論文 / 著書情報
Article / Book Information

論題(和文)	マハラノビス距離を用いた日本語話し言葉音声の音響的特徴の分析
Title(English)	
著者(和文)	中村匡伸, 岩野公司, 古井貞熙
Authors(English)	Masanobu Nakamura, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2005年春季講演論文集, Vol. , No. 2-1-4, pp. 231-232
Citation(English)	, Vol. , No. 2-1-4, pp. 231-232
発行日 / Pub. date	2005, 3

2—1—14 マハラノビス距離を用いた日本語話し言葉音声の音響的特徴の分析 *

◎中村匡伸　岩野公司　古井貞熙 (東工大)

1 はじめに

話し言葉音声の音響的特徴の分析は，話し言葉音声の認識性能の向上や，音声合成の品質向上に役立つと考えられ，非常に重要である．我々はすでに，日本語話し言葉コーパス (以下 CSJ と呼ぶ) に収録されている同一話者による話し言葉音声 (学会講演音声，模擬講演音声，対話音声) と読み上げ音声 (学会講演音声の再読み上げ) において各音素のケプストラムに関する比較を行った [1]．その結果，話し言葉音声では読み上げ音声に比べてケプストラム空間が縮小する傾向があることが明らかになった．

しかし，読み上げ音声に対して話し言葉音声の認識性能が劣化する理由として，ケプストラム空間が縮小することの他に，各音素のケプストラムの分散が大きくなることも考えられる．そこで，本稿では2音素間のケプストラムの分布の比較をマハラノビス距離を用いて行う．なお，音声データは多数話者による 4 つの発話タイプの音声 (朗読音声，学会講演音声，模擬講演音声，対話音声) を用いる．さらに，発話タイプの異なる音声データを用いた認識を行い，マハラノビス距離と音素正解精度との関係を明らかにする．

2 音声データ

分析には CSJ に含まれる，発話タイプの異なる朗読音声，学会講演音声，模擬講演音声，対話音声を用いる．音声データは 16kHz でサンプリングされており，1 講演は約 15 分のデータである．実験に際して，まず転記ファイルをもとに音声データを 400ms 以上の無音区間で区切り，区切られた区間を「発話単位」として定義した．発話単位が 1 秒未満の場合には，後続する発話単位と接続し，1 つの発話単位とみなした．

3 音響特徴量の抽出

それぞれの発話タイプごとに，各音素のケプストラムの平均と分散を求める．本実験において分析対象とする音素は，表 1 のリストにある 31 種（母音 10 種・子音 21 種）とした．分析対象データの話者は，男女各 5 名とし，無作為に選択した．朗読音声には「講演の再読み上げ」「対話形式エッセーの朗読」などの，対話音声には「インタビュー」「自由対話」などの異なる複数種類の音声が存在している．そこで，この 2 つの発話タイプの音声については，これらの種類をなるべく網羅するように，分析対象データを選択した．なお，学会講演音声と模擬講演音声については，CSJ のテストセット中から選択を行った．これらの話者の音声データを用い，発話タイプごとに性別非依存で分析を行う．図 2 に，分析対象

表 1. 音素のリスト

母音	/a,i,u,e,o,a:,i:,u:,e:,o:/
子音	/w,y,r,p,t,k,b,d,g,j,ts,ch, z,s,sh,h,f,N,N:,m,n/

表 2. 音声データの発話時間および音素サンプル数

発話タイプ	発話時間 (分)	音素サンプル数
朗読音声 (R)	82	44,311
学会講演音声 (A)	132	78,594
模擬講演音声 (S)	102	53,813
対話音声 (D)	126	55,160

データにおける発話タイプごとの発話時間と音素サンプル数を示す．ここで各音素のケプストラムの平均・分散は，以下のようにして抽出される．

(1) 音声データから MFCC 12 次元とその一次微分，二次微分成分，対数パワーの一次微分，二次微分成分の計 38 次元の音響パラメータを抽出する．分析周期は 10ms, 分析窓幅は 25ms とし，発話単位ごとに CMS 処理を行っている．

(2) 各発話タイプごとに，分析対象データを用いて 1 混合 monophone HMM を学習する．全ての音素モデルは，3 状態の left-to-right 型 HMM とする．

(3) 出来上がった monophone HMM のうち，分析対象音素の HMM の第 2 状態から 12 次元 MFCC の平均ベクトルと分散ベクトルを取り出す．

4 マハラノビス距離

音素 i と j のマハラノビス距離 D_{ij} を以下のように定義する．

$$D_{ij} = \sqrt{\frac{K\sum_{k=1}^{K}(\mu_{ik}-\mu_{jk})^2}{\sum_{k=1}^{K}\sigma_{ik}^2+\sum_{k=1}^{K}\sigma_{jk}^2}} \quad (1)$$

K は MFCC ベクトルの次元数 ($K = 12$) である．μ_{ik} および σ_{ik}^2 は，それぞれ音素 i の平均および分散 MFCC ベクトルの k 次元目の要素である．

図 1 は，各講演タイプにおける全音素間のマハラノビス距離の相対累積度数である．x 軸をマハラノビス距離，y 軸を相対累積度数とし，朗読音声，学会講演音声，模擬講演音声，対話音声をそれぞれ R, A, S, D とする．図 1 より，自発度の高い順 (D ≫ S > A ≫ R) に 2 音素間のマハラノビス距離が小さ

* Analysis of cepstral features of Japanese spontaneous speech using Mahalanobis distance
By Masanobu Nakamura, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology)

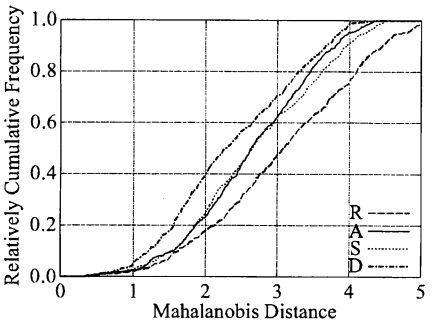

図 1. マハラノビス距離の相対累積度数

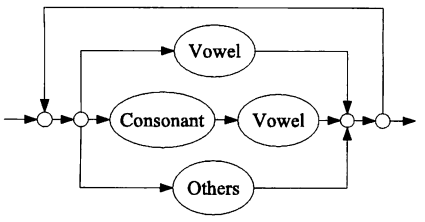

図 2. 音素ネットワーク

くなる傾向がある．朗読音声，学会講演音声，模擬講演音声，対話音声の全音素間のマハラノビス距離の分布の平均値は，それぞれ 3.07，2.65，2.70，2.38 となった．発話タイプ間の母平均の差の検定を行ったところ，学会講演音声と模擬講演音声を除き，有意水準 1 % で，2 音素間のマハラノビス距離の分布に差があることが示された．

5 マハラノビス距離と音素正解精度

発話タイプの違いによる，2 音素間のマハラノビス距離の分布の違いは，音素正解精度に現れると考えられる．そこで実際に音素正解精度を求め，マハラノビス距離と音素正解精度の関係を示す．

認識精度を算出するために用いる学習データは，無作為抽出した CSJ の学会講演音声と模擬講演音声の男女各 100 名の音声データ (学会講演音声約 50 時間，模擬講演音声約 40 時間) を用いる．音響特徴量は計 38 次元を用い，1 混合 monophone HMM を学習する．評価対象の音声データは，3 節で用いたデータと同じものを用いる．これらは学習データには含まれていない．本実験ではネットワークを用いて音素認識を行い，音素正解精度を算出する．図 2 に，用いた音素ネットワークを示す．**Vowel** は母音と長母音，**Others** は /N, N:, q, sp/, **Consonant** は **Others** を除いた子音を表す．音素認識を行う際には，挿入ペナルティは発話タイプごとに最適な値を用いる．

図 3 に，マハラノビス距離と音素正解精度の関係を示す．発話タイプごとのマハラノビス距離の値は，全音素間のマハラノビス距離の平均値である．またマハラノビス距離と音素正解精度の相関係数は 0.92

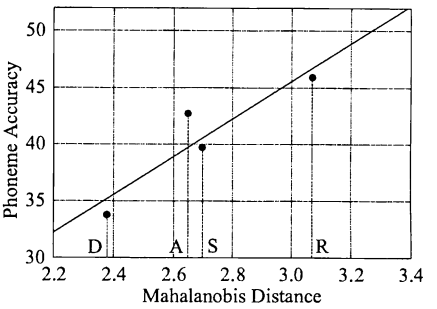

図 3. マハラノビス距離と音素正解精度の関係

となった．図中の直線は最小自乗法を用いて，点列を一次関数で近似したものである．これらの結果より，全音素のマハラノビス距離の平均と音素正解精度には強い相関関係があることが分かった．すなわちこの結果は，2 音素間のマハラノビス距離の分布を調べることで認識精度の予測が可能であることを意味している．

6 まとめ

本稿では CSJ の朗読音声，学会講演音声，模擬講演音声，対話音声という発話タイプの異なる話者非依存の音声データを用いて，その全音素間のマハラノビス距離の平均を計測した．その結果，発話の自発性が高くなるとマハラノビス距離が小さくなるという傾向が明らかになった．また，全音素間のマハラノビス距離の平均値と音素正解精度には 0.92 の相関が見られた．

本稿で取り上げた発話タイプによるマハラノビス距離の分布の違いは，音声の「自発性」を示す指標であると考えられる．この仮定に一般性があるかどうか調べるためには，CSJ 以外の他のコーパスに対しても同様の分析を行う必要がある．さらに，音声データの「自発性」を判断するのにどの程度のデータが必要であるかを調べるために，今回と同様の分析結果が得られる最低限のデータ量を調査する必要がある．また，今回得られた知見を，話し言葉音声の認識性能の向上に役立てることができるかどうか，検討する必要がある．

謝辞

本研究は文部科学省 21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」の一環として実施されました．

参考文献

[1] 中村匡伸, 岩野公司, 古井貞熙 “日本語話し言葉コーパスを用いた話し言葉音声の音響的特徴の分析,” 情報処理学会研究報告, 2004-SLP-53 (2004-10).