論文 / 著書情報
Article / Book Information

論題(和文)	音声認識における耐性向上の研究
Title	
著者(和文)	張志鵬, 古井貞熙
Author	SADAOKI FURUI
出典(和文)	日本学術振興会未来開拓学術研究推進事業研究プロジェクト「音声言語による人間－機械対話システムの研究」成果報告会資料集, Vol. , No. , pp. 23-30
Journal/Book name	, Vol. , No. , pp. 23-30
発行日 / Issue date	2001,

日本学術振興会未来開拓学術研究推進事業研究プロジェクト
「音声言語による人間-機械対話システムの研究」
成果報告会資料集、平成13年3月17日、学士会館分館、pp. 23-30

音声認識における耐性向上の研究

張 志鵬　古井 貞熙
東京工業大学大学院情報理工学研究科計算工学専攻
zzp@furui.cs.titech.ac.jp, furui@furui.cs.titech.ac.jp

1. はじめに

近年音声認識技術が著しく進展し、数多くのアプリケーションが実現されるようになった。しかし、まだ多くの研究課題が残っている。認識誤りの主たる原因は、学習データと認識すべき音声の間になんらかのずれがあることである。その背景には、話者の個人差が極めて大きいこと、種々の雑音とマイクロホンや伝送系などの歪みなどがあげられる。これらの音響変動に対応する耐性の向上技術が極めて重要である。

話者適応技術は音声認識において話者の個人差の問題に対処する重要な手段である。この問題に対応すべく、多くの改善手法が提案された。例えば、話者の少量データを用いて適応化する場合、尤度最大基準に基づいて HMM パラメータを線形変換により適応化する MLLR 法 1)とベイズ適応から導かれるMAP法 2)などが有効である。

また雑音に関しては種々の変動に対応できる基本的な適応化法として有望なのが HMM 合成法 3)である。この手法では、音声の HMM と雑音の HMM の畳み込みをスペクトル空間で行って、雑音が重畳した音声の HMM を合成する。しかしながら学習に使う雑音データと認識環境の雑音が同じでなければならないという制約などがある。

本論文では、話者適応及び雑音適応に関する研究成果をまとめて報告する。実際に放送されたニュース音声の認識をタスクとしている。

2. 話者適応化法の研究

2.1 話者変動の検出を含む適応法

多くの音声認識のアプリケーションでは話者の交代が頻繁に起こり、しかも未知の話者の音声が入力されるので、オンラインの教師なし適応を行うことが必要である。例えば、ニュース音声にはスタジオのアナウンサーによる発声だけでなく、中継先の記者や VTR 映像にあわせて原稿を読み上げた発声など様々な話者の発声が含まれている。この話者適応に関しては、次のことを考慮することが必要かつ有効と考えられる。

(1) 事前に話者情報を得ることができないので、オンライン即時型適応が必要。
(2) 同じ話者が複数の文を続けて発声することが多いので、逐次型適応が有効。
(3) 話者の交代情報を得ることができないので、自動的に検出することが必要。

本研究では、このような観点から、話者境界を自動的に検出しながらオンライン即時・逐次型教師なし話者適応を行う方法について検討した。

不特定話者の音素モデルを尤度最大化規準で特定の話者に適応化した場合、同じ話者の異なる音声に対するそのモデルの尤度は、不特定話者のモデルの尤度よりも高くなると期待される。逆に、新しい話者の音声の声質がそれ以前の話者の音声と異なる場合には、新しい話者の音声は、以前の話者に適応化したモデルよりもむしろ不特定話者のモデルに適合すると考えられる。従って、適応化モデルと不特定話者モデルの尤度を比較することによって話者境界を検出し、高い尤度を示すモデルを用いて、新しい話者に適応させるのが適当と考えられる。そして、同じ話者が複数の文を継続して発声していると判定される間は、そのモデルを逐次適応化して行くことにより、認識性能が向上すると予想される。さらに、新しい話者が検出された後でも、以前の話者が再度発声することが考えられるの

で、ある程度の数の話者に適応したモデルをそれぞれ保存しておき、活用するのが適当であろう。このような適応化法はニュース音声認識だけでなく、対話システム、会議など、話者交代を伴う多くの場合に使えると考えられる。本研究では、計算時間を考慮し、尤度比較するモデルとして、直前の話者、現在の話者、および不特定話者の三つのモデルを使うことにした。

適応手法に関してはまず MLLR および MAP 法によりモデルパラメータの変換行列を求め、その後 VFS 4) により移動ベクトルを平滑化する方法を用いた。すべての音素を共通の行列で変換する場合と、音素による違いを考慮して、無音、子音、各 5 母音、計 7 つのクラスタに分類し、各クラスタに対する変換行列を用いる場合について検討した。

音響モデルは、tree-based clustering によって状態共有化を行った不特定話者文脈依存音素 HMM 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。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。

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モデルを学習した。単語出現頻度上位 2 万語を認識語彙とした 5)。

教師なし適応の結果を図 1 に示す。図には、適応化を行う前(Baseline)、音素を 1 クラスタで適応化した場合(1 cluster)、および 7 クラスタ(7 clusters)の結果が示してある。言語モデルとして、bigram と trigram を用いた。全ての評価セットにおいて、適応化により単語正解精度が向上していることが分かる。1 cluster の場合、bigram モデルにおいて平均 7.8%, trigram では平均 4.2%誤り率が減少している。7 clusters で適応化した場合、適応化前に比べて、誤り率が男女平均で約 12%減少している。

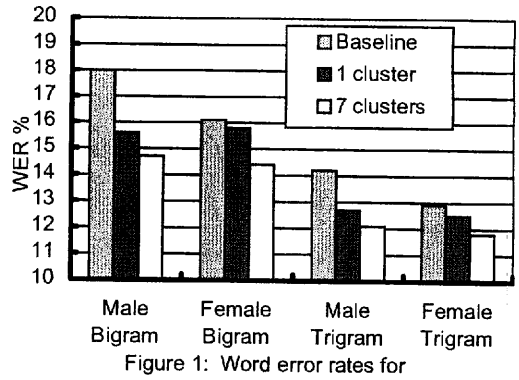

Figure 1: Word error rates for 1 cluster and 7clusters.

参考のための比較実験として、正しい話者境界を与え、そこで適応の初期モデルとして不特定話者モデルを強制的に用いる実験を行った。7 clusters の実験結果を図 2 に示す。自動検出の方が、平均 3.6%誤り率が小さい。声質の似ている話者のモデルを初期モデルとして適応する方法が、不特定話者モデルを初期モデルとして使うより有効だと言える。

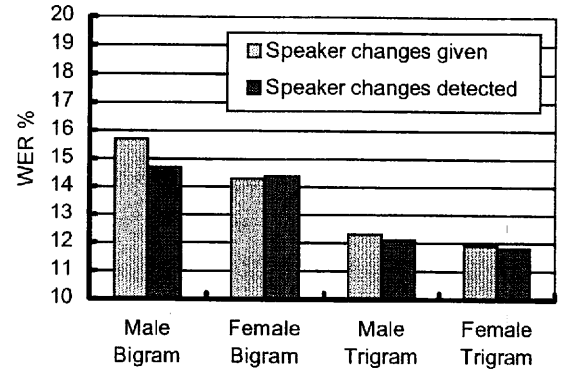

Figure 2: Comparison of word error rates for detected and given speaker changes.

2.2 GMM による改善法

話者変動を検出しながらオンライン適応を行う手法の有効性を確認したが、この方法には、複数の音素 HMM セットによる認識を並列的に行うため、計算量が大きくなるという問題があった。そこで HMM の代りに 1 状態の混合ガウス分布 (GMM; Gaussian mixture model) を用いて、話者交代を検出することにより、計算量を削減する方法を提案する。

テキスト独立形話者認識において、これまで GMM が広く使われている 6)。GMM を話者交代の

検出に用いるためには、GMMが話者適応化後の音素HMMの個人差を十分に表現している必要がある。本研究においては、音素HMMの話者適応は基本的にMLLR法によって行なわれる。このため、音素HMMの適応化と並行して、GMMをHMMと同じ変換行列で適応化することにした。ただし、音素HMMは7つの音素クラスタに分けて変換しているが、GMMは音素独立のため、音素HMMを1クラスタで変換する変換行列を求め、これを用いてGMMを変換する。オンラインの適応の流れを図3に示す。

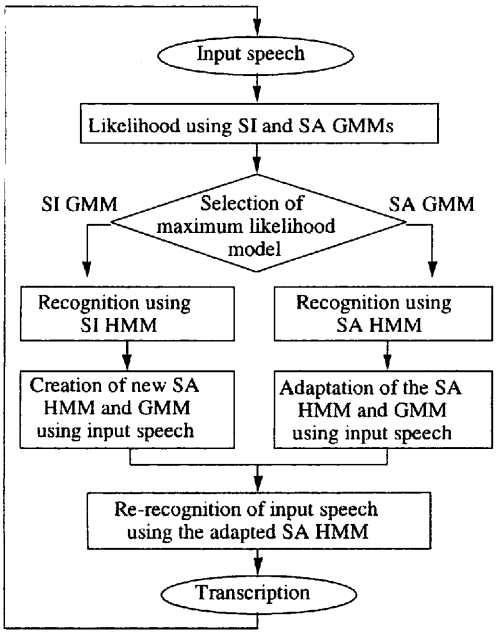

Figure 3: Online incremental speaker adaptation process including automatic speaker-change detection (SI: speaker independent; SA: speaker adapted).

入力音声に対しては、まず不特定話者用（SI）GMMとそれまでの話者に適応化した（SA）GMMに対する尤度を求める。もしSA GMMに対する尤度の方が大きければ、同じ話者が継続していると判断し、その話者に適応化したSA HMMで音声認識する。そのデコーディング結果と入力音声を用いて、音素HMMとGMMをさらに話者適応化する。もしSA GMMよりもSI GMMに対する尤度の方が大きければ、話者が交代したと判断し、SI HMMで音声認識する。そのデコーディング結果と入力音声を用いて、新たなSA HMMとSA GMMを作る。前述の実験と同様に、尤度比較するモデルとして、最新の2人の話者と不特定話者の三種類のモデルを使うことにした。

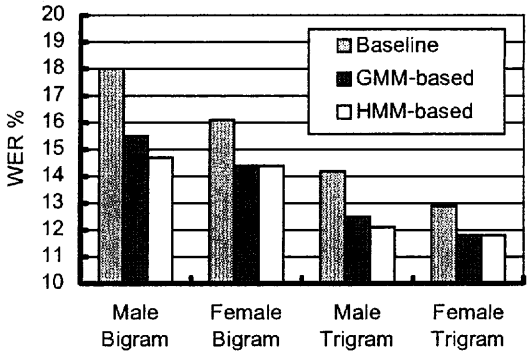

Figure 4:Word error rates by baseline and HMM-based, GMM-based methods.

実験結果を図4に示す。図には、適応化を行う前(Baseline)、HMMによって話者交代を検出する方法、GMMによって話者交代を検出する方法による単語誤り率を示す。いずれの評価セットに対しても、適応化により誤り率が低下している。"HMM"法に比べ、"GMM"法でも、女性では性能の低下はなく、男性でもわずかの低下ですむことがわかる。"GMM"法により、適応化前に比べて誤り率は男女平均で10.0%低下している。

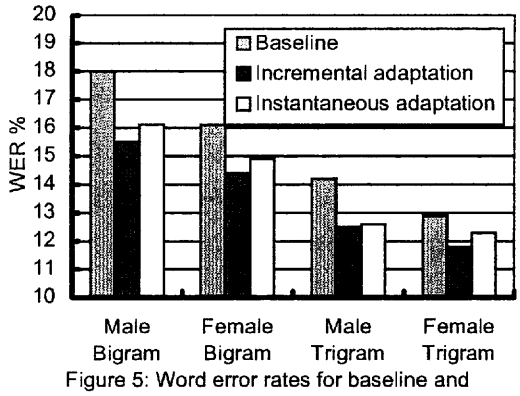

Figure 5: Word error rates for baseline and incremental/instantaneous adaptation.

次に、逐次的に適応を行わず、入力音声ごとに不特定話者HMMを用いてオンライン即時適応を行ったときの比較実験を行った。実験結果を図5に示す。図には、適応化を行う前(Baseline)、即時適応法("Instantaneous")、提案した逐次適応法("Incremental")による誤り率を示す。いずれの条件においても、即時適応法より逐次適応法の方が、誤り率が低下することが確認された。

2.3 話者クラスタによる初期モデル選択法

ここまで述べた手法には不特定話者モデルを

初期モデルとして用いたが、話者クラスタに基づく初期モデルを用いた話者適応の方が有効だと考えられる。話者クラスタによる話者適応法は、各クラスタに対応するモデルから、入力音声に対して尤度最大となるモデルを選ぶだけで済み、パラメータ変換の必要がない。選ばれたモデルをそのまま用いて認識するか、或いは幾つかのモデル間の内挿によって新しいモデルを構築して認識に用いる。この手法の問題点として、認識する際、計算量が膨大で、必ずしも適切でないモデル選択をする恐れがある。

ここでは、話者クラスタによって作成したモデルを用いる話者適応法について種々の検討を行った。計算量を削減するために、HMM でなく GMM の尤度比較による方法を検討した。

話者クラスタ化は、各特定話者モデル間の距離に基づいて行う。SPLIT 法 7)で用いられたクラスタリングアルゴリズムを用いた。この手法は一般的な LBG 法とは異なり、歪みが最大となるクラスタを順次分割するため、任意の数のクラスタが作成できる。

クラスタリングする前に各話者間の距離行列を作成する。あらかじめ尤度の閾値或いはクラスタ数を指定すれば自動的にクラスタリングの結果が得られる。SI HMM の学習に使われるのと同じ男性 53 名、女性 56 名の話者からなるデータを用いた。

まず女性テストセットを用いて各クラスタ数における認識実験を行った。各クラスタに属するあらゆる話者のデータを用いて、Baum-Welch アルゴリズムによる連結学習法で各クラスタの HMM モデルを構築する。認識する際、まず SI HMM を用いて尤度を計算すると同時に、認識結果としてのラベルファイルを作成する。次に、各クラスタの HMM モデルを用いてこのラベルファイルに対しリスコアリングして、各クラスタの尤度を計算する。SI HMM と各クラスタの尤度から最大尤度を示す HMM モデルを選んで、認識を行う。この手法により尤度最大化が保証される。

実験結果を図 6 に示す。図には、各クラスタ数における単語誤り率を示す。クラスタ数が 4 の場合に最も良い結果が得られる。

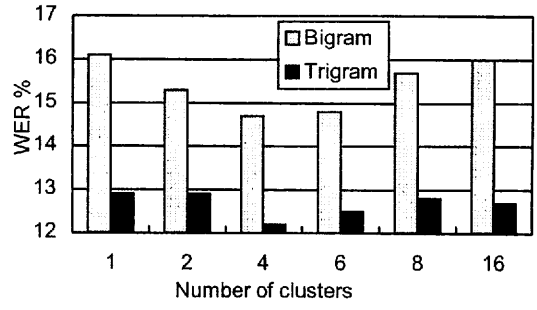

Figure 6: Word error rates for various number of clusters (female test data).

前述した手法では、モデル選択の際、SI HMM による認識を行った後に各クラスタの HMM でリスコアリングすることが必要なので、膨大な計算量がかかる。この問題を対応するために GMM を尤度比較に用いることを試みた。このため、まず HMM の学習と同じデータで不特定話者と各クラスタの GMM を構築する。入力音声に対して、不特定話者と各クラスタの GMM に関する尤度を計算する。その中の最大尤度を示す GMM モデルに対応する HMM を選んで認識を行う。

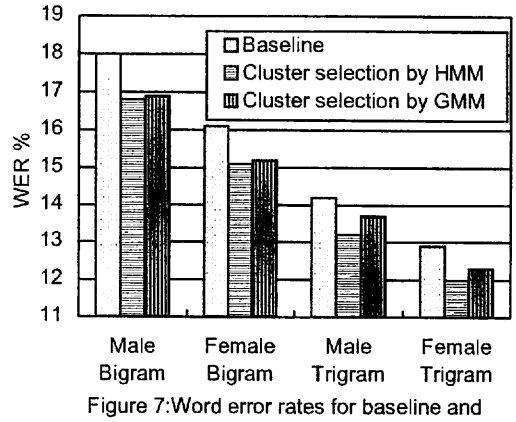

Figure 7:Word error rates for baseline and GMM/HMM-based cluster selection methods.

実験結果を図 7 に示す。4 クラスタの場合の男女別の実験結果を示す。図には、適応化を行う前(baseline)、HMM の尤度比較、提案する GMM の尤度比較による単語誤り率を示す。いずれの評価セットに対しても、適応化により誤り率が低下していることが分かる。"HMM"法により、適応化前に比べて誤り率は男女平均で 7.0%低下している。"GMM"法により、適応化前に比べて誤り

率は男女平均で4.1%低下している。

次に、オンライン逐次適応を行った。まずGMMによってクラスタを選ぶ。SI HMMの代わりに尤度最大のクラスタのHMMを初期モデルとして、オンライン逐次適応を行った。実験結果を図8に示す。図には、適応化を行う前(baseline)、不特定話者モデルを初期モデルとしての逐次適応法(1 cluster)、提案したモデル選択法(4 clusters)による誤り率を示す。モデル選択法により誤り率が低下することが確認された。

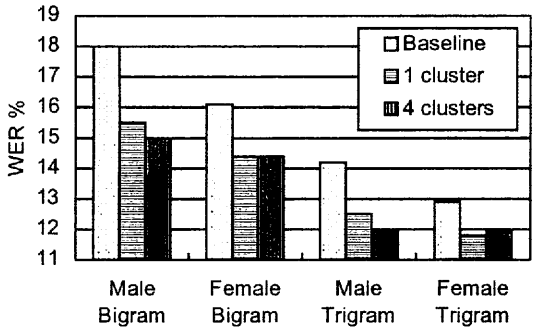

Figure8: Word error rates for baseline and incremental adaptation with 1/4 clusters.

2.4 MLLRにおける適応データ量に応じたクラスタ数選択法

MLLR、MAP及びVFS法を併用する話者適応化における音素クラスタ数の決定法について検討した。上述のように、MLLRにおいて、適応用の入力音声として一文を用いた場合でも、音素クラスタごとに変換を行う方が単一変換行列を用いるより性能が良い。一文ずつ用いる教師なし逐次オンライン適応の場合、7クラスタの条件で一番良い結果が得られている。適応データが更に増加すれば最適クラスタ数も変動すると考えられる。一般に、モデルの自由パラメータ数の多い複雑なモデルを用いると、データ量が少ない時に性能が悪化するが、モデルの自由パラメータ数の少ない簡単なモデルを用いると、データ量が多い時に性能が低く抑えられてしまう。そこで適応データ量と適応による改善効果を考察し、MDL基準により適応データ量に応じてクラスタ数を決定する手法を検討した。

MDL(minimum description length)基準 8)はモデルの最適化によく利用される基準である。この基準はデータに対し、最適な確率モデルを選択する方法として有効である。MDL基準では、モデル $i = 1, \ldots, M$ のうち、データ $X^N = \{x_1, \ldots, x_N\}$ の記述長を最小にするモデルが最適なモデルであると考える。記述長は以下の式で表される。

$$l^{(i)} = -\log P_{\theta(i)}(X^N) + \frac{k_i}{2}\log N + \log M$$

ここで、k_iはモデルiの次数、$P_{\theta(i)}(X^N)$はデータX^Nに対するモデルiのパラメータ$\theta(i)$の最尤推定量である。第一項はデータに対する対数尤度を記号反転させたもの、第二項はモデルの複雑さを表す量である。モデルが複雑になるほど、第一項は小さくなり、第二項は大きくなる。このように両者の間にトレードオフがあり、ある適当な複雑さを持ったモデルが最適なモデルとして選択される。

適応データ量に応じてMLLRのクラスタ数を決定するために、モデルの次数k_iとしてクラスタ数の対数$\log C$を用い、パラメータαを加えてMDLの式を次のように修正する。

$$l^{(i)} = -\log P_{\theta(i)}(X^N) + \alpha \times \frac{\log C}{2}\log N + \log M$$

あるNでαの値を実験的に決めれば、$l^{(i)}$の最小化によって最適なクラスタ数Cが決定される。

実際に放送されたニュース音声から、評価用として、スタジオで収録された女性のクリーンな発話18文を抽出した。適応データは同じ話者の1000文余りを用いた。

まず1クラスタの場合について適応文の数を変えて、MLLRの性能を考察した。結果は図9に示す通りである。図にはSD(特定話者)モデルの結果を合わせて示す。MLLRを用いるとき、適応文数が100の場合に一番良い結果が得られることが分かる。100文を超えると適応データを増やしても性能が上がらず、かえって下がる。この

ことは、クラスタ数が少ないときのMLLRのような変換手法の限界を示していると考えられる。

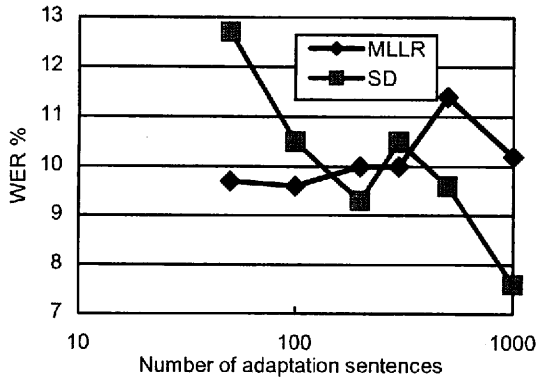

Figure 9: Word error rates for various sizes of adaptation data.

従って適応データ量によって最適なクラスタ数を決める必要がある。

適応文数を変えた時に最適なクラスタ数が達成できるαの値を実験的に求めた。その結果、MDL基準により、各適応文数の条件で最適なクラスタ数を求めると、図10の中の星印に示す結果が得られた。図には各適応文数の条件で各クラスタ数における単語誤り率の二次近似曲線を示してある。MDL基準によって最適条件に近いクラスタ数が得られることがわかる。

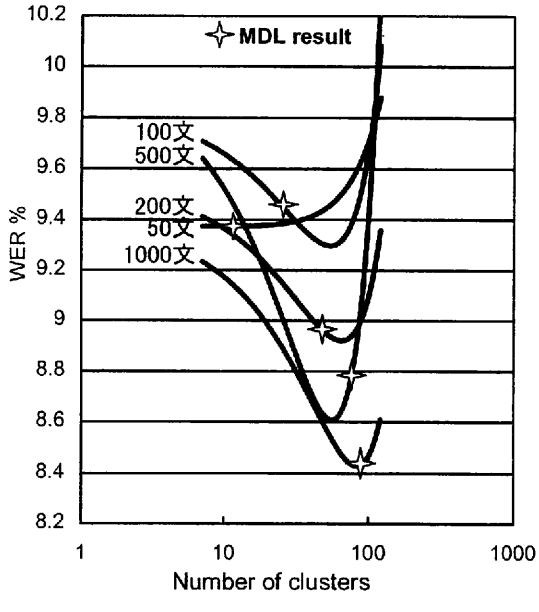

Figure 10: Word error rates as a function of the number of clusters.

3. 雑音適応

3.1 雑音適応手法

ニューラルネットワークを用いたHMMの雑音適応について検討した。またHMM合成法とMLLRなどの手法との比較を行った。

提案した手法では、HMMの各状態の出力確率分布に対し、ニューラルネットワークを用いて雑音に適応した値を推定する非線形演算を行う。学習時に雑音HMM、音声HMM、及び雑音情報を含むHMM(目標HMM)が必要である。音声HMMは,クリーンな環境の下で学習された多数話者の音声から作成した不特定話者HMMを使用する。雑音HMMは,各雑音重畳音声データから,音声を含まない部分を雑音データとして抽出して作成する。ニューラルネットワークの構造を図11に示す。

HMM合成法(PMC;parallel model combination)は、音声のHMMと雑音のHMMの畳み込みをスペクトル空間で行って、雑音が重畳した音声のHMMを合成するものである。従って、HMMパラメータをスペクトル空間に変換する必要があるため、単純にはCMS(Cepstral mean subtraction)を用いたHMMを合成法に利用することができない。

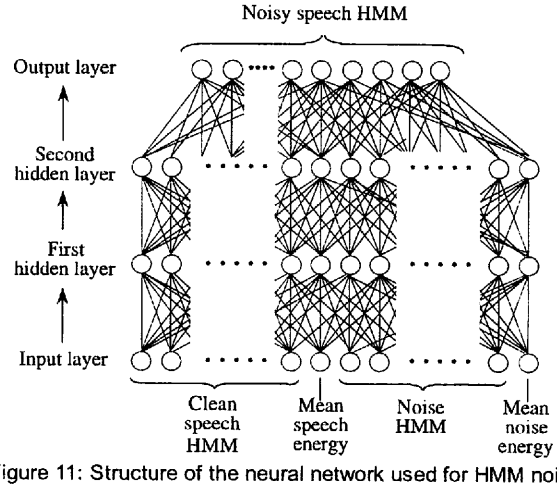

Figure 11: Structure of the neural network used for HMM noise adaptation.

MLLRはHMMのガウス分布の平均値を尤度最大化の基準に基づいた線形変換により適応化する方法で、話者適応によく用いられる手法であるが、雑音の適応にも有効である。前述したニューラルネットワーク法とHMM合成法には雑音

HMM が必要になるが、MLLR の場合は雑音 HMM が要らない。

評価用データは 1996 年 7 月に実際に放送されたクリーンなニュース音声に，人工的に雑音を付加させるものである。一人の男性話者による 14 文を抽出し，うち 4 文をニューラルネットワーク法における目標 HMM の作成用データ，残り 10 文はすべての手法において評価用データとして使用した。

3. 2 認識実験結果

雑音 HMM は各雑音データより Baum-Welch アルゴリズムを用いて作成した。1 状態 1 混合とし音声 HMM と同様の音響特徴量を用いた。目標 HMM は MLLR-MAP 及び VFS の手法を併用し，音声 HMM の平均ベクトルを各雑音重畳音声データに適応させることで作成した。よって音響特徴量，モデルの状態数は音声モデルと同一で，混合分布の重みを修正せずに各雑音に適応した HMM を得ることができる。なお、CMS を適用している。

以上より得られる各種 HMM を用いて，音声 HMM 及び雑音 HMM の 16 次の LPC ケプストラムとそれぞれの平均パワーを入力，目標 HMM の 16 次元の LPC ケプストラムを出力とするニューラルネットワークを学習した。

二種類の雑音（人ごみと展示場）の雑音条件で適応実験を行った。各SN比における認識結果を図12, 13に示す。MLLR法で教師なしのオンライン適応を行った結果も併せて示してある。ほとんどの場合、ニューラルネットワーク法の方がMLLR法より性能が良いことが分かる。

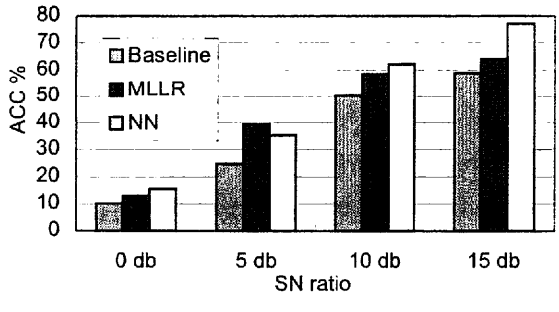

Figure 12: Word accuracy for baseline, MLLR and NN methods.
("hitogomi" noise)

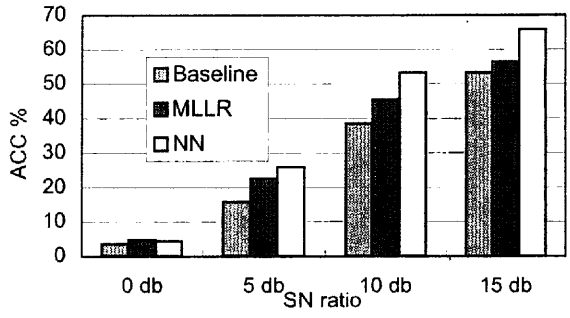

Figure 13: Word accuracy for baselin,MLLR and NN methods
("tenji" noise).

図 14, 15 に HMM 合成法を用いた各 SN 比における適応効果を示す。雑音 HMM はニューラルネットワーク法と同じ構造である。HMM 各状態のガウス分布の平均値だけを適応した。本手法では CMS を用いていないことから、ベースラインの認識率がニューラルネットワーク法より劣っている。そのため、適応後の結果もニューラルネットワーク法の方が高い認識率を示した。

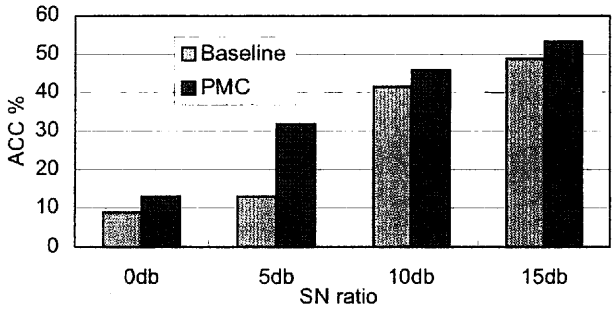

Figure 14: Word accuracy for Baselin and PMC method ("hitogomi" noise).

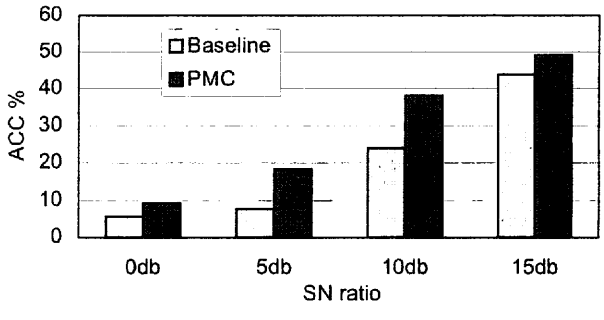

Figure 15: Word accuracy for baselin and PMC method ("tenji" noise).

4. まとめ

音声認識における耐性向上を目指して、音

響モデルすなわち音素 HMM の話者適応と雑音適応に関する研究を進めた。話者適応に関して、まず尤度比較による話者交代検出に基づく教師なしオンライン即時・逐次型適応化法を検討した。音素クラスタを用いる MLLR-MAP-VFS 法で単語誤り率が、男女平均で約 12%減少することを確認した。計算量を削減するために不特定話者用混合ガウス分布モデル(GMM)と、話者に適応した GMM に対する入力音声の尤度比較によって話者交代を検出する手法を提案した。複数の HMM セットを用いる方法に比べて大幅に計算量を削減しながら、わずかの性能低下ですみ、適応化前に比べて単語誤り率が男女平均で約 10%減少することを確認した。さらに即時適応法との比較によって逐次適応法の効果が確認された。次に、話者クラスタによって作成した初期モデルを用いる話者適応法について検討した。計算量を削減するために混合ガウス分布モデル(GMM)の尤度比較によるクラスタ選択方法を提案した。適応化実験によって提案する手法の有効性が確認された。さらに、MLLR における適応データ量と改善効果の関係を考察した。MDL 基準によって適応データ量に応じたクラスタ数を決定する手法を提案した。MLLR のクラスタに適用するため MDL 基準を修正し、パラメータを追加した。実験によって提案した手法の有効性が確認された。

雑音適応に関しては、ニューラルネットワークを用いた雑音適応の手法を提案した。HMM 合成法、MLLR 法との性能比較を行った。実験結果からニューラルネットワーク法はほとんどの条件で優れた適応性能を有することがわかった。ニューラルネットワークによる方法が HMM 合成法よりも高い認識性能を示す主たる理由は、HMM 合成法では CMS を適用できないために、ベースラインの性能が劣化していることにある。これに対処するために拡張HMM合成法 9)が提案されているが、計算量が大きくなる問題がある。今後はこのような手法も比較の対象として検討を進めていきたい。

音声認識における耐性向上は極めて重要な研究課題であり、音声認識が実際の場で広く実用化されるためには、ハンズフリー入力を前提とした研究などの一層の推進が必要である。また、これまでの音声認識研究では、書き言葉の言語モデルや、書き言葉を読み上げた音声から作成した音響モデルを主として用いてきたが、話し言葉をターゲットとした大規模コーパスの構築と、それに基づくモデル化の研究の推進が必須である。

参考文献

1) C. J. Leggetter et al., Computer Speech and Language, 9, pp.171-185 (1995)
2) J.-L. Gauvain et al., IEEE Trans. Speech and Audio Processing, 2, 2, pp. 291-298 (1994)
3) F. Martin et al., 信学技報, SP92-96 (1992)
4) 大倉 他, 信学論, J76-D-II, 12, pp.2468-2476 (1993)
5) 桜井 他, 音学春季講論, 2-1-3 (1999)
6) 松井 他, 信学論, J77-A, 4, pp. 601-606 (1994)
7) 管村 他, 信学技報, S82-64 (1982)
8) K. Shinoda et al., Journal of ASJ (E), 19, 2, pp.79-86 (2000)
9) 南 他, 信学論, J80-A, 7, pp.1179-1186 (1997)

発表文献

[1] Z. P. Zhang and S. Furui, "On-line incremental speaker adaptation with automatic speaker change detection", *Proc. ICASSP2000*, pp. 961-964 (2000)
[2] Z. P. Zhang and S. Furui, "An online incremental speaker adaptation method using speaker-clustered initial models", *Proc. ICSLP2000*, pp. III-694-697 (2000)
[3] S. Furui and D. Itoh, "Noise adaptation of HMMs using neural networks", *Proc. ISCA ITRW ASR2000*, pp. 160-167 (2000)
[4] 張 志鵬, 古井 貞熙, "MLLRにおける適応データの量に応じたクラスタ数の選択法", 音学秋季講論, 1-5-8 (2000)