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

2004－SLP－54（53）
2004／12／22

混合ガウス分布による多言語音声系統樹の構成

朱　世イ†　山本　幹雄‡　板橋　秀一‡

†筑波大学大学院理工学研究科 〒305-8573　つくば市天王台 1-1-1
‡筑波大学大学院システム情報工学研究科 〒305-8573　つくば市天王台 1-1-1
E-mail: †zhu@mibel.cs.tsukuba.ac.jp, ‡{myama, itahashi}@cs.tsukuba.ac.jp

あらまし　本論文では，多言語音声のクラスタリングに関する手法を提案する．ここでは，言語を音声特徴あらメータを生成する情報源とみなし，この情報源の確率・統計的な性質を確率モデルによりモデルかする．次に，確率モデル間に距離尺度を導入し，この尺度に基づく多言語音声の系統樹を構成する
キーワード　混合ガウスモデル，多言語音声系統樹，クラスタリング

Constructing a Family Tree of Multilingual Speech Based on Gaussian Mixture Models

Shiwei ZHU†　Mikio YAMAMOTO‡　and　Shuichi ITAHASHI‡

†Master's Program in Science and Engineering, University of Tsukuba　1-1-1 Tennodai, Tsukuba, 305-8573 Japan
‡Graduate School of Systems and Information Engineering, University of Tsukuba　1-1-1 Tennodai, Tsukuba, 305-8573 Japan
E-mail: †zhu@mibel.cs.tsukuba.ac.jp, ‡{myama, itahashi}@cs.tsukuba.ac.jp

Abstract　This paper proposes a method for automatically clustering multilingual speech. We consider that language is the source of information which generates the speech feature parameter, and the probability and the statistical characteristics of these information are modeled by GMMs. Next, a distance measure between GMMs is proposed. Based on this we construct a family tree of multilingual speech.
Keyword　Gaussian Mixture Model, Family Tree of Multilanguage Speeches, Clustering

1．まえがき

近年，大量の多言語音声コーパスが利用可能になったことや，計算機の性能が大幅に向上したことから，コーパス・データを利用した確率モデルの音声認識や多言語音声識別，そして方言音声識別への応用の研究が活発に行われている．確率モデルは，従来，音声，言語処理などの工学分野での有効性を実証してきたが，多言語，方言などの言語学の各分野においても有用な手法を提供するものと思われる．

本研究は，言語学の分野での確率モデルの有用性を示す一例として，多言語音声系統樹の構成を取り上げる．また，多言語音声の間の音響類似性を実験的に検証してみたい．ここでは，言語を音声特徴パラメータを生成する情報源とみなし，この情報源の確率・統計的な性質を確率モデルによりモデル化する．次に，確率モデル間に距離尺度を導入し，この尺度に基づく多言語音声のクラスタリングを行う方法を提案する．

2．関連研究

確率・統計的手法に基づき，言語の比較を計量的に行う研究は，従来から広く行われている．文献[1]にはtrigram モデルを用い，確率モデルを用いた言語クラスタリングの方法を述べ，その結果がかなりの程度で言語学での分類と一致することを示した．

ここで，trigram 或は N-gram モデルは文字列をモデル化するために適した確率モデルであるので，言語らしさを測ることができる．けれども，それは音響的な特徴パラメータと直接には対応していない．そのため，多言語音声識別において有効な確率モデルが必要である．

多言語音声識別の従来手法は音声認識するかしないかにより 2 種類に分けられる．今まで一番高い識別率が得られると言われた方法は「音声認識＋言語モデル」[4]である．即ち，未知音声データを音声認識

システムにより，文字列または音素列などに転換し，言語モデルを用いて識別するものである．これらのシステムの識別率は高いが，音声認識システムでは各言

語の各音素の確率音響モデル[1]を訓練データにより推定し，また，識別するとき，全ての言語と比較しなければならないので，計算量はかなり多く，リアルタイムの言語識別に適応できない．また，訓練データは，セグメント化や音素ラベリング[2]されなければならない．特に，音素認識の精度を保証するために，音素ラベリングは基本的に手作業で行うことが必要である．

それに対して，音声認識をせず，単純な音響情報だけを利用する識別方法も続けて研究されている．その中で代表的な方法は GMMs(Gaussian Mixture Models) classifier[4]，VQ クラスタリング[3][5]などがある．特に GMMs による多言語音声識別は効率的かつ識別率の高い手法として，いくつかの論文の中で紹介された．

3. 混合ガウスモデルによる多言語音声のクラスタリング

本研究で提案する方法は，まず各言語の音声データから特徴パラメータを抽出して，その特徴パラメータの確率モデルを自動的に学習し，次に確率モデル間に距離を導入することにより，言語間の距離を定義する．その距離を群平均法(UPGMA: Unweighted Pair-Group Method using Average)と呼ばれるクラスタリング・アルゴリズムによって，音声の樹状図を作成する．確率モデルとしては，ガウス混合モデル(GMMs)を用いた．

3.1. 音声特徴パラメータ:メルケプストラム (Mel-Cepstrum)

今回の実験では，メルケプストラムを利用して音声の特徴を表す．音声生成の基本的な構造は，声帯(音源)振動により生成された音が，声道(咽頭腔，口腔，鼻腔)の形により形成される音響的なフィルタ(調音フィルタ)を通過することでさまざまに変化し，口または鼻から「放射」されることで説明している[7]．ケプストラムは音源と調音フィルタの特性を分離する信号処理である．特に低次数部分のケプストラムは調音フィルタ，即ち声道の特徴を表現できる．音声認識においては低次数部分のケプストラムを用いた HMM(Hidden Markov Model)法などが広く用いられている．そのうえ，メルケプストラムは非線形周波数スケールが使用されるので，聴覚システムの動きにより似ている．今回の実験では，サンプリング周波数 16KHz の音声データに対して 20 次ケプストラムを一つのベクトルとして抽出した．各パラメータの値は下に示されている．

- 分声フレーム長: 25ms
- フレームシフト: 10ms
- 分析窓: HAMMING
- チャネル数: 36
- カットオフ周波数: [300Hz:3400Hz]

3.2. ベクトル量子化

ベクトル量子化は，特徴ベクトルにクラスタ化の手法を適用し，代表パターンの集合を生成する方法である．例えば：いくつか n 次元ベクトル $Y_j=(y_{0j},y_{1j},\cdots,y_{(n-1)j})^T$, j=0,…,J から成る n 次元空間において I 個の代表ベクトル $Y_0^*,Y_1^*,\ldots,Y_{(I-1)}^*$をとり，I 個のクラスタとする[3]．ここで代表ベクトルをクラスタの平均値として，各クラスタの分散や共分散が求められる．

その平均値と分散は GMMs モデルを推定するための EM アルゴリズムの初期値とすることができる．そうすることで，EM アルゴリズムの計算量を減らすことができる．

3.3. 混合ガウスモデル

ガウスモデルは正規分布とも呼ばれ，自然界や人間社会の中の数多くの現象に対してあてはまり，多くの統計研究に応用されている．ガウス分布の密度関数は次式で表される：

$$p(x)=N(x\mid\mu,\Sigma)=\frac{1}{(2\pi)^{d/2}\mid\Sigma\mid^{1/2}}\exp[-0.5(x-\mu)^T\Sigma^{-1}(x-\mu)]$$

μ：平均値

Σ：分散共分散行列

d: ベクトルの次元数

一方，ガウス混合モデルは複数の正規分布の密度関数を重みつきで足し合わせた密度関数で表現する：

$$p(x_i)=\sum_{j=1}^{M}p(x_i\mid j)w(j)=\sum_{j=1}^{M}N(x_i\mid\mu_j,\Sigma_j)w(j)$$

$$\sum_j w(j)=1$$

w(j): j 番目ガウス成分の重み(確率ともいえる)

離散パターンにおいての対数ゆう度は：

$$\log\{\prod_{i=1}^{N}p(x_i)\}=\sum_{i=1}^{N}\log p(x_i)=\sum_{i=1}^{N}\log\{\sum_{j=1}^{M}p(x\mid j)w(j)\}$$

のように計算する．

混合ガウス分布を推定するために EM アルゴリズムを用いた．

Step 1: k=0, M, μ_j^k, Σ_j^k, w^k (j)に初期値を与える．

Step 2: ゆう度が収束するまで，p(j|x)，μ_j，Σ_j，w(j)を繰り返して計算する：

[1]隠れモルコフモデル(HMM)

[2] ラベリング(labeling):各音素の時間位置と名前(ラベル)を示すこと．

[3] クラスタ化するには，k-means アルゴリズムや LBG アルゴリズム[6]などがある．

	ame	hun	chi	ita	eng	rus	fre	ger	gre	hin	ara	ind	dut	jap	kor	pol	fin	spa	swe	tha
ame	0.000																			
hun	0.062	0.000																		
chi	0.060	0.049	0.000																	
ita	0.083	0.064	0.080	0.000																
eng	0.031	0.038	0.058	0.061	0.000															
rus	0.073	0.058	0.077	0.071	0.050	0.000														
fre	0.071	0.040	0.070	0.092	0.048	0.074	0.000													
ger	0.032	0.046	0.062	0.083	0.032	0.082	0.058	0.000												
gre	0.093	0.047	0.088	0.084	0.077	0.088	0.085	0.087	0.000											
hin	0.081	0.070	0.076	0.107	0.068	0.088	0.073	0.070	0.107	0.000										
ara	0.136	0.059	0.127	0.106	0.094	0.089	0.087	0.116	0.053	0.164	0.000									
ind	0.069	0.049	0.068	0.100	0.049	0.074	0.067	0.069	0.086	0.095	0.116	0.000								
dut	0.052	0.074	0.063	0.135	0.059	0.108	0.082	0.051	0.103	0.075	0.183	0.074	0.000							
jap	0.077	0.052	0.060	0.088	0.052	0.081	0.084	0.082	0.090	0.086	0.136	0.042	0.051	0.000						
kor	0.060	0.044	0.057	0.086	0.089	0.077	0.063	0.083	0.110	0.077	0.112	0.051	0.046	0.032	0.000					
pol	0.059	0.053	0.066	0.051	0.035	0.043	0.077	0.071	0.079	0.076	0.089	0.058	0.089	0.051	0.055	0.000				
fin	0.070	0.045	0.072	0.053	0.055	0.054	0.054	0.068	0.090	0.071	0.080	0.073	0.108	0.087	0.070	0.050	0.000			
spa	0.051	0.050	0.065	0.063	0.036	0.066	0.032	0.064	0.086	0.083	0.122	0.047	0.064	0.053	0.043	0.038	0.071	0.000		
swe	0.075	0.071	0.128	0.075	0.073	0.076	0.076	0.123	0.090	0.090	0.077	0.117	0.065	0.128	0.103	0.077	0.062	0.074	0.000	
tha	0.080	0.051	0.053	0.101	0.051	0.077	0.069	0.076	0.088	0.090	0.122	0.035	0.090	0.071	0.066	0.073	0.079	0.070	0.113	0.000

図 1. 20 言語の尤度距離

for each i, j :

$$p^k(j \mid x_i) = \frac{p^k(x_i \mid j) w^k(j)}{\sum_i p^k(x_i \mid p^k(j))}$$

for each j :

$$\mu_j^{k+1} = \frac{\sum_i p^k(j \mid x_i) x_i}{\sum_i p^k(j \mid x_i)}$$

$$\Sigma_j^{k+1} = \frac{\sum_i p^k(j \mid x_i) \| x_i - \mu_j^{k+1} \|}{d \sum_i p^k(j \mid x_i)}$$

$$p^{k+1}(j) = \frac{\sum_i p^k(j \mid x_i)}{N}$$

$$k = k + 1$$

ここで，ガウス成分の数はベクトル量子化されたクラスタの数と同じにし，各クラスタの平均と分散共分散行列を各ガウス成分の初期値とした．また，簡単のために，今回の実験では各クラスタの分散共分散行列は対角行列にした．

3.4. 混合ガウスモデル間の距離

混合ガウスモデル間の距離は，隠れマルコフモデル(Hidden Markov Model, HMM)間の距離として定義された距離を用いた[2]．

いま，言語 L_1 および言語 L_2 の音声の特徴パラメータの集合として，それぞれ D_1, D_2 が与えられているとする．$|D_i|$は D_i の数と表記する．またパラメータ集合 D_i から作成された言語モデルを M_i で表す．

まず，言語モデル M_1 および M_2 に対し，距離尺度 $d_0(M_1,M_2)$を次のように定義する：

$$d_0(M_1, M_2) = \frac{1}{|D_2|}[\log p(D_2 \mid M_2) - \log p(D_2 \mid M_1)]$$

上の式では，L_1 と L_2 の間の距離を，L_2 のモデル M_2 からデータ D_2 が生成される確率と，L_1 のモデル M_1 からデータ D_2 が生成される確率の差に基づいて決める．L_1 と L_2 が類似しているなら，距離は小さくなるし，類似していなければ，距離も大きくなる．

上の距離は，モデル M_1 と M_2 に対し，非対称($d_0(M_1,M_2) \neq d_0(M_2,M_1)$)であるので，$d_0(M_1,M_2)$と $d_0(M_2,M_1)$の平均を取り，対称形にする．したがって，モデル M_1 と M_2 の間の距離 $D(M_1,M_2)$は次のように定義される．

$$D(M_1, M_2) = \frac{1}{2}[d_0(M_1, M_2) + d_0(M_2, M_1)]$$

3.5. 群平均法(UPGMA)

上記により計算されたモデル間の距離に対し，階層的クラスタ分析をするために，群平均法 UPGMA(Unweighted Pair-Group Method using Average)と呼ばれる方法を用いた．群平均法は，広い範囲において良い結果を与えるクラスタ分析法であると良く言われている．それは，N 個のモデルの距離行列(N×N)の中で最小な距離のモデル M_i と M_j をまとめて新しいモデル $M_{(N+1)}$とし，M_i と M_j の代わりに距離行列に入る．この新しいモデル $M_{(N+1)}$と残りのモデルの間の距離を計算し，距離行列を更新し，(N-1)×(N-1)の行列になる．この操作を繰り返すうちに，モデルが次々にまとめられてゆき，最後にひとつになる．このようにし

て，系統樹が作成される．

4. 実験について

4.1. コーパス

今回の実験で用いた音声コーパスはNTT多言語音声コーパス「NTT Speech Data Base For Telephonometry 1994」である．その中で英語(米)，英語(英)，アラビア語，中国語(標準語)，オランダ語，フィンランド語，フランス語，ドイツ語，ギリシャ語，ハンガリー語，ヒンディー語，インドネシア語，イタリア語，日本語，韓国語，ポーランド語，ポルトガル語，ロシア語，スペイン語，スウェーデン語，タイ語などの20言語が収集されている．各言語ずつ，男女各4名のネイティブ話者，一人24文の音声データが収録された．音声データはサンプリング周波数16kHzでデジタル化されている．なお，今回の実験では男性の音声データだけを用いた．

4.1. 分類の結果と考察

今回の実験では，各言語から，4万個ほどの特徴パラメータを抽出し，16混合の混合ガウスモデルを作成した．各言語のGMMsモデル間の距離行列は図1のようになっている．

実験により得られた結果は，図2に示すようになっている．

実験の結果を評価するため，言語学上の分類を基準として実験の結果を考察した．

まず，評価実験で用いた言語は，言語学上で以下のように大きく分類される：

(A) Indo-European Family :
- (A-1) Indo-Iranian→Hind
- (A-2) Hellenic→Greek
- (A-3) Italic→Latin
  - ① Proto-Italo-Romance→Italian
  - ② Proto-western-Romance→
    - French
    - Spanish
    - Portuguese
- (A-4) Balto-Slavic
  - ① West Slavic→Polish
  - ② East Slavic→Russian
- (A-5) Germanic
  - ① North Germanic→Swedish
  - ② West Germanic→
    - English
    - Dutch
    - German

(B) Uralic language Family :
- (B-1) Finnic→Finnish
- (B-2) Ugric→Hungarian

(C) Afro-asiatic Family : Arabic

(D) Altaic Language Family :
- (D-1) Japanese
- (D-2) Korean

(E) Sino-Tibetan Family : Chinese

(F) Kadaic Family : Thai

(G) Malayo-Polynesian Family : Indonesian

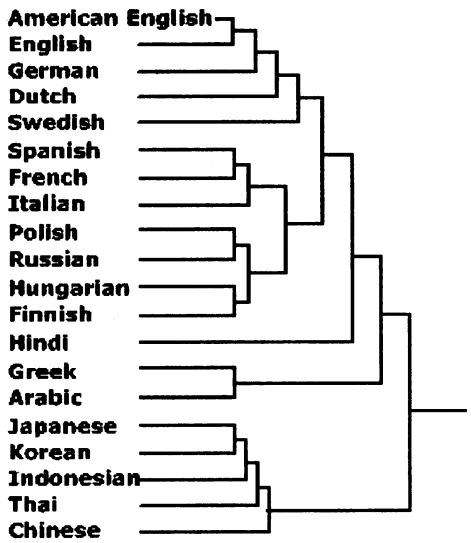

図 2. UPGMAによる多言語音声系統樹

今回の分類は，音響情報だけを用いたので，言語学上の分類と完全に一致することはないが，図に示すように，実験により得られた結果は，大まかに上記の大分類を反映したものになっている．

次に，言語間のより細かな関係について調べた．最初に一つのクラスタとしてまとめられたのはアメリカ英語とイギリス英語である．同じ英語であるので，その結合は極めて妥当であるといえる．また，スペイン語，フランス語とイタリヤ語の結合，フィンランド語とハンガリー語の結合，ポーランド語とロシア語の結合も妥当といえる．アラビア語とギリシャ語の結合はちょっと不自然であるが，その以外の分類はかなりの部分で言語学での分類と一致になっているので，提案した分類手法が有効なものであることを示している．

5. まとめ

今回，ガウス混合モデルによる多言語音声系統樹の構成手法を提案した．また，それに基づいた実験を行い，実験結果を言語学での分類と比較することにより，

提案した手法の有効性を示した．

今後，女性話者の音声データに対する評価実験を予定している．また：

- 分類の精度を更に向上するために，音素配列情報も含めて，より正確な言語モデル作る．
- ケプストラム以外の特徴パラメータを利用する．
- 音声データを増してクラスタリング実験を行う．

などを考えている．

文　献

[1] 北研二,"確率的言語モデルに基づく多言語コーパスからの言語系統樹の再構築," 自然言語処理, July, 1997. (雑誌例 1) 山上一郎，山下二郎，“パラメトリック増幅器，”信学論(B), vol.J62-B, no.1, pp.20-27, Jan.1979.

[2] B. H. Juang and L. R. Rabiner "A Probabilistic Distance Measure for Hidden Markov Models," AT&T Technical Journal, Vol.64, No.2, February 1985.

[3] Masahide Sugiyama "Automatic Language Recognition Using Acoustic Features," Technical report TR-I-0167, ATR Interpreting Telephony Research Laboratories, 1991.

[4] M. Zissman, E. Singer "Automatic Language Identification of Telephone Speech Messages using Phoneme Recognition and N-Gram Modelling," Proc. ICASSP-94, Minneapolis, USA, 1994.

[5] T. Nagarajan, Hema A. Murthy "A pairwise multiple codebook approach to implicit language identification," In WSLP-2003, Mumbai, Indo, 2003..

[6] 中川聖一著\quad"確率モデルによる音声認識," 電子情報通信学会, 1988.

[7] 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄 "音声認識システム," オーム社, 2001