.9)	87.8(84.0)	91.5(89.0)	92.7(89.4)	92.0(89.5)	93.0(88.3)

() の中は JNAS 音響モデルによる高齢者音声の単語認識率

表 6: 高齢者+JNAS 音響モデルでの認識実験結果

	モノフォン	PTM	トライフォン
JNAS	81.5(79.9)	91.7(90.6)	94.4(93.3)
高齢者	79.2(76.4)	86.5(83.8)	88.6(86.8)

() の中は単語正解精度 (word accuracy)

3.3 MLLR による話者適応実験

高齢者音響モデルと JNAS 音響モデルの音響的差異の比較のため、MLLR[5] による話者適応実験を行った。実験では、各評価用話者について評価用データとは異なる音声データを用い、音響モデルのガウス分布の平均値に対して適応を行なった[3]。JNAS 音響モデルから高齢者評価話者への適応と、高齢者音響モデルから高齢者評価用話者への適応を、話者毎に行ない、各話者で前節と同様の認識実験を行なっている。実験結果を表 7 に示す。

表 7: 高齢者音声の MLLR による適応結果

	初期モデル	適応前	10 文章	50 文章
モノフォン	(JNAS)	(75.5)	(80.1)	(82.2)
	高齢者	79.5	81.2	82.1
PTM	(JNAS)	(83.5)	(86.6)	(88.2)
	高齢者	87.1	88.6	88.8

[3]平均と分散の両者に対して適応を行なった場合、平均のみを適応した場合と比べてモノフォンモデルでは約 1%認識率が高いが、PTM では効果が無かった。本稿では実験の簡略化のため平均のみを適応した。線形変換を行なったサブスペースの数は、モノフォン (50 文章) では高齢者全話者の平均で 23.3、PTM(50 文章) では 28.6 であった。

適応用データとして 10 文章を使用した場合には、適応元となった音響モデルによる差が、モノフォンモデルでは 1.0%、PTM モデルでは 2.0%あり、適応用データとして 50 文章を使用した場合には、モデルによる差は、モノフォンモデルでは差はなく、PTM モデルでは 0.6%である。これらの結果より、MLLR による話者適応を行なう場合においても、初期モデルが重要であり、高齢者音響モデルが有効であることが分かる。

3.4 充足統計量による話者適応実験

充足統計量と話者距離を用いた教師なし学習手法 [6] による話者適応実験を行った。充足統計量は、高齢者+JNAS 音響モデルを作成するのに使用した話者 (561 人) について作成し、話者選択には 64 混合 1 状態の GMM を用いた。各評価用話者について、評価用音声とは異なる音声データ 1 文に対する尤度を、話者選択用 GMM について求めた。尤度の大きい話者 N 人の充足統計量から音響モデルを合成し、この合成された音響モデルを用いて各話者について認識実験を行なった。話者数 N は 20,40,60,80,100 について実験を行ない、最も認識率の良い話者数 N の場合の結果を表 8 に示す。この話者数 N はモノフォンモデルでは 20 人、PTM モデルでは 40 人である。

図 1 に評価用話者に対する GMM 尤度上位 100 人の話者候補を示す。データベース全体を、高齢者男性話者, 高齢者女性話者,JNAS 男性話者,JNAS 女性話者に分類した上で、各評価話者に対して尤度が高いと判断された学習用話者が同一のカテゴリである場合をカウントしてある。また尤度上位 100 人は 10 人毎に区切られている。例として、高

表 8: 充足統計量による話者適応結果 (高齢者+JNAS 音響モデル)

	JNAS 評価話者		高齢者評価話者	
	適応前	適応後	適応前	適応後
モノフォン	81.5	85.8	79.2	82.2
PTM	91.7	92.0	86.5	88.0

齢者男性話者の上位 1 位から 10 位では、高齢者評価用男性話者 23 人について、尤度の高いと判断された学習用話者 10 人のうち、9 人が高齢者男性話者であったことを示している。図より高齢者話者への適応時に、高い確率で高齢者話者の充足統計量が選択されていることが分かる。

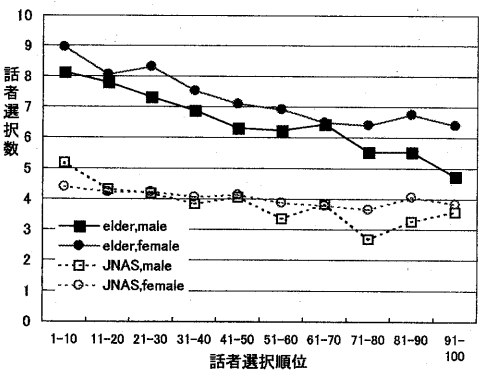

図 1: 話者選択結果

高齢者音響モデル,JNAS 音響モデルについても話者適応実験を行なった。高齢者評価話者に対して、音響モデル作成用話者 301 人のうち尤度の高い話者の充足統計量を、モノフォンモデルでは上位 20 人、PTM モデルでは上位 40 人選択して、各話者に適応した音響モデルを合成した。同様の実験を、JNAS 音響モデルでは音響モデル作成用の 260 人の充足統計量を用いて行なった。高齢者音響モデルでの認識実験結果を表 9 上段に、JNAS 音響モデルでの認識実験結果を下段に示す。これらの結果より、充足統計量と話者距離を用いた教師なし学習手法において、適応元となる音響モデルに高齢者音響モデルを用いることが高齢者音声認識に有効であることが分かる。

表 8 と表 9 の比較により、充足統計量による話者適応を行なう場合においても、初期モデルには JNAS+高齢者音響モデルよりも高齢者音響モデルを用いる方が高齢者音声の認識には優れていることが分かる。JNAS 評価話者についても同様である。

また、表 7 と表 9 の比較により、高齢者音響モデルを初期モデルとする話者適応において、MLLR(50 文章) と、充足統計量と話者距離を用いた教師なし学習手法が同程度の適応効果が得られていることが分かる。

表 9: 充足統計量による話者適応結果 (高齢者音響モデル、JNAS 音響モデル)

初期モデルは	JNAS 評価話者		高齢者評価話者	
高齢者モデル	適応前	適応後	適応前	適応後
モノフォン	78.5	81.9	79.5	81.8
PTM	86.6	89.1	87.1	88.4
初期モデルは	JNAS 評価話者		高齢者評価話者	
JNAS モデル	適応前	適応後	適応前	適応後
モノフォン	83.8	86.4	75.5	77.0
PTM	92.5	92.5	83.5	84.5

図 2 に高齢者音響モデルによる各高齢者話者の認識率と、話者適応を行なった後の認識率を示す。図中の結果は適応前の認識率で並べ変えてある。認識率の特に悪い話者が大幅に改善され、全体として認識率が向上していることが分かる。

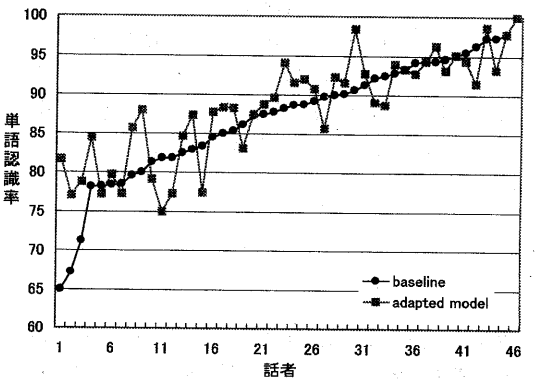

図 2: 話者毎の適応効果

4 おわりに

大規模な高齢者音声データベースを用いて音響モデルの学習を行なった。3種類の音響モデルについてJNSA音響モデルとの比較を大語彙連続音声認識システムJuliusで行なった結果、PTMで5.2%の単語認識率の改善が得られた。また、高齢者話者認識では特に男性話者の認識が困難であることが分かった。

高齢者データベースとJNASデータベースを合成したデータベースで音響モデルの学習を行ない、高齢者評価話者,JNAS評価話者それぞれについて認識実験を行なった。モノフォンモデル,PTMモデルで認識率の低下が確認され、高齢者向け音響モデル,成人向け音響モデルが必要であることが分かった。

高齢者音響モデルによるMLLRを使った話者適応実験を行なった。認識実験結果では、JNAS音響モデルから適応を行なった場合と比較して、最大で2.0%(PTM,10文章)の差があり、MLLRのよる話者適応を行なう場合に初期モデルが重要であり、高齢者音響モデルが有効であることが確認された。

充足統計量と話者距離を用いた教師なし学習手法を、高齢者音響モデル,JNAS音響モデル,高齢者+JNAS音響モデルに対して行なった。高齢者+JNAS音響モデルでの実験において、高齢者評価話者に対して高齢者の充足統計量が高い確率で選択されており、高齢者話者に対してJNASデータベースよりも高齢者音声データベースの方がマッチしていることがわかる。高齢者音響モデル,JNAS音響モデルでの結果との比較により、高齢者音響モデル,JNAS音響モデルから適応した場合には、MLLRによる適応結果と同程度の高い適応効果が得られた。

謝辞 本研究は、NEDO:新エネルギー・産業技術総合開発機構の援助を受けて行われた。高齢者音声データベースの収集・データ編集は本プロジェクトの一環としてTIS株式会社によって行なわれた。ご協力いただいた関係各位に感謝します。

参考文献

[1] 小沼智浩, 桑野裕康, 木村達也, 渡辺泰助, "高齢者音声の解析と認識評価", 日本音響学会講演論文集, 2-Q-1, pp.117-118, 1997.

[2] 李 晃伸, 河原達也, 武田一哉, 鹿野清宏, "Phonetic Tied-Mixture モデルを用いた大語彙連続音声認識", 電子情報通信学会論文誌, J83-D-II No.12, pp.2517–2525, 2000.

[3] 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, J82-D-II No.1, pp.1–9, 1999.

[4] Katsunobu Itou, Mikio Yamamoto, Kazuya Takeda, Toshiyuki Takezawa, Tatsuo Matsuoka, Tetsunori Kobayashi, Kiyohiro Shikano, Shuichi Itahashi, "JNAS:Japanese speech corpus for large vocabulary continuous speech recognition research", The Journal of the Acoustical Society of Japan (E), Vol.20, No.3, pp.199-206, 1999.

[5] C.J.Leggetter, P.C.Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, vol.9, pp.171-185, 1995.

[6] 芳澤伸一, 馬場朗, 松浪加奈子, 米良祐一郎, 山田実一, 鹿野清宏, "充足統計量と話者距離を用いた音韻モデルの教師なし学習", 情処研報, SLP-34-15, pp.83-88, 2000.