論文 / 著書情報
Article / Book Information

論題(和文)	音声認識のためのq ガウス分布を用いた音響モデル
Title(English)	
著者(和文)	周澤西, 岩野公司, 篠田浩一
Authors(English)	Takusei Syuu, Koji Iwano, Koichi Shinoda
出典(和文)	日本音響学会講演論文集, , , pp. 175-178
Citation(English)	, , , pp. 175-178
発行日 / Pub. date	2013, 3

音声認識のための q ガウス分布を用いた音響モデル*

◎周澤西 (東工大), 岩野公司 (都市大), 篠田浩一 (東工大)

1 まえがき

音声認識の性能は周囲雑音の影響によって著しく劣化する．劣化の原因として，音響モデルを作成する際に用いられる音声と認識対象の音声データとの間での雑音環境が異なることが考えられる．この問題を解決するために音響モデルを雑音環境に適応させる方法がある．従来手法として MLLR[1] や MAP[2] などが広く用いられている．ところが，これらの手法で作成された音響モデルは雑音下の音声特徴量の分布を十分に表現できない．その理由として，雑音の重畳した音声特徴量の分布が長いすそ（ロングテール）をもつことがあげられる．

物理学や経済学において「複雑系」における現象がしばしばロングテールをもつ分布に従うことが報告されている [3].「複雑系」とは多種多様な因子が相互作用しているために未来予測が困難であるシステムのことである．雑音環境下での音声もこれに属すると考えられる．近年この分布を表現するために q ガウス分布が提唱された [4]．q ガウス分布とは通常のガウス分布がもつ 2 つのパラメータに加え,「分布の裾の広さ」を表すパラメータ q をもつ．

本論文では，q ガウス分布を用いて雑音が加わった音声特徴量の分布の広がりを表現した新しいモデルによる耐雑音手法を提案する．

2 q ガウス分布

2.1 非示量性統計

「系内の各因子は一様で，かつ，独立である」ことを前提にして，中心極限定理は導出された．この枠組みにおいて，ある 1 つのマクロな系の状態量と，それの部分系の状態量の和は等しい．これを「相加性」といい，また相加性が成り立つ状態量を「示量性変数」という [4]．この定理に従う分布として，ガウス分布は幅広い分野で応用されている．

ところが，この前提から導出された理論を用いても解明できない現象が多く存在する．このような現象では，異なる因子間の相関が非常に大きい．そのためシステムをモデル化するにあたり，因子が互いに及ぼす影響を加味した考察が必要になる．非示量性統計は，こうしたミクロな系の一様性，独立性を前提としない理論から派生した．

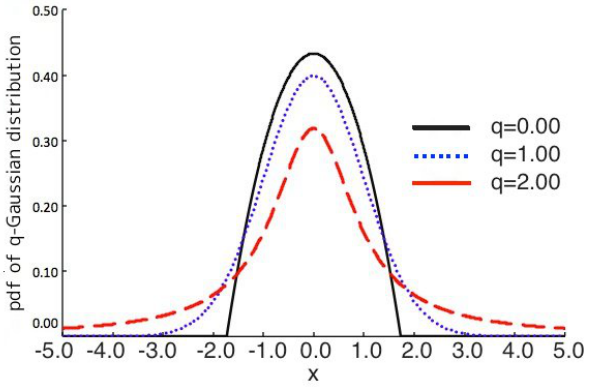

Fig. 1 The q-Gaussian distribution[5].

2.2 q ガウス分布の導出

q ガウス分布とは，ガウス分布を一般化した確率分布であり，非示量性統計の 1 つである Tsallis 統計 [4] から導出される確率分布である．ガウス分布は平均や分散に関する制約条件のもとで Boltzmann-Shannon エントロピー

$$S = -\int p(x)\log p(x)dx \tag{1}$$

を最大化する確率分布として導出されるのに対し，q-ガウス分布は Tsallis エントロピー

$$S_q = -\frac{1}{1-q}\left(1-\int p(x)^q dx\right) \tag{2}$$

を最大化する確率分布として導出される．ここで，q は実数値 ($q > 0$) をとり，$q = 1$ とした時に Tsallis エントロピーは Boltzmann-Shannon エントロピーに一致する．Tsallis エントロピーはある 2 つの系 A, B と，それらの合成系 $A \cup B$ に関して，

$$S_q^{A\cup B} = S_q^A + S_q^B + (1-q)S_q^A S_q^B \tag{3}$$

を満すため，示量性を満たさない．両エントロピーの違いは式 (3) の最後の項にあり，これがシステム内の因子の相関を表す．

q 値を変化させた際の q ガウス分布を Fig. 1 に示す．q ガウス分布は $q > 1$ の時，ガウス分布よりも裾が長い分布となり，$q = 1$ とするとガウス分布に一致する．

*q-GMM based acoustic model for speech recognition. by Zhou Zexi (Tokyo Institute of Technology), Koji Iwano (Tokyo City University) and Koichi Shinoda (Tokyo Institute of Technology)

2.3 多次元 q ガウス分布

次元数が D である q ガウス分布は以下のように定義される [6].

$$\mathcal{N}_q(X|\mu_q,\Sigma_q) = \begin{cases} \frac{1}{Z_{q,D}}\left(1-\frac{(1-q)}{(D+2-Dq)}(X-\mu_q)^T\Sigma_q^{-1}(X-\mu_q)\right)^{\frac{1}{1-q}}, \\ \quad \text{if } (X-\mu_q)^T\Sigma_q^{-1}(X-\mu_q) < \frac{(D+2-Dq)}{(1-q)} \\ 0, \quad \text{otherwise} \end{cases} \tag{4}$$

ここで μ_q , Σ_q はそれぞれ q ガウス分布における平均ベクトル，および分散行列を表し，$Z_{q,D}$ は式 (4) の X に関する積分を 1 とするための正規化定数である．上記の各パラメータは μ , Σ をそれぞれガウス分布の平均ベクトル，および分散行列として，以下の式で表現される．

$$\mu_q = \mu \tag{5}$$

$$\Sigma_q = \frac{(D+4)-(D+2)q}{(D+2)-Dq}\Sigma \tag{6}$$

$$Z_{q,D} = \begin{cases} \left(\frac{D+2-Dq}{q-1}\right)^{\frac{D}{2}} \frac{\pi^{\frac{D}{2}}\Gamma(\frac{1}{q-1}-\frac{D}{2})}{\Gamma(\frac{1}{q-1})}|\Sigma_q|^{\frac{1}{2}}, \\ \quad \text{for } 1<q<(1+\frac{2}{D}) \\ \left(\frac{D+2-Dq}{1-q}\right)^{\frac{D}{2}} \frac{\pi^{\frac{D}{2}}\Gamma(\frac{2-q}{q-1})}{\Gamma(\frac{2-q}{1-q}+\frac{D}{2})}|\Sigma_q|^{\frac{1}{2}}, \\ \quad \text{for } q<1 \end{cases} \tag{7}$$

3 耐雑音性モデル作成手法

3.1 q-混合ガウス分布

音声特徴量の分布の裾を広げる方法として，分散の値に底上げをする方法がある [7]．この手法はヒューリスティクスであり，その根拠は必ずしも明らかでない，そこで，q ガウス分布により観測される特徴量に近い分布を再現することを目指す.

本研究では HMM の出力確率分布として q-混合ガウス分布 (q-GMM) を用いる．q-GMM は q ガウス分布の混合分布であり，確率密度関数は以下で与えられる．

$$p_q(X|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}_q(X|\mu_{q,k},\Sigma_{q,k}) \tag{8}$$

ここで K は混合数，w_k は混合重み，$\theta = \{w_k,\mu_{q,k},\Sigma_{q,k}\}_{k=1}^{K}$ は q-GMM のパラメータ集合である．

3.2 学習アルゴリズム

GMM や HMM など，隠れ変数を持った確率モデルのパラメータ推定には Expectation Maximization (EM) アルゴリズムが広く用いられている．しかし，q-GMM に EM アルゴリズムを適用すると M-ステップが解析的に解けない．そこで本研究では，通常の GMM の M-ステップにおける更新式をそのまま q-GMM に用いる近似を導入する．以下アルゴリズムに沿って説明する．

1. クリーンな学習データから音声特徴量を抽出し，その集合を $X=\{x_i\}_{i=1}^{N}$ とする．
2. パラメータ $\hat{\theta}=\{\hat{w}_k,\hat{\mu}_k,\hat{\Sigma}_k\}$ を $\{1$, $\mathbf{0}$,$\mathbf{1}\}$ で初期化する．
3. (E-ステップ) 以下で与えられる負担率 c_{ik} を計算する．

$$c_{ik} = \frac{\hat{w}_k\mathcal{N}_q(x_i|\hat{\mu}_k,\hat{\Sigma}_k)}{\sum_{k=1}^{K}\hat{w}_k\mathcal{N}_q(x_i|\hat{\mu}_k,\hat{\Sigma}_k)} \tag{9}$$

4. (M-ステップ) まず q-GMM に対して EM アルゴリズムを適用することを考える．このとき，重みの総和を 1 とする制約条件を満たすためにラグランジュ乗数 λ を導入した評価関数

$$Q(\theta) = \log\prod_i N_q(x_i|\theta) + \lambda\left(1-\sum_{k'}w_{k'}\right) \tag{10}$$

を最大化するように各パラメータを更新したい．すなわち，平均 μ_k，分散 Σ_k，重み w_k はそれぞれ式 (10) を偏微分することで得られる導関数

$$\frac{\partial}{\partial\mu_k}Q(\theta) = \Sigma_k^{-1}\sum_{i=1}^{N} d_{ik}c_{ik}(x_i-\mu_k) \tag{11}$$

$$\frac{\partial}{\partial\Sigma_k}Q(\theta) = \frac{1}{2}\Sigma_k^{-2}\sum_{i=1}^{N} c_{ik}\left(d_{ik}(x_i-\mu_k)(x_i-\mu_k)^T-\Sigma_k\right) \tag{12}$$

$$\frac{\partial}{\partial w_k}Q(\theta) = \frac{1}{w_k}\sum_{i=1}^{N} c_{ik} - \lambda \tag{13}$$

を 0 とする値を求めたい．

ここで，

$$d_{ik} = \frac{2}{(D+2-Dq)-(1-q)(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)} \tag{14}$$

である．

GMM の場合との違いは式 (11) と式 (12) に変数 d_{ik} が新しく含まれていることである．式 (14) が示す

ようにこの d_{ik} は分母中に解くべき変数を含むため，最急降下法などの数値計算手法を用いると計算時間が多くかかる．そこで，$q \simeq 1$ のときに $d_{ik} = 1$ と近似できることを用い，以下のように各パラメータを更新する．

$$\hat{\mu}_k = \frac{1}{C_k} \sum_{i=1}^{N} c_{ik} x_i, \tag{15}$$

$$\hat{\Sigma}_k = \frac{1}{C_k} \sum_{i=1}^{N} c_{ik} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T, \tag{16}$$

$$\hat{w}_k = \frac{C_k}{\sum_{k=1}^{K} C_k}, \tag{17}$$

ここで，$C_k = \sum_i c_{ik}$ である．

5. 式 (5) および (6) から，$\mu_{q,k}$，$\Sigma_{q,k}$ を更新する．

$$\mu_{q,k} = \hat{\mu}_k \tag{18}$$

$$\Sigma_{q,k} = \frac{(D+4) - (D+2)q}{(D+2) - Dq} \hat{\Sigma}_k \tag{19}$$

6. 対数尤度値が収束するまで，3~5 のステップを繰り返す．

ここでの D は特徴量の次元数を表す．また，q-ガウス分布が 2 次のモーメントをもつのは $q < 1 + \frac{2}{D}$ の範囲に限られるため，ここでも $q < 1 + \frac{2}{D}$ の場合のみ考える．さらに，q-GMM に対し MLLR 適応を行うことも検討する．

4 評価実験

4.1 データセット

評価データとして，AURORA-2 [8] を用いた．タスクは雑音重畳音声を対象とした最大で 7 桁の連続 11 数字 (“zero” , … , “ten”) の認識である．学習は男女各々55 名の静かな環境下の 8440 発話で行う．2 つのテストセット A,B の音声は自動車内や展示会などの定常雑音 4 種類と，道路や空港などの非定常雑音 4 種類が付加された男女各々52 人の 1101 発話である．雑音の大きさは SNR で 0 dB から 20 dB まで 5dB おきにとったものと clean を合わせて計 7 種類を用いて実験を行う．

4.2 実験条件

HMM の出力確率分布として分散の底上げを行った GMM，q-GMM，前述の GMM に対し MLLR 適応を行ったもの (GMM + MLLR)，q-GMM に対し MLLR 適応を行ったもの (q-GMM + MLLR) の 4 種類を比較する．評価は．雑音の大きさごとに最もよい q 値を選んだときの認識率について平均をとることで行う．

Table 1　評価セット A , B における単語認識率 (%)

HMM / SNR	SNR20	SNR15	SNR10	SNR5	SNR0
GMM	93.7	83.8	62.7	35.7	15.8
q-GMM	94.5	84.8	64.4	37.0	16.5
GMM + MLLR	97.6	95.5	90.2	76.6	44.9
q-GMM + MLLR	98.0	96.1	91.5	78.8	49.2

特徴量として， 0 次を除いた 12 次元 MFCC とパワー，およびその一次と二次の差分の 39 次元 MFCC を用いた．HMM は各数字あたり 16 状態，各状態に対角共分散 3 混合の GMM または q-GMM をもつ．

なお，q の値は q ガウス分布が 2 次のモーメントをもつ範囲のうち，1.000 から 1.048 まで 0.001 の間隔で逐次的に動かして音響モデルを作成した．

4.3 実験結果

評価セット A , B を合わせた結果を Table 1 に示す．雑音の大きさに関わらず，HMM の出力確率分布として q-GMM を用いた方がより GMM を用いる場合よりも高い認識率が得られた．また MLLR と合わせることでさらに性能が改善した．

雑音の種類，大きさの違いと性能の関係を Table 2 に示す．表の数字は q-GMM 認識率から GMM の認識率を引いた値である．その左の数字はそのときの q-GMM の認識率が最も高いときの q の値である．

Subway や Station では雑音の大きさにらず q の最適値はほぼ一定であるが，Babble や Airport では q の最適値は雑音の大きさにより変化する．q を推定するためには，雑音の性質を考えなければならない．

5 まとめ

非示量性統計から導出される q ガウス分布を HMM の出力確率分布として用いる手法を提案した．その結果，より精密に分布を再現することが可能になり，高い認識率を得ることができた．また，MLLR による適応の効果も確認した．

今後，雑音に応じたパラメータ q の自動最適化法を検討する．

参考文献

[1] M. J. F. Gales and P.C.Woodland, “Mean and Variance Adaptation within the MLLR Framework,” In *Computer Speech and Language*, vol. 10, iss. 4, pp. 249–264, 1996.

Table 2　様々な雑音環境下における qGMM と GMM の比較．左欄：q-GMM-HMM の認識率 - GMM-HMM の認識率，右欄：q-GMM 音響モデルの最適な q 値

	Clean		SNR20		SNR15		SNR10		SNR5		SNR0	
Subway	0.1	1.021	0.3	1.024	0.6	1.026	2.4	1.031	1.8	1.028	1.9	1.033
Babble	0.0	1.008	2.0	1.024	2.4	1.024	1.4	1.024	0.4	1.003	0.0	1.000
Car	0.2	1.018	0.2	1.002	0.1	1.002	0.1	1.001	0.0	1.000	0.0	1.000
Exhibition	0.0	1.023	0.7	1.014	1.4	1.030	4.3	1.030	5.3	1.024	1.6	1.024
Restaurant	0.1	1.021	1.7	1.024	2.7	1.024	2.5	1.024	0.6	1.002	0.0	1.000
Street	0.0	1.008	0.6	1.011	0.8	1.028	2.8	1.024	2.7	1.024	1.2	1.024
Airport	0.2	1.018	0.7	1.024	1.0	1.002	0.9	1.023	1.2	1.002	0.6	1.001
Station	0.0	1.023	0.3	1.001	0.3	1.002	0.8	1.002	0.0	1.000	0.4	1.002

[2] J. L. Gauvain *et al.*, “Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” In *IEEE Transactions on Speech and Audio Processing*, vol. 2, No. 2, April 1994.

[3] S. Picoli Jr *et al.*, “q-Distributions in Complex Systems: a Brief Review,” In *Braz. J. Phys*, vol.39, no.2A, 2009.

[4] C. Tsallis, “Possible Generalization of Boltzmann-Gibbs Statistics,” In *Journal of Statistical Physics*, vol. 52, pp. 479–487, 1988.

[5] N. Inoue and K. Shinoda, “q-Gaussian Mixture Models for Video Semantic Indexing,” 信学技報, vol.112, no.197, pp.31–36, 2012.

[6] D.Ghoshdastidar *et al.*, “q-Gaussian based Smoothed Functional algorithms for stochastic optimization,” In *Information Theory Proceedings*, July 2012.

[7] S. Young *et al.*, “The HTK Book,” Cambridge University Engineering Department, edition 3.4, 2006.

[8] D. Pearce and H. Hirsch, “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions,” In ISCA ITRW ASR2000, Paris, pp.181–188, 2000.