社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－ＳＬＰ－69（15）
2007／12／20

# 十分統計量を用いた教師なし話者適応における話者選択法

谷 真宏† 江森 正† 大西 祥史† 越仲 孝文† 篠田 浩一‡

†日本電気株式会社 中央研究所 〒211-8666 川崎市中原区下沼部 1753
‡東京工業大学 情報理工学研究科 〒152-8552 東京都目黒区大岡山 2-12-1
E-mail: m-tani@bu.jp.nec.com

**あらまし** 十分統計量を用いた教師なし話者適応において，選択する話者の数を決定する手法を提案する．音声認識における高速な教師なし話者適応の一つとして，話者毎の十分統計量を用いた手法が提案されている．これは，予め用意した複数の話者の中から，評価話者に音響的な特徴が近い話者を選択し，選択された話者の十分統計量を用いて，評価話者に適応した音響モデルを構築する手法である．従来手法では，評価話者に音響的な特徴が近い話者を選択する際，複数の話者の中から，予め定められた数だけ選択する．提案手法では，評価話者と予め用意した話者との音響特徴量空間における話者間距離を基準に，選択する話者の数を決定する．電話による対話音声を用いた認識実験において，従来手法に比較し，単語正解精度が 0.74 ポイント向上した．特に，音響的な特徴が近い話者が少ない評価話者に対して有効であることを確認した．
**キーワード** 教師なし話者適応，十分統計量，話者選択

# Speaker Selection for Unsupervised Speaker Adaptation based on HMM Sufficient Statistics

Masahiro TANI† Tadashi EMORI† Yoshifumi OHNISHI† Takafumi KOSHINAKA† and Koichi SHINODA‡

†Central Research Laboratories, NEC Corporation
1753, Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa, 211-8666 Japan
‡Graduate School of Information Science and Engineering, Tokyo Institute of Technology
2-12-1, Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: m-tani@bu.jp.nec.com

**Abstract** We propose a new speaker selection method for the unsupervised speaker adaptation based on HMM sufficient statistics. The adaptation technique of using HMM sufficient statistics has been proposed as one of the rapid unsupervised speaker adaptation techniques in speech recognition. The procedure is as follows: First the training speakers acoustically close to the test speaker are selected. Then, the acoustic model is trained using the HMM sufficient statistics of these selected training speakers. In this technique, the number of selected training speakers is always constant. In our proposed speaker selection method, the number of speakers is determined by the distances between the test speaker and each training speaker. In our recognition experiments using spoken dialogue data, the proposed method improved word accuracy by 0.74 points. It was confirmed that the proposed method particularly effective when there are not many training speakers around the test speaker in acoustic space.
**Keyword** Unsupervised Adaptation, HMM Sufficient Statistics, Speaker Selection

## 1. ま え が き

音声認識における高速な教師なし話者適応の一つとして，話者毎の十分統計量を用いた手法が提案されている[1]．これは，予め用意した複数の話者（学習話者）の中から，評価話者に音響的な特徴が近い話者を，予め定められた数だけ選択し，選択された話者に対応する十分統計量を用いて評価話者に適応した音響モデル（適応モデル）を構築する手法である．MLLR（Maximum Likelihood Linear Regression）[2]や MAP（Maximum A Posteriori）推定[3]等の手法に比較し，適応のための発声データが極少量で良い点，そして，予め計算した十分統計量を用いるため，高速に適応モデルを構築できる点が本手法の特長である．しかしながら，評価話者の音声そのものを用いて適応モデルを構築しないため，予め用意した学習話者の中から，評価話者に適合する話者だけを正確に選択することが重要となる．

本稿では，評価話者毎に，選択する話者の数（選択話者数）を適切に決定することで，話者適応による認識性能の向上を目指す．提案手法では，評価話者毎に選択話者数を決定し，予め用意した話者の中から，音響的な特徴が近い話者のみを選択する．音響的な特徴が近い話者のみの発声データを用いて適応モデルを構築すること，言い換えると，音響的に似ていない話者の発声データは用いずに適応モデルを構築することは，認識性能の向上の面からも重要であると考えられる．本稿では，電話による対話音声データベースを用いた認識実験により，提案手法の有効性について検証する．

本稿の構成は次の通りである．第2章で十分統計量を用いた教師なし話者適応の概要と問題点について述べる．第3章で提案する話者選択法について述べる．第4章で認識実験結果を示す．この結果を踏まえ，第5章で考察し，第6章でまとめる．

## 2. 十分統計量を用いた教師なし話者適応

### 2.1. 概要

本節では，提案手法のベースとなる十分統計量を用いた教師なし話者適応について説明する[1]．本従来手法は，話者毎の音響的特徴を表すGMMにより選択された十分統計量を用いた話者適応法であり，基本的には次の4ステップから構成される．以下，話者毎の音響的な特徴を表す統計モデル（ここでは，GMM）を話者モデルと呼ぶ．第1ステップでは，話者毎に十分統計量が計算される．また，話者毎に話者モデルが学習される．第2ステップでは，第1ステップで予め学習された話者モデルを用い，評価話者の入力音声に対する尤度が計算される．尤度は話者モデル毎に計算される．第3ステップでは，尤度が高い話者モデルに対応する話者と評価話者の音声には，音響的に似た性質があると仮定し，複数の話者の中から，尤度が高い話者モデルに対応する話者が，予め定められた数だけ選択される．第4ステップでは，第3ステップで選択された話者に対応する十分統計量を用い，評価話者に適応した音響モデルが構築される．

従来手法の特に第4ステップを改良した手法がいくつか提案されている．例えば，選択された話者毎の十分統計量を重み付け統合する手法[4]や，不特定話者の十分統計量と選択された話者の十分統計量を線形補間する手法[5]が提案されている．

しかしながら，いずれの手法も第3ステップにおける選択話者数を決定する仕組みはない．つまり，選択話者数は，評価話者に依らず固定されている．

### 2.2. 選択話者数固定の問題点

従来手法において，評価話者に依らず選択話者数が固定されていることによる問題点について検討する．

選択話者数が固定の従来手法は，理想的な環境において，大きな適応効果が期待できる．理想的な環境とは，予め用意した話者1名あたりの発声データが大量に存在すること，そして，評価話者の入力音声と音響的に似た性質を持つ話者が用意した話者の中に必ず存在し，その話者を正確に選択できることである．このような環境では，極少数の話者を選択すれば，精度の良い適応モデルが構築され，話者適応の効果は大きくなると考えられる．

しかしながら，実際の利用場面に即した環境では，話者数は多いが話者1名あたりの発声データを大量に用意できない場合が多い，また，評価話者の音声と音響的に似た性質を持つ話者が存在するとは限らない．このような環境では，複数の話者を選択することでデータ量は確保できる．一方で，評価話者の音声と音響的に似た性質を持つ話者が殆ど存在しない場合がある．その結果，選択話者数が固定の従来手法では，音響的に性質の似ていない話者まで選択され，適応モデルの精度が劣化し，話者適応の効果は小さくなると考えられる．

## 3. 話者選択法

本稿では，前述した問題点を解決するために，評価話者毎に選択話者数を決定する方法を検討する．この話者選択法の基本的な考え方は，評価話者の音声と音響的に似た性質を持つ話者が多数存在する場合は，多くのデータ量を確保するように，選択話者数を大きくし，少数しか存在しない場合は，必要最低限のデータ量を確保しながら，音響的に性質の似ていない話者は可能な限り選択しないように，選択話者数を小さくするものである．

### 3.1. 話者間距離

まず，評価話者と予め用意した話者の音声が，音響的に性質が似ているか否かの尺度として，話者間距離を定義する．話者間距離は，評価話者と予め用意した話者との音響特徴量空間における統計的モデル間距離である．統計モデルは，評価話者については，入力音声を用いて学習され，予め用意した話者については，話者毎の発声データを用いて学習される．統計モデルは，多次元正規分布とする．具体的には，話者間距離$d$を式（1）により定義する．

$$d=\sum_{i=1}^{D}\frac{\left(\mu_i^p-\mu_i^q\right)^2}{\sigma_i^p\sigma_i^q} \quad (1)$$

ここで，$\mu^p$及び$\sigma^p$は，それぞれ，評価話者$p$に対応する多次元正規分布の平均及び分散であり，$\mu^q$及び$\sigma^q$は，それぞれ，予め用意した話者$q$に対応する多次元正規分布の平均及び分散である．また，$D$は，特徴量の次元数である．$d$の値が小さいことは，話者$p$と$q$の音声の音響的な性質が似ていることを表し，$d$の値が大きいことは，評価話者$p$と用意した話者$q$の音声の音響的な性質が似ていないことを表す．

次に，評価話者の音声と音響的に似た性質を持つ話者が多数存在するか否かの尺度として，平均話者間距離を定義する．平均話者間距離は，評価話者と予め用意した全ての話者の話者間距離$d$の平均値である．具体的には，平均話者間距離$\bar{d}(p)$を式（2）により定義する．

$$\bar{d}(p)=\frac{1}{N_{\mathrm{all}}}\sum_{q=1}^{N_{\mathrm{all}}}\sum_{i=1}^{D}\frac{\left(\mu_i^p-\mu_i^q\right)^2}{\sigma_i^p\sigma_i^q} \quad (2)$$

ここで，$\mu^p$，$\sigma^p$，$\mu^q$及び$\sigma^q$は，式（1）と同様，それぞれ，評価話者$p$に対応する多次元正規分布の平均，分散，予め用意した話者$q$に対応する多次元正規分布の平均及び分散である．$D$は，特徴量の次元数である．$N_{\mathrm{all}}$は，予め用意した話者の総数である．$\bar{d}(p)$の値が小さいことは，評価話者$p$の音声と音響的に似た性質を持つ話者が多く，音響特徴量空間において，予め用意した話者が評価話者$p$の近傍に密集していることを表す．このような状態を密の状態と呼ぶことにする．一方，$\bar{d}(p)$の値が大きいことは，評価話者$p$の音声と音響的に似た性質を持つ話者が少なく，音響特徴量空間において，予め用意した話者が評価話者$p$から散在していることを表す．このような状態を疎の状態と呼ぶことにする．

### 3.2. 選択話者数の決定

提案する話者選択法では，従来手法で固定されている選択話者数を，平均話者間距離$\bar{d}(p)$を用いて，評価話者毎に決定する．この方法は，$\bar{d}(p)$の値が小さい（密の状態にあるような）話者に対しては，多くのデータ量を確保するように，選択話者数を$N_{\max}$まで大きくし，$\bar{d}(p)$の値が大きい（疎の状態にあるような）話者に対しては，必要最低限のデータ量は確保するように，選択話者数を$N_{\min}$まで小さくするように選択話者数を決定する．$N_{\max}$は，選択話者数の上限値であり，予め用意した話者の総数とする（$N_{\max}=N_{\mathrm{all}}$）．$N_{\min}$は，選択話者数の下限値であり，必要最低限のデータ量を確保するために必要な選択話者数とする．ここでは，上限値$N_{\max}$と下限値$N_{\min}$をシグモイド関数で補間する（式（3））．即ち，式（4）に示す平均話者間距離$\bar{d}(p)$の関数により選択話者数$N$を決定する．ここで，$k$は，$\bar{d}(p)=a$の点での勾配である．

$$\frac{N_{\max}-N\left(\bar{d}(p)\right)}{N_{\max}-N_{\min}}=\frac{1}{1+\exp\left(-k\left(\bar{d}(p)-a\right)\right)} \quad (3)$$

$$N\left(\bar{d}(p)\right)=N_{\max}-\frac{N_{\max}-N_{\min}}{1+\exp\left(-k\left(\bar{d}(p)-a\right)\right)} \quad (4)$$

## 4. 実　　験

### 4.1. 実験条件

評価実験には，電話による対話音声データベースを利用する．対話音声データベースは1075名（約54時間）の音声データより構成され，話者1名あたりの発声時間は3分である．話者毎に発声内容は異なっている．学習セットには1015名の音声データを用いる．学習セットは，男性話者716名と女性話者359名の音声データから成る．評価セットには学習セットに含まれない60名の音声データを用いる．評価セットは，男性話者34名と女性話者26名の音声データから成る．

特徴量には，分析周期10msで抽出した，12次元のMFCCとこの一次差分（Δ）及び二次差分（ΔΔ）量，パワーの一次差分及び二次差分量，ピッチとこの一次差分量を用いる（計40次元）．特徴量抽出後，CMN（Cepstrum Mean Normalization）を行う．

音響モデルには，評価話者60名を除いた1015名の音声データを用いて構築した，性別依存の状態共有HMM（3000状態，32混合）を用いる．言語モデルには，対話音声の書き起こしにより構築した，trigram言語モデルを用いる．

### 4.2. 話者間距離の分析

評価話者毎に平均話者間距離$\bar{d}(p)$の傾向を確認するため，評価セットについて調査した．具体的には，評価セットにおける平均話者間距離$\bar{d}(p)$の度数分布を求めた．度数分布を図2に示す．図2より，評価話者毎に平均話者間距離にばらつきがあることが確認で

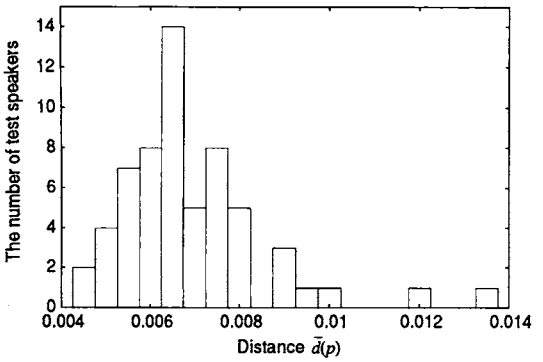
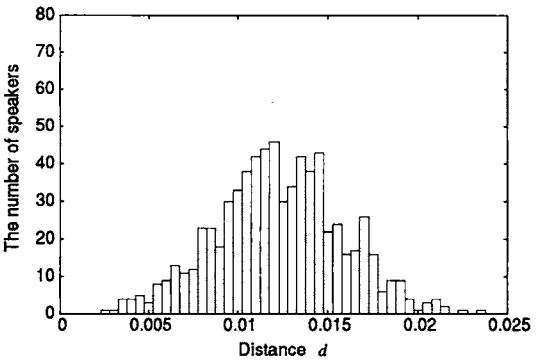
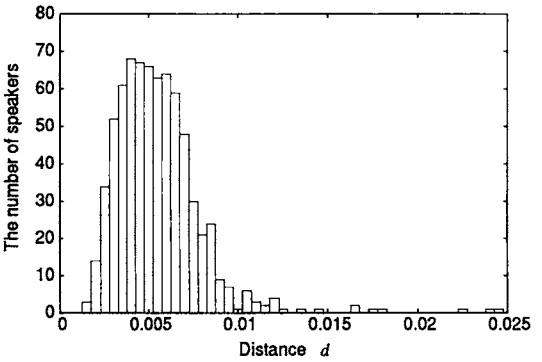



図 2　平均話者間距離の度数分布




(a)　評価話者 A：疎の状態




(b)　評価話者 B：密の状態

図 3　話者間距離の度数分布の一例

きる．平均話者間距離が大きい（図 2 の右側に位置する）評価話者は，音響的に似た性質を持つ話者が少なく，平均話者間距離が小さい（図 2 の左側に位置する）評価話者は，音響的に似た性質を持つ話者が多いと考えられる．これを確認するため，評価話者と予め用意した話者との話者間距離 $d$ の度数分布を求めた．度数分布の一例を図 3 に示す．図 3 は，ある評価話者から一定距離の区間に，用意した話者が何名存在するかを表している．図 3（a）に示す評価話者 A は，$\bar{d}(p_A)=0.012$ であり，（b）に示す評価話者 B は，$\bar{d}(p_B)=0.0053$ である．図 3 より，平均話者間距離が大きい評価話者は，疎の状態にあることが確認できる（図は割愛するが，平均話者間距離が大きい A 以外の評価話者についても同様の傾向を確認している）．また，平均話者間距離が小さい評価話者は，密の状態にあることが確認できる（平均話者間距離が小さい B 以外の評価話者についても同様の傾向を確認している）．

次に，選択話者数と認識性能の関係を評価セットについて調査した（図 4）．図 4 の×印 1 つは，評価話者 1 名に対応する．縦軸は，選択話者数である．選択話者数を数点振って認識実験し，認識率が最も高くなった点をプロットしている．横軸は，評価話者と用意した話者との平均話者間距離 $\bar{d}(p)$ である．

図 4 より，平均話者間距離 $\bar{d}(p)$ の値が大きい（疎の状態にある）話者に対しては，選択話者数を小さく設定した方が良いことが確認できる．一方，$\bar{d}(p)$ の値が大きい（密の状態にある）話者に対しては，疎の状態ほどの明瞭な傾向は出ていないものの，選択話者数を大きく設定した方が良いことが確認できる．

図 4 の破線は，従来手法である $N=300$（選択話者数固定）を表している．$N$ の値は，評価セットを用い，後述の認識実験で認識率が最も高い 300 名である．図 4 の実線は，提案手法である式(4)を表している．$N_{\min}$，$k$，$a$ の値は，同様に評価セットを用い，後述の認識実験で認識率が高くなるように決定した．それぞれ，$N_{\min}=133$，$k=1025$，$a=0.0089$ とした．

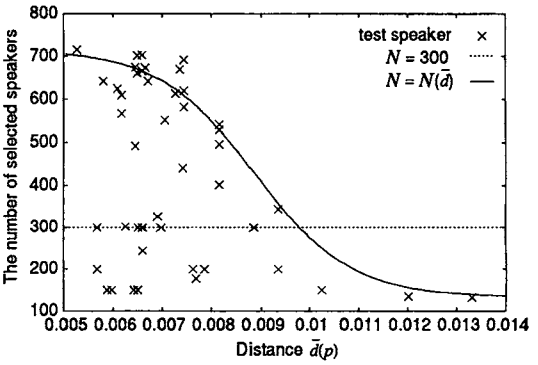



図 4　選択話者数と認識性能の関係

表 1　全評価話者 60 名の認識率.

| | Percent correct [%] | Word accuracy [%] |
|---|---|---|
| no adaptation | 64.06 | 52.78 |
| fix $N$ | 64.48 | 52.89 |
| best $N$ | 65.23 | 54.17 |
| $N(\bar{d})$ | 64.67 | 53.63 |

表 2　疎の状態にある評価話者 4 名の認識率.

| | Percent correct [%] | Word accuracy [%] |
|---|---|---|
| no adaptation | 59.13 | 50.53 |
| fix $N$ | 60.74 | 52.32 |
| best $N$ | 62.11 | 53.62 |
| $N(\bar{d})$ | 61.92 | 53.19 |

### 4.3. 認識実験

4.1.節で述べた評価セットを用い，認識実験を行った．全評価話者 60 名の認識率を表 1 に，疎の状態にある評価話者 4 名の認識率を表 2 に示す．この評価話者 4 名は，平均話者間距離 $\bar{d}(p)$=0.013，0.012，0.010，0.0093 である話者（図 1 の平均話者間距離が大きい右から 4 名）とする．認識率は，単語正解率（Percent correct）及び単語正解精度（Word accuracy）である．ここで，「no adaptation」は，話者適応はせず，性別依存の音響モデルを用いた認識率(性別は既知)である．「fix $N$」は，選択話者数を固定した場合の認識率である．選択話者数を 10，50，100，150，200，300，全話者（男性は 716 名，女性は 359 名）として，それぞれ認識実験を行い，認識率が最も高い 300 名の結果を示している．「best $N$」は，上記選択話者数の中から，評価話者毎に最も認識率が高い選択話者数を採用した場合の認識率である．「$N(\bar{d})$」は，平均話者間距離 $\bar{d}(p)$ を用いて，式（4）により，選択話者数を決定した場合の認識率である．式（4）では，$N_{\min}$=133，$k$=1025，$a$=0.0089 とした．

表 1，表 2 の結果より，ベースラインとなる選択話者数固定「fix $N$」と比較し，「$N(\bar{d})$」の方が，単語正解率，単語正解精度ともに向上していることが分かる．

## 5. 考　　察

平均話者間距離 $\bar{d}(p)$ を用いて選択話者数を決定する手法は，選択話者数固定のベースラインと比較し，全評価話者に関し，単語正解率で 0.19 ポイント（64.48%から 64.67%），単語正解精度で 0.74 ポイント（52.89%から 53.63%）向上しているが，疎の状態にある評価話者に関しては，単語正解率で 1.18 ポイント（60.74%から 61.92%），単語正解精度で 0.87 ポイント（52.32%から 53.19%）向上している．従って，平均話者間距離を用いて選択話者数を決定する手法は，疎の状態にある評価話者，即ち，予め用意した話者の中に，音響的な特徴が近い話者が少ない評価話者に対して特に有効であると考えられる．評価話者毎に最も認識率の高い選択話者数を設定した「best $N$」の結果と比較しても，全評価話者に関しては，まだ，認識率に差はあるものの，疎の状態にある評価話者に関しては，認識率の差が小さい．

## 6. ま　と　め

本稿では，十分統計量を用いた教師なし話者適応において，選択する話者の数を決定する手法を提案した．本提案手法は，評価話者と予め用意した話者との音響特徴量空間における話者間距離を基準に，選択する話者の数を決定する．電話による対話音声を用いた認識実験において，選択話者数固定の従来手法に比較し，単語正解精度が 0.74 ポイント向上した．特に，疎の状態にある評価話者，即ち音響的な特徴が近い話者が少ない評価話者に対して有効であることを確認した．

今回，提案手法では，多次元正規分布を用いて話者間距離を計算したが，今後 GMM を用いて距離計算を行う予定である．また，話者間距離以外の選択話者数の制御パラメータを検討する予定である．

## 文　　献


[1] 芳澤伸一，馬場朗，松浪加奈子，米良祐一郎，山田実一，李晃伸，鹿野清宏，“十分統計量と話者距離を用いた音韻モデルの教師なし学習法，” 信学論(D), vol.J85-D, no.3, pp.382-389, Mar. 2002.

[2] C.J. Leggetter and 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.

[3] J.-L. Gauvain and C.-H. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains,” IEEE Trans. Speech and Audio Process., vol.2, no.2, April 1994.

[4] R. Gomez, T. Toda, H. Saruwatari, K. Shikano, “Improving the rapid unsupervised speaker adaptation through HMM-sufficient statistics weighting,” Acoustical Society of Japan, pp.155-156, Mar. 2006.

[5] R. Gomez, T. Toda, H. Saruwatari, K. Shikano, “Improving rapid unsupervised speaker adaptation based on HMM sufficient statistics,” ICASSP, pp.1001-1004, 2006.