TR-I-0306

マイク入力音声で学習した混合連続分布 HMM の電話音声認識への適用
Telephone speech recognition using continuous HMMs with clean speech data

片岸一起、藤原紳吾、杉山雅英
Kazuki KATAGISHI , ShiNgo FUJIWARA , Masahide SUGIYAMA

1993.3.12

概要

マイク入力によって収録された標準音声を用いて学習した混合連続分布 HMM 音声認識システムを用いて、電話音声のように周波数特性がマイクのそれとは異なる系からの入力音声の認識手法を提案し、24 音素認識実験および文節認識実験により評価した。男性話者一名 (MAU) のマイク入力音声に対して 24 音素認識率 99.0%(99.9%) の音素 HMM を用いた電話音声中の音素認識率は 27.5%(80.2%) に低下した。標準 HMM のガウス分布の平均および分散を補正することにより、74.2%(96.8%) の音素認識率に向上した。また、同一話者のマイク入力音声に対して文節認識率 87.5%(98.6%) の音素 HMM を用いた電話音声中の文節認識率は 2.9%(8.2%) に低下した。先の場合と同じ補正値を用いることにより、47.0%(68.5%) の文節認識率に向上した。以上のことから、本手法の有効性が確認された。

ATR 自動翻訳電話研究所
ATR Interpreting Telephony Research Laboratories

目次

図目次

表目次

1 まえがき

近年、大規模な音声データベースが作成されつつあるが、実際の音声認識の応用においては入力音声と異なる周波数特性や雑音環境で動作させる必要があり、そのすべての環境に対する音声データベースを作成することは一般的には不可能に近い。そこで、標準音声で作成された音声認識モデルを適応する必要がある。文献 [1] では雑音に関する各種の方法を提案しその有効性を示した。また、電話音声の認識についてはその必要性から多くの研究がなされている [4, 2, 3] が、電話音声への適応問題については多くない [5]。本稿では、標準音声を用いて学習した混合連続分布 HMM 音声認識システムを用いて、電話音声のように周波数特性の異なる系からの入力音声の認識手法を提案し、音素認識実験および文節認識実験により評価する。本手法は電話音声に限らず周波数伝達特性が異なる音声の補正に対して一般的に適用可能である。

2 電話入力音声の HMM

電話音声は通過帯域が 0.3 - 3.3 kHz に制限されており、2種類の歪みが発生する。(1) 通過帯域の周波数特性に対応する線形歪み、(2) 話者の音量や口と送話機との位置関係（距離、角度）などの音圧レベル変化による非線形な歪み。後者の影響は小さいものとして本報告では線形フィルターで対応できる部分のみを扱うことにする。ただし線形フィルターの周波数特性については時間的に変化しないものとする。

このような仮定をおいた上で、マイク入力のような標準周波数特性から得られた音声（以下、標準音声と言う）と電話周波数特性から得られた音声との間の音響特徴量 (LPC ケプストラム係数、Δ ケプストラム係数) について考察する。

2.1 LPC ケプストラム係数について

図1は、標準音声および標準音声から生成される電話音声を LPC 分析して各種音響特徴量を得る過程を示している。ここでは、LPC 分析によって LPC ケプストラム係数を求める過程について考察する。

標準音声波形 (x_n) から線形フィルター (W) された 電話音声波形 $(\hat{x}_n)$ は、LPC 分析 (L) により $\hat{c}_n$ に変換される。波形 (x_n) から得られるものを c_n とする。この時、電話音声に対するピリオドグラム $\hat{P}(\lambda)$ は (x_n) のピリオドグラム $P(\lambda) = |X(e^{-j\lambda})|^2$ (ここで $X(z)$ は (x_n) の Z 変換多項式) と q 次の FIR 型線形フィルター $w(z) = \sum_{m=0}^{q} w_m z^{-m}$ $(w_0 = 1)$ により以下の式で計算される。

$$\hat{P}(\lambda) = P(\lambda)|w(e^{-j\lambda})|^2 = P(\lambda)W(\lambda). \tag{1}$$

ここで、$W(\lambda) = |w(e^{-j\lambda})|^2$ である。これと同様の関係が LPC スペクトルの間にも成り立つと仮定し、

$$L(\hat{P}(\lambda)) = L(P(\lambda))W(\lambda), \tag{2}$$

とする。この時、電話音声の対数 LPC スペクトラムは

$$\log \hat{f}(\lambda) = \log L(\hat{P}(\lambda)) = \log L(P(\lambda)) + \log W(\lambda),$$

$$\log \hat{f}(\lambda) = \log f(\lambda) + \log W(\lambda). \tag{3}$$

のように、標準音声の対数スペクトラムと w_n の対数スペクトラムの和として求められる。

LPC ケプストラム係数 c_n は対数 LPC スペクトラム $\log f(\lambda)$ のフーリエ係数として計算される。従って、電話音声に対するケプストラム係数 $\hat{c}_n$ は式 (3) を用いて以下の式で計算される。

$$\begin{aligned}
\hat{c}_n &= \int_{-\pi}^{\pi} \log \hat{f}(\lambda) e^{jn\lambda} \frac{d\lambda}{2\pi} && (4)\\
&= \int_{-\pi}^{\pi} \log(f(\lambda)W(\lambda)) e^{jn\lambda} \frac{d\lambda}{2\pi} \\
&= \int_{-\pi}^{\pi} \log f(\lambda) e^{jn\lambda} \frac{d\lambda}{2\pi} + \int_{-\pi}^{\pi} \log W(\lambda) e^{jn\lambda} \frac{d\lambda}{2\pi} \\
&= c_n + \int_{-\pi}^{\pi} \log W(\lambda) e^{jn\lambda} \frac{d\lambda}{2\pi}. && (5)
\end{aligned}$$

従って、h_n を

$$h_n = \int_{-\pi}^{\pi} \log W(\lambda) e^{jn\lambda} \frac{d\lambda}{2\pi}, \tag{6}$$

とおけば、$\hat{c}_n$ は

$$\hat{c}_n = c_n + h_n, \tag{7}$$

のように求められる。式 (6) から $-h_n$ は q 次の全極型スペクトル $1/W(\lambda)$ に対する LPC ケプストラム係数であることが分かる。$W(\lambda) = |w(e^{-j\lambda})|^2, w(z) = \sum_{m=0}^{q} w_m z^{-m}$ $(w_0 = 1)$ であるので、w_m を線形予測係数としてケプストラム係数を計算すれば、$-h_n$ が求められることになる。

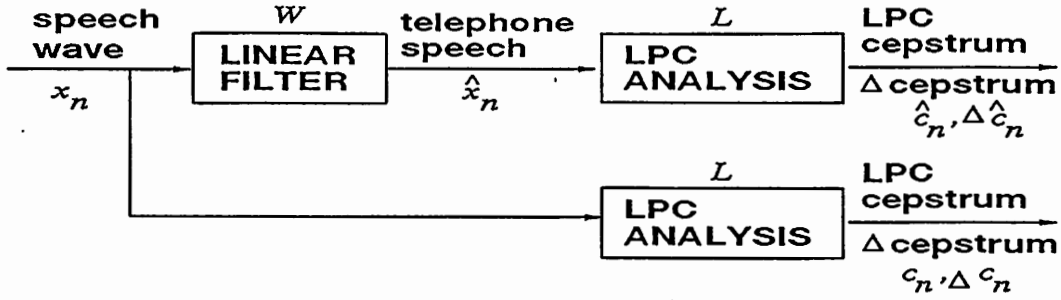


図 1: 電話音声作成の手順

2.2 対数パワーについて

本来、対数パワーは

$$\hat{u} = \int_{-\pi}^{\pi} \hat{f}(\lambda) \frac{d\lambda}{2\pi}, \tag{8}$$

なる式で定義される $\hat{u}$ の対数であり、式 (4) において $n = 0$ とおいた c_0 とは異なった量である。

電話音声のパワー $\hat{u}$ は

$$\hat{u} = \int_{-\pi}^{\pi} f(\lambda) W(\lambda) \frac{d\lambda}{2\pi}, \tag{9}$$

のように表される。一方、r_n を LPC 分析における標準音声の相関係数とすれば、$f(\lambda)$ および $W(\lambda)$ はそれぞれ

$$f(\lambda) = u \sum_{n=-\infty}^{+\infty} r_n e^{-jn\lambda}, \tag{10}$$

$$W(\lambda) = |\sum_{n=0}^{q} w_n e^{-jn\lambda}|^2, \tag{11}$$

で与えられる。すると、$\hat{u}$ は

$$\hat{u} = \int_{-\pi}^{\pi} f(\lambda) W(\lambda) \frac{d\lambda}{2\pi} = u \sum_{l=m-q}^{m} r_l \sum_{m=0}^{q} w_{m-l} w_m, \tag{12}$$

のように表され、電話音声に対する対数パワーは

$$\log \hat{u} = \log u + \log \left(\sum_{l=m-q}^{m} \sum_{m=0}^{q} r_l w_{m-l} w_m \right), \tag{13}$$

のように求められる。

2.3 Δ ケプストラム係数について

Δ ケプストラム係数は、ケプストラム係数の時系列を線形回帰分析した時の重み付き回帰直線の傾きとして求められる。重み付け回帰窓として、三角窓が一般的に用いられており、これを $a_n = -|n| + N, (n = -N, -N+1, .., 0, .., N-1, N)$ とすると、電話音声の第 i フレームにおける Δ ケプストラム係数 $\Delta\hat{c}_n^{(i)}$ は

$$\Delta\hat{c}_n^{(i)} = \frac{\sum_{k=-N}^{N} k a_k \hat{c}_n^{(i+k)}}{\sum_{k=-N}^{N} k^2 a_k} \tag{14}$$

のように求まり、式 (7) を用いることによって

$$\widehat{\Delta c}_n = \Delta c_n. \tag{15}$$

なる関係が導かれる。従って、本手法において電話音声に対する Δ ケプストラム係数は補正する必要がなく、対応する Δ ケプストラム係数をそのまま使用できることが分かる。

2.4 Δ 対数パワーについて

電話音声に対する Δ 対数パワー ($\Delta \log \hat{u}$) についても、前節と同様にして

$$\begin{aligned} \Delta \log \hat{u}^{(i)} &= \frac{\sum_{k=-N}^{N} k a_k \log \hat{u}^{(i+k)}}{\sum_{k=-N}^{N} k^2 a_k} \\ &= \Delta \log u^{(i)} + \frac{\sum_{k=-N}^{N} k a_k \sum_{m=0}^{q} w_m \sum_{l=m-q}^{m} r_l^{(i+k)} w_{m-l}}{\sum_{k=-N}^{N} k^2 a_k} \end{aligned} \tag{16}$$

のように求められる。ここで、添字の (i) は i フレームを表す。

3 混合連続分布 HMM への適用

LPC ケプストラム係数、Δ ケプストラム係数を特徴量とし、音素を混合連続分布型 HMM $M = (\pi, A, B)$, $A = (A_i)$, $B = (B_i)$, $B_i = \sum_{m=1}^{M} w_m N(\mu_{i,m}, \Sigma_{i,m})$ で表すことにする。この時、$\mu_{i,m}$, $\Sigma_{i,m}$ はベクトル量であり、LPC ケプストラムの分析次数を p とすれば、それぞれ $\mu_{i,m} = {}^t(\mu_{i,m}^{(0)}, \mu_{i,m}^{(1)}, ..., \mu_{i,m}^{(p)}, \mu_{i,m}^{(p+1)}, \mu_{i,m}^{(p+2)}, ..., \mu_{i,m}^{(2p+2)})$, $\Sigma_{i,m} = {}^t(\Sigma_{i,m}^{(0)}, \Sigma_{i,m}^{(1)}, ..., \Sigma_{i,m}^{(p)}, \Sigma_{i,m}^{(p+1)}, \Sigma_{i,m}^{(p+2)}, ..., \Sigma_{i,m}^{(2p+2)})$ で与えられる。なお、添字 (0) は対数パワーに、(1) ～ (p) は LPC ケプストラムに、$(p+1)$ は Δ 対数パワーに、$(p+2)$～$(2p+2)$ は Δ ケプストラムに対応している。

ここでは、標準音声音素 HMM の電話音声音素 HMM への適用方法について考察する。

3.1 h_n を用いた標準音声音素 HMM のケプストラム項の平均値の補正

式 (7) から分かるように、まず h_n を用いて LPC ケプストラム係数を補正することが考えられる。これは図 2 に示すように、標準音声音素 HMM のケプストラム項の平均のみを補正することに対応し、適応後は $\widehat{M} = (\pi, A, \widehat{B})$, $\widehat{B}_i = \sum_{m=1}^{M} w_m N(\mu_{i,m} + \hat{\mu}, \Sigma_{i,m})$ と表される。ここで $\hat{\mu}$ を $\hat{\mu} = {}^t(\hat{\mu}^{(0)}, \hat{\mu}^{(1)}, ..., \hat{\mu}^{(p)}, \hat{\mu}^{(p+1)}, \hat{\mu}^{(p+2)}, ..., \hat{\mu}^{(2p+2)})$ と表せば、ここでの補正項は $\hat{\mu}_{i,m} = {}^t(0., h_1, h_2, ..., h_p, 0., 0., ..., 0.)$ となる。

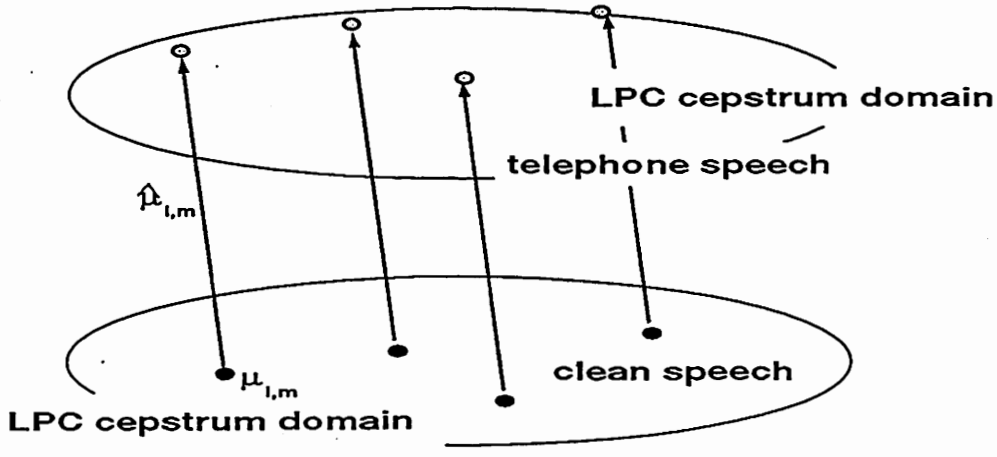


図 2: h_n を用いたケプストラム領域における標準音声音素 HMM の補正の概念

3.2 標準音声音素 HMM のケプストラム項の平均値・分散の補正

前節の適応方法は式 (7), (15) に基づくものであり、そこには幾つかの仮定や近似 (2) がある。その近似の精度は h_n の分布を算出することで推定される。一方、文献 [6] は 2 つの HMM (音声 HMM と雑音 HMM) に関してガウス分布の合成手法を提案し、雑音下の音声認識における有効性を示している。そこで、誤差ベクトル h をガウス分布型 HMM で表現することにより上で述べた方式の性能の向上が期待できる。今その分布が単一ガウス分布 $N(\hat{\mu}, \hat{\Sigma})$ で表現され、しかも入力音声と無相関の場合、適応後の分布は $N(\mu_{im} + \hat{\mu}, \Sigma_{i,m} + \hat{\Sigma})$ のように表される。ここでは、$\hat{\mu} =^t (0., h_1, h_2, ..., h_p, 0., 0., ..., 0.)$ とし、$\hat{\Sigma} =^t (0., \hat{\Sigma}^{(1)}, ..., \hat{\Sigma}^{(p)}, 0., 0., ..., 0.)$ とする。この場合の概念図を図 3 に示す。

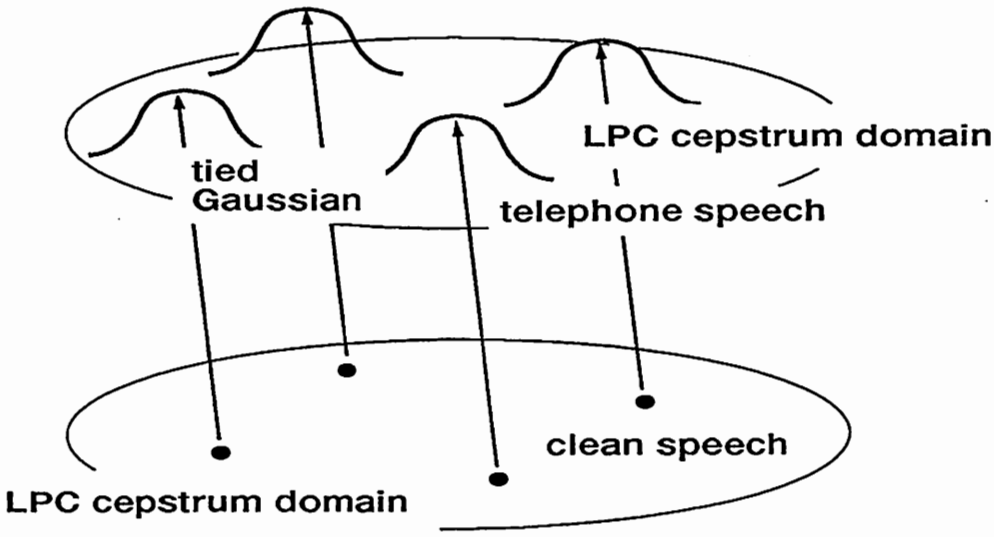


図 3: ケプストラム領域における標準音声音素 HMM の平均値と分散の補正の概念

本方法は文献 [6] のケプストラム領域に限定した応用と位置付けされる。一方、パワー項に関しては何も補正していないが、本来、適応後の分布としてパワー項も考慮する必要がある。この場合、適応後の分布における平均値および分散の補正項はそれぞれ、$\hat{\mu} =^t (\mu^{(0)}, h_1, h_2, ..., h_p, 0., 0., ..., 0.)$、$\hat{\Sigma} =^t (\hat{\Sigma}^{(0)}, \hat{\Sigma}^{(1)}, ..., \hat{\Sigma}^{(p)}, 0., 0., ..., 0.)$ そこで、次節では $\hat{\mu}^{(0)}$ と $\hat{\Sigma}^{(k)}; (k = 1, 2, ..., p)$ の推定方法について述べる。

3.3 標準音声音素 HMM における平均値と分散の推定

前節の標準音声音素 HMM の平均値と分散を補正において、まず対数パワー項の平均値の補正に関しては、式 (13) に従って求める必要がある。ここでは、学習単語のすべてのフレームに対して、電話音声の対数パワーから対応する標準音声のそれを引いたものの総和を計算し、それを総フレーム数で割ったものを対数パワー項の平均値の補正値として用いることにする。また、標準音声音素 HMM のケプストラム項の分散の補正に関しては、学習単語のすべてのフレームに対して電話音声のケプストラムベクトルから対応する標準音声のそれを引いたものの分散を次元毎に求め、次元毎のそれぞれの値を分散の補正値として用いる。なお、標準音声音素 HMM の対数パワー項の分散の補正に関しては、ケプストラム項の分散の補正方法と同様である。

4 認識実験による性能評価

4.1 音素認識性能

本節では、これまで述べてきた標準音声音素 HMM の補正の仕方が電話入力音声の音素認識性能にどのように反映されるかについて考察する。

4.1.1 実験条件

図4に実験に用いた電話音声作成用フィルタ (W) の周波数特性を示す。ここで、FIR フィルタの次数は $q = 31$ であり、その係数を付録 A.1 に示す。また、音素認識実験の条件を表1に示す。

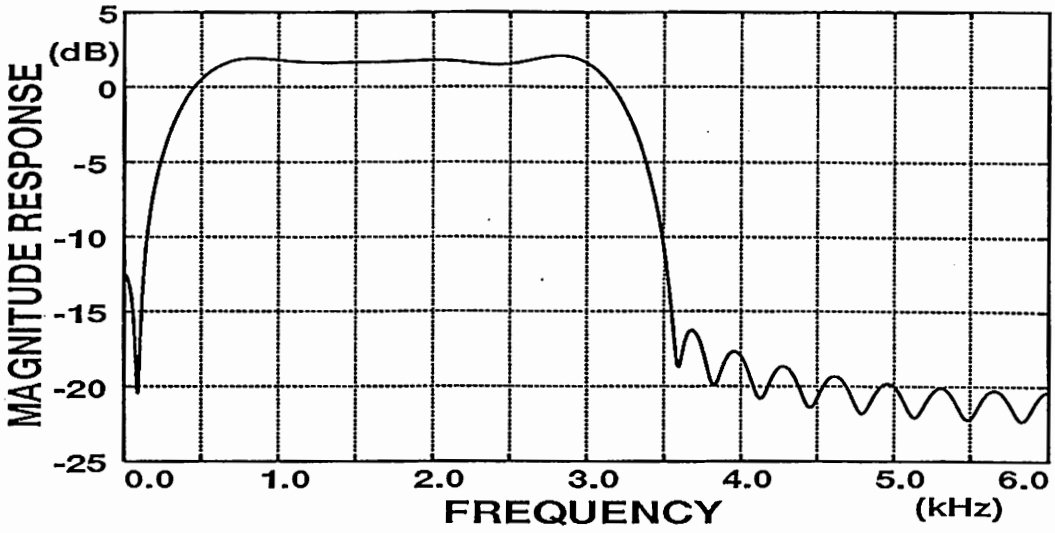


図 4: 電話音声作成用時不変型線形フィルタ (W) の周波数特性

表 1: 24 音素認識実験条件

LPC 分析次数	$p = 16$
フィルター次数	$q = 31$
音素モデル数	25
HMM 状態数	4 状態 (3 loop)
混合数	10
話者数	男性 1 名
学習音声	5240 単語 (奇数)
評価音声	5240 単語 (偶数)

4.1.2 h_n を用いた標準音声音素 HMM のケプストラム項の平均値の補正による効果

ここでは、3.1 節で述べた標準音声音素 HMM のケプストラム項における平均値のみを $h_n(n = 1, 2, ..., p\ ; p = 16)$ を用いて補正した場合の認識性能について考察する。ただし、$34(= 2p + 2)$ 次元から成る特徴ベクトルには対数パワー、Δ 対数パワー、Δ ケプストラムを含むが、これらの補正は本節では行なっていない。認識実験結果を表2に示す。なお、h_n については、付録 A.2 に記す。

表 2: h_n を用いた補正による認識性能

方式		認識率 (%)		
		24 音素	子音	母音
標準音声入力		98.0(99.9)	97.5(99.9)	98.4(99.9)
電話音声入力	特性補正なし	27.5(80.2)	9.5(59.3)	44.4(99.7)
	ケプストラム項の平均値の補正	62.5(92.3)	46.0(86.6)	78.0(97.5)
	パワー・ケプストラム項の平均値の補正	62.6(92.5)	46.3(87.1)	77.8(97.6)

( ) : 第5位までの累積認識率

マイク入力音声に対して認識率 98.0% の標準音声音素 HMM を用いた電話音声中の音素認識率は 27.5% に低下した。h_n を用いて標準音声音素 HMM のガウス分布の平均値を補正することにより、62.5% の認識率に向上した。さらに、補正の効果を音素毎に見るために、図 5 に補正前 (without normalization) と補正後 (MEAN(h_n)) の認識性能の比較を示す。これより、/u/ および /j/ を除く各音素に対しては認識率が改善されることが分かる。

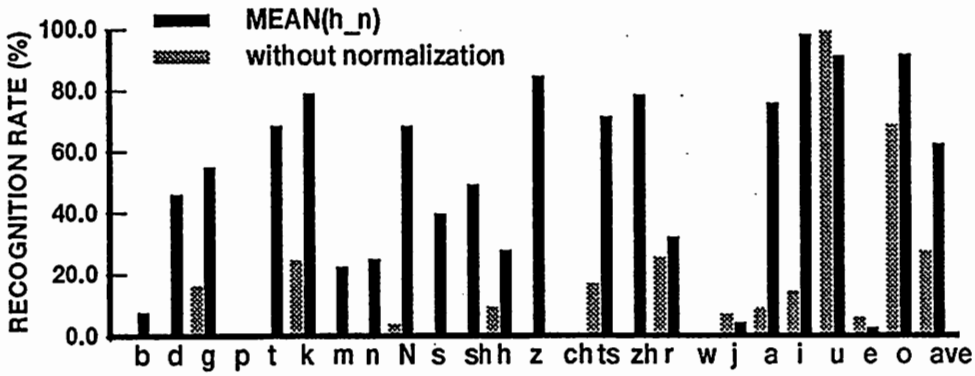


図 5: h_n を用いたケプストラム項の平均値の補正による音素毎の効果

4.1.3 標準音声音素 HMM のケプストラム項の平均値・分散の補正による効果

ここでは、差分ベクトル h が単一ガウス分布によって表現されるものとして、その平均値と分散を用いて標準音声音素 HMM の補正を行なう。表 3 に認識実験結果を示す。なお、$\hat{\mu}^{(0)}$ と $\hat{\Sigma}^{(k)}$, $(k = 0, 1, ..., 16)$ の推定値は付録 A.3 に記す。

表 3: 平均値と分散の補正による認識性能

方式		認識率 (%)		
		24 音素	子音	母音
標準音声入力		98.0(99.9)	97.5(99.9)	98.4(99.9)
電話音声	特性補正なし	27.5(80.2)	9.5(59.3)	44.4(99.7)
	ケプストラム項の平均値・分散の補正	74.7(97.2)	61.7(96.5)	86.8(97.8)

( ) : 第 5 位までの累積認識率

これより、ケプストラム項の平均値のみならず分散も補正することによって、前節に比べて認識性能がさらに改善されることが分かる。補正の効果を音素毎に見るために、図 6 に補正前 (without normalization) と補正後 (with normalization) の認識性能の比較を示す。ケプストラムの平均値の補正だけでは全く認識されなかった /p/,/w/,/ch/ が分散も補正することによってかなり認識されていることが分かる。また /b/,/m/,/n/,/e/ についても改善効果は大きいと思われる。

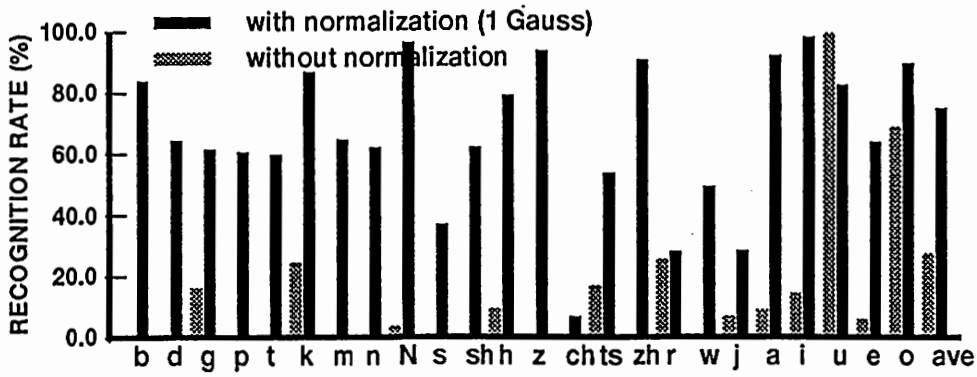


図 6: ケプストラム項の平均値・分散の補正による音素毎の効果

4.1.4 標準音声音素 HMM のパワー・ケプストラム項の補正による効果

前節まではケプストラム項のみの補正を行なってきたが、本節ではケプストラム項の補正とパワー項の補正を併用した場合の効果について考察する。表 4 に認識実験結果を示す。また、パワー項を補正する効果を音素毎に見るために、図 7 にケプストラム項のみの補正とパワー・ケプストラム項の補正に対する認識性能の比較を示す。

表 4: 平均値と分散の補正による認識性能

方式		認識率 (%)		
		24 音素	子音	母音
標準音声入力		98.0(99.9)	97.5(99.9)	98.4(99.9)
電話音声入力	特性補正なし	27.5(80.2)	9.5(59.3)	44.4(99.7)
	ケプストラム項の平均値・分散の補正	74.7(97.2)	61.7(96.5)	86.8(97.8)
	パワー・ケプストラム項の平均値・分散の補正	74.2(96.8)	63.2(96.3)	84.4(97.3)

( ) : 第 5 位までの累積認識率

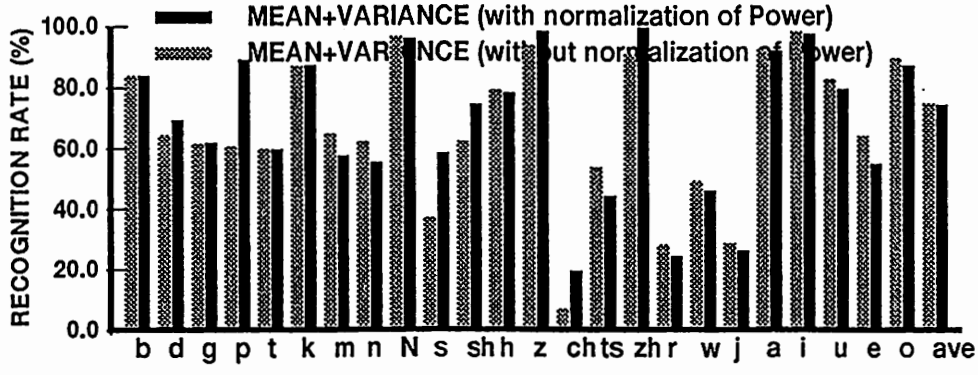


図 7: パワー・ケプストラム項の平均値・分散の補正による音素毎の効果

これより、音素毎に見た場合には、/d/,/g/,/p/,/k/,/s/,/ch,/sh/,/z/,/zh/ のみ認識率の改善がなされ, パワー項の補正の効果が子音に対して認められることが分かる。

4.1.5 標準音声音素 HMM のパワー・ケプストラム項の平均値の補正と分散の拡大による効果

前節までの結果から、ケプストラム項の平均値および分散の補正は認識性能の向上に有効であり、パワー項の補正は認識性能の改善にそれほど寄与していないことが分かった。特にケプストラム項の平均値の補正が認識性能の改善に大きく寄与しているものと考えられる。

そこで、パワー・ケプストラム項の平均値の補正は行なわずに分散のみを定数倍することによって補正した場合の認識結果と、パワー・ケプストラム項の平均値の補正を行ない、しかも分散を定数倍することによって補正した場合の認識結果を図 8 に示す。

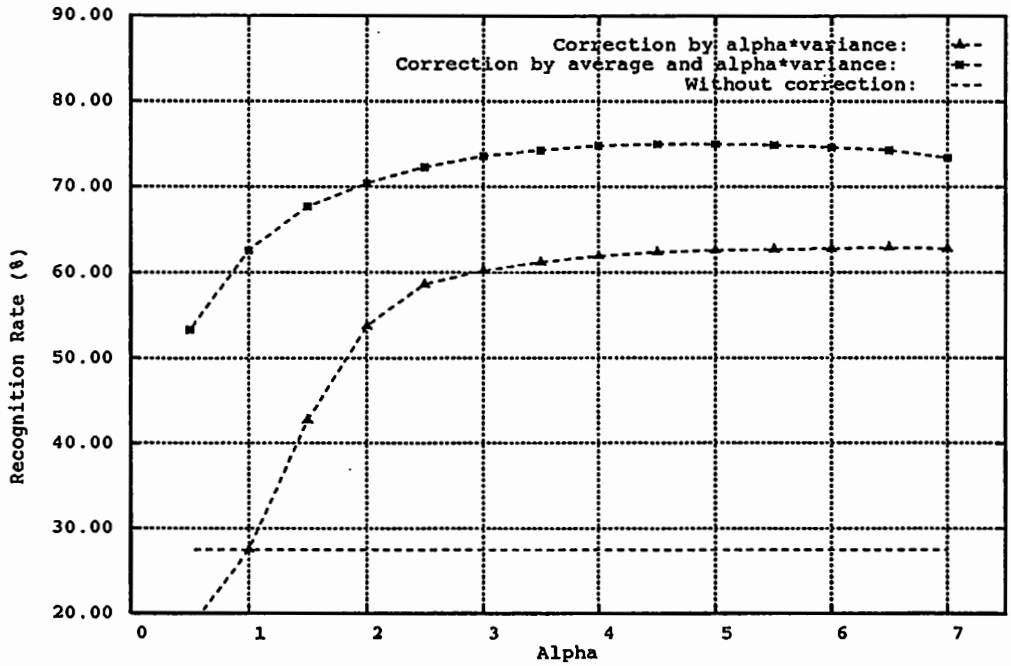


図 8: 分散の拡大補正による効果

これより、分散を拡大するだけでも改善効果が認められることが分かる。しかしながら、単に分散を拡大するだけでは平均値の補正と分散の拡大を併用する場合に比べて認識性能が低くなる。図 8 に最大の認識率を得た場合のそれぞれの結果を示す。

表 5: 分散の拡大補正による認識性能 (平均値の補正ある・なし)

方式		認識率 (%)		
		24 音素	子音	母音
標準音声入力		98.0(99.9)	97.5(99.9)	98.4(99.9)
電話音声入力	特性補正なし	27.5(80.2)	9.5(59.3)	44.4(99.7)
	$\alpha * \Sigma_{i,m}$ $(\alpha = 8.5)$	63.0(88.4)	40.1(77.0)	84.3(99.0)
	平均値と $\alpha * \Sigma_{i,m}$ $(\alpha = 4.5)$	75.0(96.7)	61.7(96.1)	87.4(97.2)

( ) : 第 5 位までの累積認識率

4.1.6 平均値および分散の推定に必要な単語数についての考察

これまでパワー項の平均値 ($\hat{\mu}^{(0)}$) と分散 ($\hat{\Sigma}^{(k)}$, $k = 0, 1, ..., 16$) の推定には学習単語 (2620 単語) を用いていた。本節では、推定に必要な単語数と認識性能との関係について考察する。ただし、使用する単語は順序づけられた 2620 単語の先頭から使うことにする。表 6 は推定に用いる単語数を変化させた時の認識性能を示しており、図 9 はそれをグラフで表したものである。

表 6: 推定に用いた単語数と 24 音素認識性能の関係

単語数	1	2	3	4	5	6	7
認識率 (%)	74.1 (96.8)	74.5 (97.1)	74.5 (97.2)	74.6 (97.0)	74.5 (96.9)	74.6 (96.9)	74.7 (96.9)
単語数	8	9	10	25	50	100	2620
認識率 (%)	74.8 (96.9)	74.8 (96.9)	74.8 (97.0)	74.8 (97.0)	74.8 (96.9)	74.6 (96.9)	74.2 (96.8)

( ): 第 5 位までの累積認識率

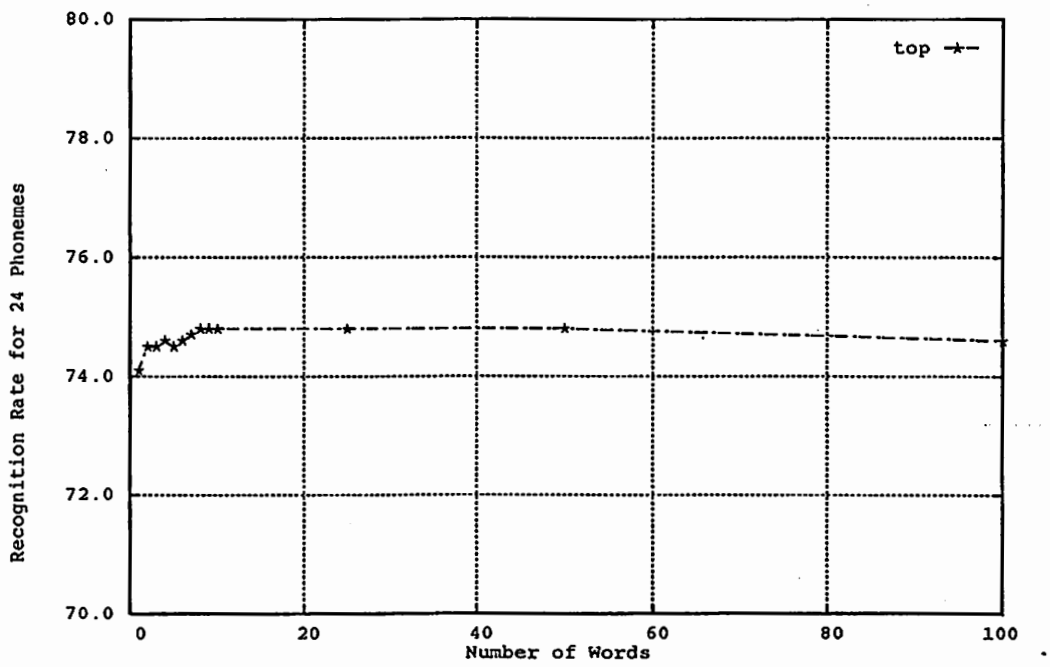


図 9: 推定に用いた単語数と 24 音素認識性能の関係

これより、学習単語をすべて用いて推定した場合と比較すると、推定に必要な単語数は 10 単語前後でよいことが分かる。

4.2 文節認識性能

文節認識実験条件を表 7 に、認識結果を表 8 に示す。この場合の補正値は、式 (6) による h_n と 3.2 節および 3.3 節で求めた $\hat{\mu}$ と $\hat{\Sigma}$ を用いた。なお、混合連続分布型 HMM に関する実験条件の詳細は文献 [7] に述べられている。

表 7: 文節認識実験条件

LPC 分析次数	p = 16
フィルター次数	q = 31
HMM 状態数	4 状態 (3 loop)
混合数	3 〜 15
継続時間長制御	なし
話者数	男性 1 名
学習音声	5240 単語 & 216 音素バランス単語
評価音声	279 文節

表 8: 文節認識結果

方式		認識率 (%)
標準音声入力		87.5 (98.6)
電話音声入力	特性補正なし	2.9 (8.2)
	平均値と分散の補正	47.0 (68.5)

( ): 第 5 位までの累積認識率

これより、文節認識実験においても認識性能の改善が顕著に行なわれていることが分かる。

5 考察

ここでは、まずパワー・ケプストラム項の平均値および分散の補正が音素認識性能にどのように反映されたかを総合的に見ることにする。表 9 はそれをまとめたものである。

表 9: 音素認識性能評価の比較

方式		認識率 (%)		
		24 音素	子音	母音
標準音声入力		98.0(99.9)	97.5(99.9)	98.4(99.9)
電話音声入力	特性補正なし	27.5(80.2)	9.5(59.3)	44.4(99.7)
	ケプストラム項の平均値の補正	62.5(92.3)	46.0(86.6)	78.0(97.5)
	パワー・ケプストラム項の平均値の補正	62.6(92.5)	46.3(87.1)	77.8(97.6)
	ケプストラム項の平均値・分散の補正	74.7(97.2)	61.7(96.5)	86.8(97.8)
	パワー・ケプストラム項の平均値・分散の補正	74.2(96.8)	63.2(96.3)	84.4(97.3)
	$\alpha * \Sigma_{i,m}(\alpha = 8.5)$	63.0(88.4)	40.1(77.0)	84.3(99.0)
	平均値と $\alpha * \Sigma_{i,m}(\alpha = 4.5)$	75.0(96.7)	61.7(96.1)	87.4(97.2)
	電話音声で学習した HMM を使用	97.3(99.9)	96.4(99.9)	98.1(99.9)

( ) : 第 5 位までの累積認識率

これより、h_n を用いたケプストラム項の平均値の補正は有効である、h_n を用いたケプストラム項の平均値の補正と推定値を用いた分散の補正を併用するとさらに有効である、ことが分かった。特に、平均値の補正が有意であり、分散に関してはその拡大効果から分かるように、一率に定数倍を掛けておけば認識性能はある程度維持できると言える。一方、パワー項の平均値および分散の補正に関しては、それほど有意性が認められなかった。その理由の一つとして、付録 B に示すように、学習単語のすべてのフレームに対して電話音声の対数パワーから標準音声のそれを引いたものの分布において、特に対数パワー項に関しては単一ガウス分布で表現するのに少し無理があることによるものと考えられる。従って、パワー項の補正に関しては、混合ガウス分布で表現し直し、推定を行なう方が良いと思

われる。また、本稿で述べた補正は HMM 音素カテゴリに依存せずに行なったが、本来音素毎に補正値を変える必要があるかもしれない。

本稿では標準音声 HMM を補正したものを電話音声認識システムへ適用することによって、音素認識性能 27.5% だったものが 75.0% まで改善されてはいるが、実際電話音声を用いて学習した HMM を用いると、認識性能が 97.3% まで改善されることが表 9 の最下段に示されている。現段階で見るとまだ約 20% の改善効果の差があり、本手法の改良によってさらに改善効果が期待できるものと考えられる。

6 むすび

本稿では、標準音声を用いて学習した混合連続分布 HMM 音声認識システムを用いて、電話音声のように周波数特性の異なる系からの入力音声の認識手法を提案し、音素認識実験および文節認識実験により性能の改善効果を評価した。その結果、本手法の有効性が確認された。今後の課題としては、考察でも述べたように、パワー項の平均値と分散の補正に関しては、電話音声におけるパワー項と標準音声のそれとの差の分布を混合ガウス分布で表現し直し、推定を行なうこと、音素毎に補正値を適応的に変えること、それから周波数特性の逐次推定などが上げられる。

謝辞

研究の機会を与えて頂いた ATR 自動翻訳電話研究所 樽松明社長、熱心に御討論していただいた音声情報処理研究室の諸氏に感謝します。

参考文献

[1] K.Ohkura, M.Sugiyama, Shigeki Sagayama, Speaker Adaptation Based on Vector Field Smoothing with Continuous Mixture Density HMMs, ICSLP92, We.fAM.1.1, pp.369-372 (1992-10).

[2] J.G.Wilpon, A Study on the Ability to Automatically Recognize Telephone Quality Speech from Large Customer Populations, AT&T Tech. J., Vol.64, No.2, pp.423-451 (1985-02).

[3] 今村, HMM による電話音声のスポッティング, 音声研究会資料, SP90-18 (1990-06).

[4] 杉山, 鹿野, WLR 尺度による帯域制限大語彙単語音声認識, 音響学会講演論文集, 2-1-16,pp.107-108 (1981-10).

[5] S.Lerner, B.Mazor, Channel Normalization of Telephone Speech for Automatic Speech Recognition, Proc. of ICASSP92, 35.8 (Mar.1992).

[6] F. Martin, K. Shikano, Y. Minami and Y. Okabe, *Recognition of Noisy Speech by Using the Composition of Hidden Markov Models*, Proc. ASJ Fall Meeting, 1-7-10 (Oct. 1992).

[7] 山口, 蟇峨山, 混合連続分布型 HMM を用いた HMM-LR 連続音声認識, 音響学会講演論文集, 1-P-5, pp.113-114 (1993.3).

付録

A 本稿における種々のパラメータの値

A.1 電話音声作成のためのフィルタ係数

本文2.1節において述べたように、本稿で用いた電話音声波形 $(\hat{x}_n)$ は標準音声波形 (x_n) を線形フィルタ (W) することによって求めている。本稿ではフィルタの次数 (q) は $q=31$ としており、フィルタ係数を w_m $(m=0,1,...,31)$ で表すと、

$$\hat{x}_n = \sum_{m=0}^{31} w_m x_{n-m}$$

であり、w_m $(m=0,1,...,31)$ は以下の通りである。

$$\begin{aligned}
w_0 &= 1.000000, \quad w_1 = 2.840699, \quad w_2 = 2.920846, \quad w_3 = -0.339155, \quad w_4 = -3.722351 \\
w_5 &= -3.176896, \quad w_6 = -0.279907, \quad w_7 = 0.398203, \quad w_8 = -1.114378, \quad w_9 = -1.176012 \\
w_{10} &= 0.426425, \quad w_{11} = 0.765578, \quad w_{12} = -0.242964, \quad w_{13} = -0.191914, \quad w_{14} = 0.802882 \\
w_{15} &= 0.742443, \quad w_{16} = -0.098820, \quad w_{17} = -0.068892, \quad w_{18} = 0.511939, \quad w_{19} = 0.378448 \\
w_{20} &= -0.121702, \quad w_{21} = -0.074903, \quad w_{22} = 0.189524, \quad w_{23} = 0.081853, \quad w_{24} = -0.095944 \\
w_{25} &= -0.010605, \quad w_{26} = 0.073143, \quad w_{27} = -0.005502, \quad w_{28} = -0.041801, \quad w_{29} = 0.013173 \\
w_{30} &= 0.017875, \quad w_{31} = -0.015443
\end{aligned}$$

A.2 h_n $(n=1,2,...,16)$ の推定値

本文中の式(6)より、$-h_n$ $(n=1,2,...,16)$ は31次の全極型スペクトル $1/W(\lambda)$ に対するLPCケプストラム係数である。h_n は以下の通りである。

$$\begin{aligned}
h_1 &= 2.840699, \quad h_2 = -1.113939, \quad h_3 = -0.995326, \quad h_4 = 0.265898, \quad h_5 = -0.073151 \\
h_6 &= -0.671348, \quad h_7 = -0.248347, \quad h_8 = 0.047857, \quad h_9 = -0.302859, \quad h_{10} = -0.351183 \\
h_{11} &= -0.042177, \quad h_{12} = -0.074599, \quad h_{13} = -0.237539, \quad h_{14} = -0.110357, \quad h_{15} = -0.021600 \\
h_{16} &= -0.134629
\end{aligned}$$

A.3 $\hat{\mu}^{(0)}$ と $\hat{\Sigma}^{(k)}$ $(k=0,1,...,16)$ の推定値

ここでは、本文3.3節に基づいて求められた $\hat{\mu}^{(0)}$ と $\hat{\Sigma}^{(k)}$ $(k=0,1,...,16)$ の推定値を記す。これらの値は、学習単語(MAU5240単語の奇数番目の単語)のすべてのフレームに対する $(\hat{x}_n - x_n)$ の次元毎の平均値および分散である。

$$\begin{aligned}
\hat{\mu}^{(0)} &= -0.270457, \quad \hat{\Sigma}^{(0)} = 1.050285, \quad \hat{\Sigma}^{(1)} = 0.311527, \quad \hat{\Sigma}^{(2)} = 0.050514, \quad \hat{\Sigma}^{(3)} = 0.039047 \\
\hat{\Sigma}^{(4)} &= 0.031627, \quad \hat{\Sigma}^{(5)} = 0.019250, \quad \hat{\Sigma}^{(6)} = 0.020295, \quad \hat{\Sigma}^{(7)} = 0.016263, \quad \hat{\Sigma}^{(8)} = 0.013466 \\
\hat{\Sigma}^{(9)} &= 0.007275, \quad \hat{\Sigma}^{(10)} = 0.007966, \quad \hat{\Sigma}^{(11)} = 0.007112, \quad \hat{\Sigma}^{(12)} = 0.006512, \quad \hat{\Sigma}^{(13)} = 0.006765 \\
\hat{\Sigma}^{(14)} &= 0.004429, \quad \hat{\Sigma}^{(15)} = 0.004007, \quad \hat{\Sigma}^{(16)} = 0.003857
\end{aligned}$$

A.4 $\hat{\mu}^{(l)}$ と $\hat{\Sigma}^{(l)}$ $(l=17,18,...,33)$ の推定値

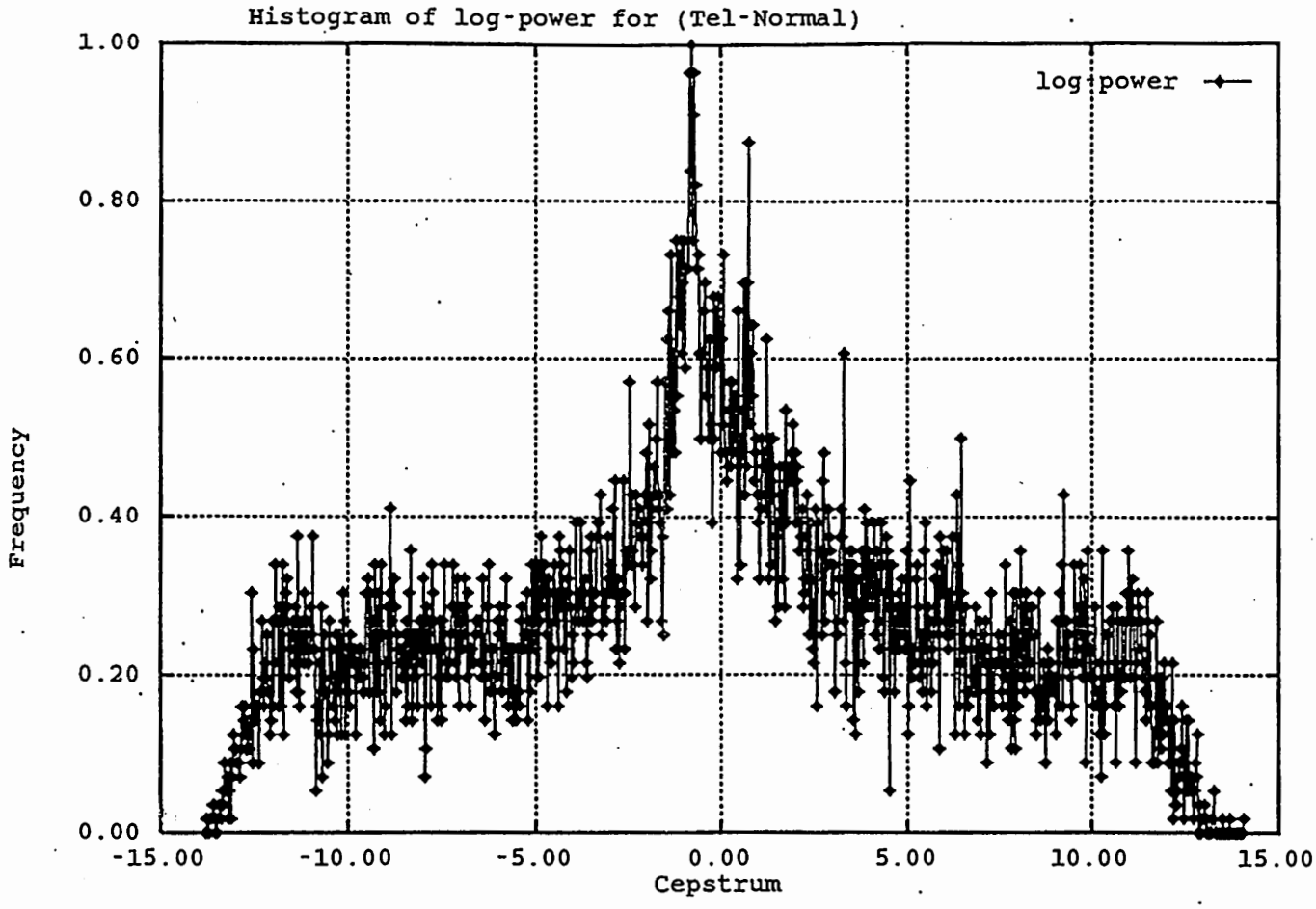
参考までに、学習単語(MAU5240単語の奇数番目の単語)のすべてのフレームに対する $(\hat{x}_n - x_n)$ のΔ対数パワーとΔケプストラムの平均値および分散を記しておく。

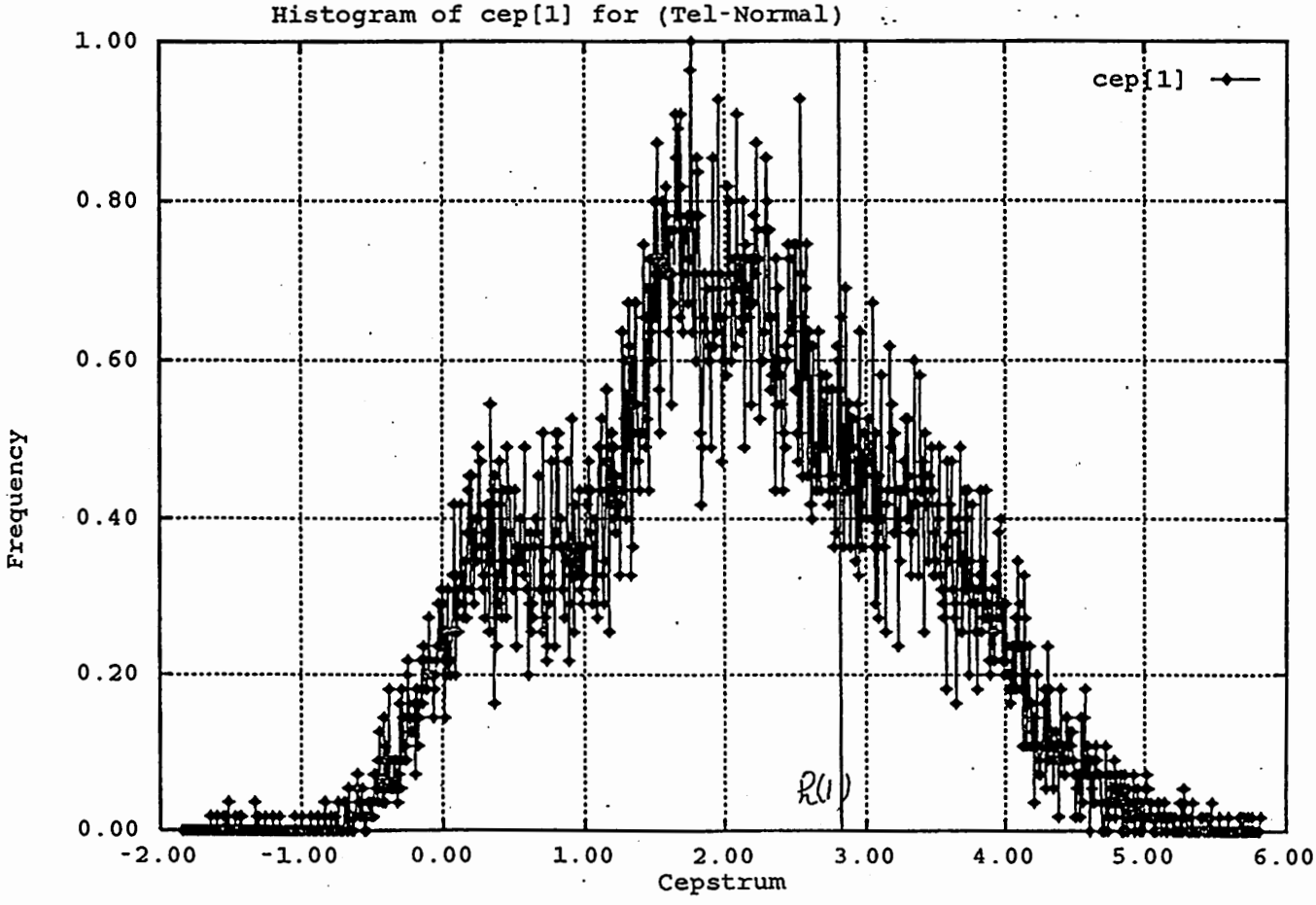
$$\begin{aligned}
\hat{\mu}^{(17)} &= -0.000020, \quad \hat{\mu}^{(18)} = 0.000017, \quad \hat{\mu}^{(19)} = 0.000005, \quad \hat{\mu}^{(20)} = 0.000004, \quad \hat{\mu}^{(21)} = 0.000006 \\
\hat{\mu}^{(22)} &= 0.000007, \quad \hat{\mu}^{(23)} = 0.000009, \quad \hat{\mu}^{(24)} = -0.000001, \quad \hat{\mu}^{(25)} = 0.000009, \quad \hat{\mu}^{(26)} = 0.000001
\end{aligned}$$

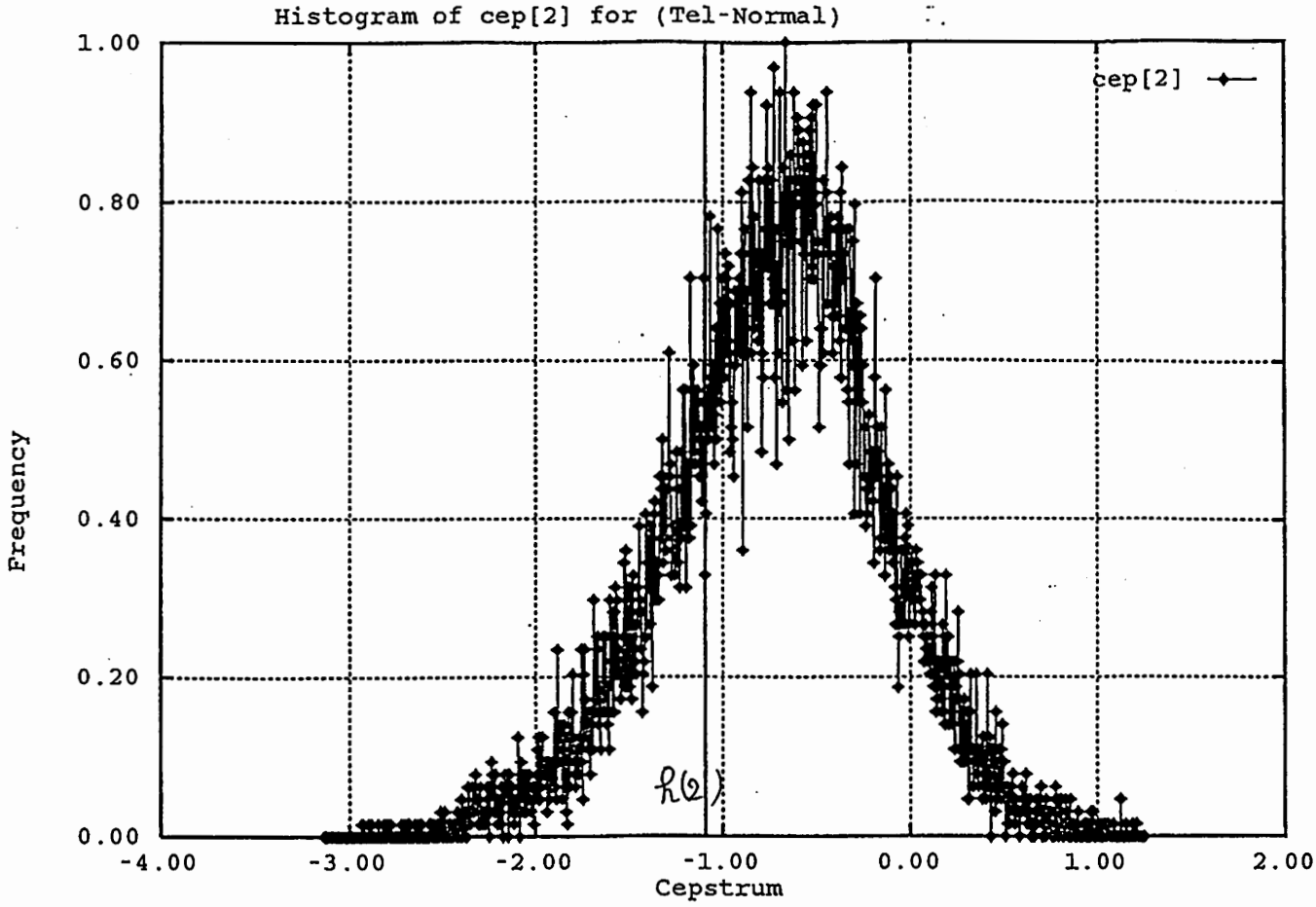
$$\widehat{\mu}^{(27)} = 0.000000,\quad \widehat{\mu}^{(28)} = 0.000004,\quad \widehat{\mu}^{(29)} = 0.000000,\quad \widehat{\mu}^{(30)} = 0.000001,\quad \widehat{\mu}^{(31)} = 0.000000$$
$$\widehat{\mu}^{(32)} = 0.000001,\quad \widehat{\mu}^{(33)} = -0.000002,\quad \widehat{\Sigma}^{(17)} = 0.006770,\widehat{\Sigma}^{(18)} = 0.001753,\quad \widehat{\Sigma}^{(19)} = 0.000289$$
$$\widehat{\Sigma}^{(20)} = 0.000275,\quad \widehat{\Sigma}^{(21)} = 0.000183,\quad \widehat{\Sigma}^{(22)} = 0.000104,\quad \widehat{\Sigma}^{(23)} = 0.000142,\quad \widehat{\Sigma}^{(24)} = 0.000113$$
$$\widehat{\Sigma}^{(25)} = 0.000084,\quad \widehat{\Sigma}^{(26)} = 0.000048,\quad \widehat{\Sigma}^{(27)} = 0.000049,\quad \widehat{\Sigma}^{(28)} = 0.000044,\quad \widehat{\Sigma}^{(29)} = 0.000041$$
$$\widehat{\Sigma}^{(30)} = 0.000042,\quad \widehat{\Sigma}^{(31)} = 0.000029,\quad \widehat{\Sigma}^{(32)} = 0.000025,\quad \widehat{\Sigma}^{(33)} = 0.000025$$

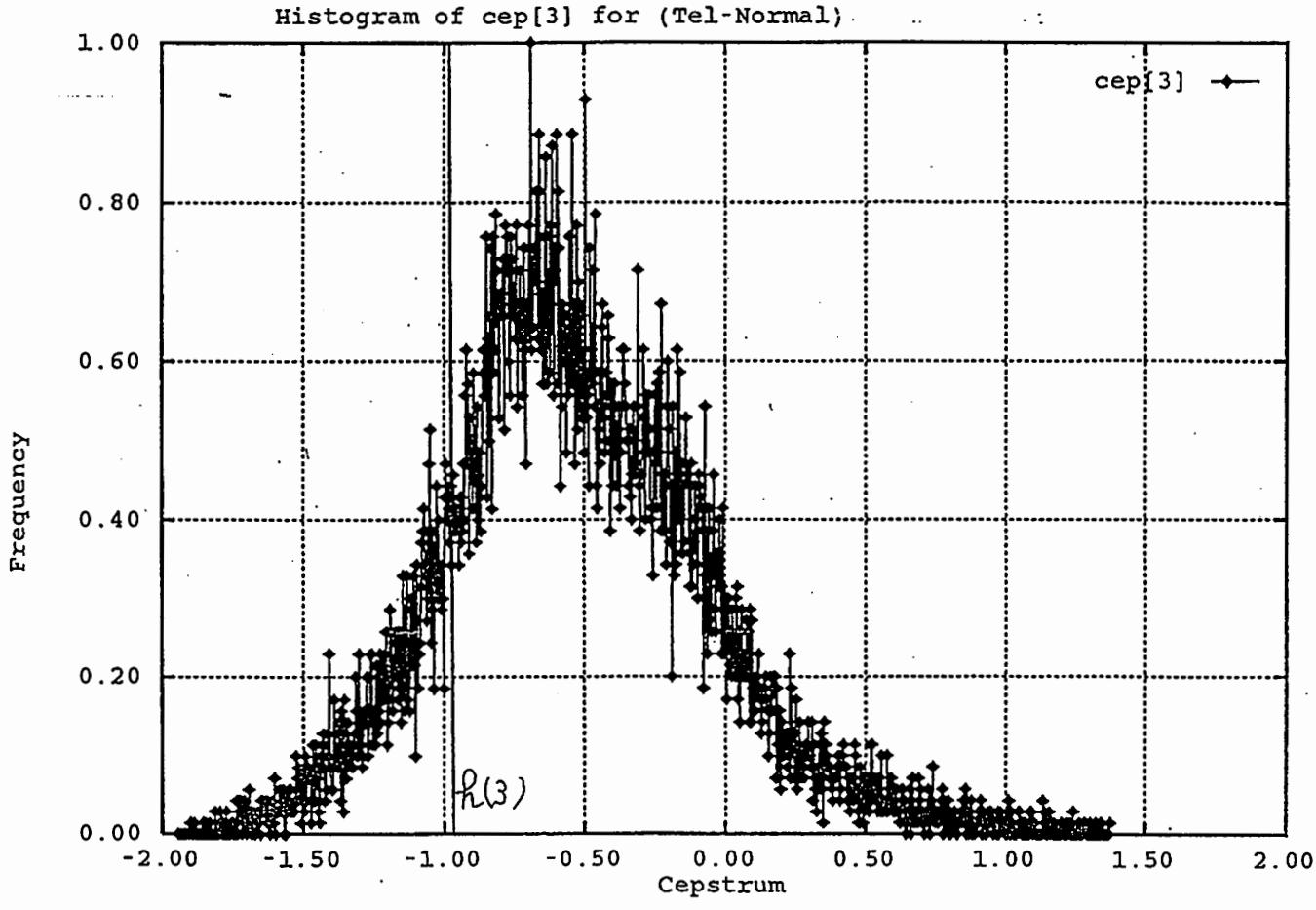
これより、それぞれの平均値および分散がほぼ零に近いとみなされるので、Δ対数パワーおよびΔケプストラム項の補正に関してほぼ必要ないと考えられる。このことは式(15)が実験的にも成立しているとも解釈できる。またΔ対数パワー項に関しても、式(16)の右辺第2項がほぼ無視できると考えられる。

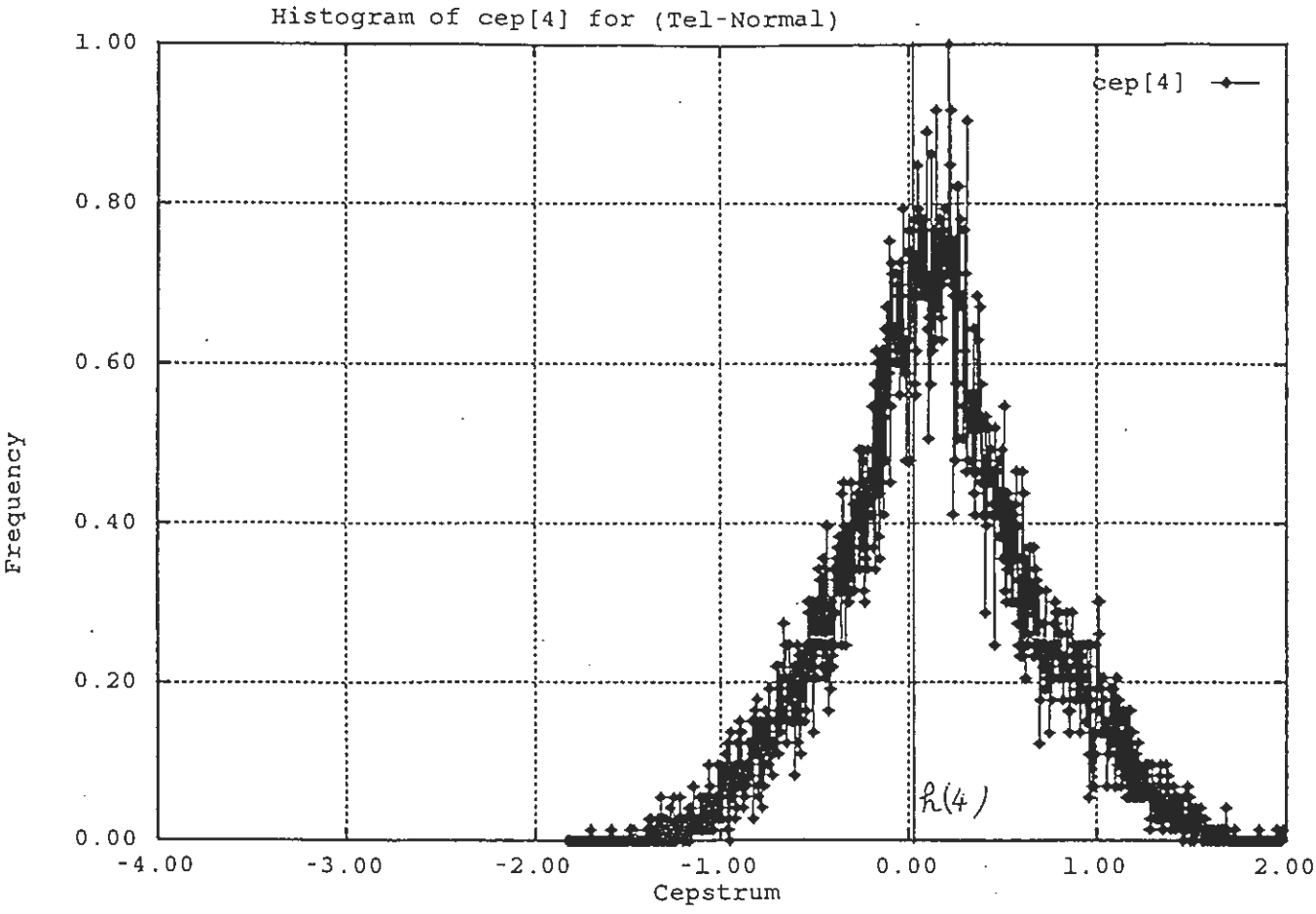
B $(\widehat{c}_n - c_n\ (n = 1, ..., 16))$ の対数パワーおよびケプストラム項の分布

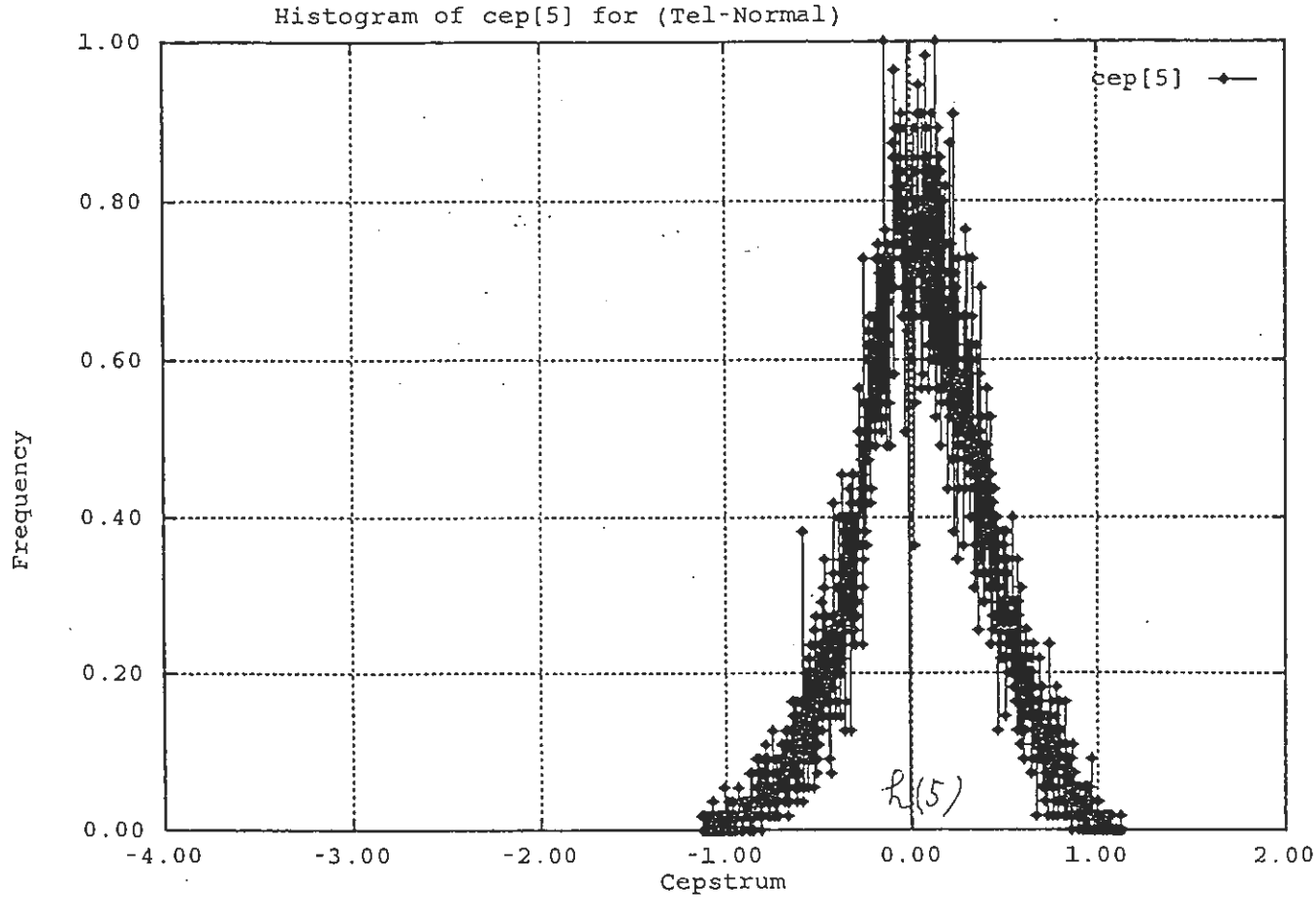


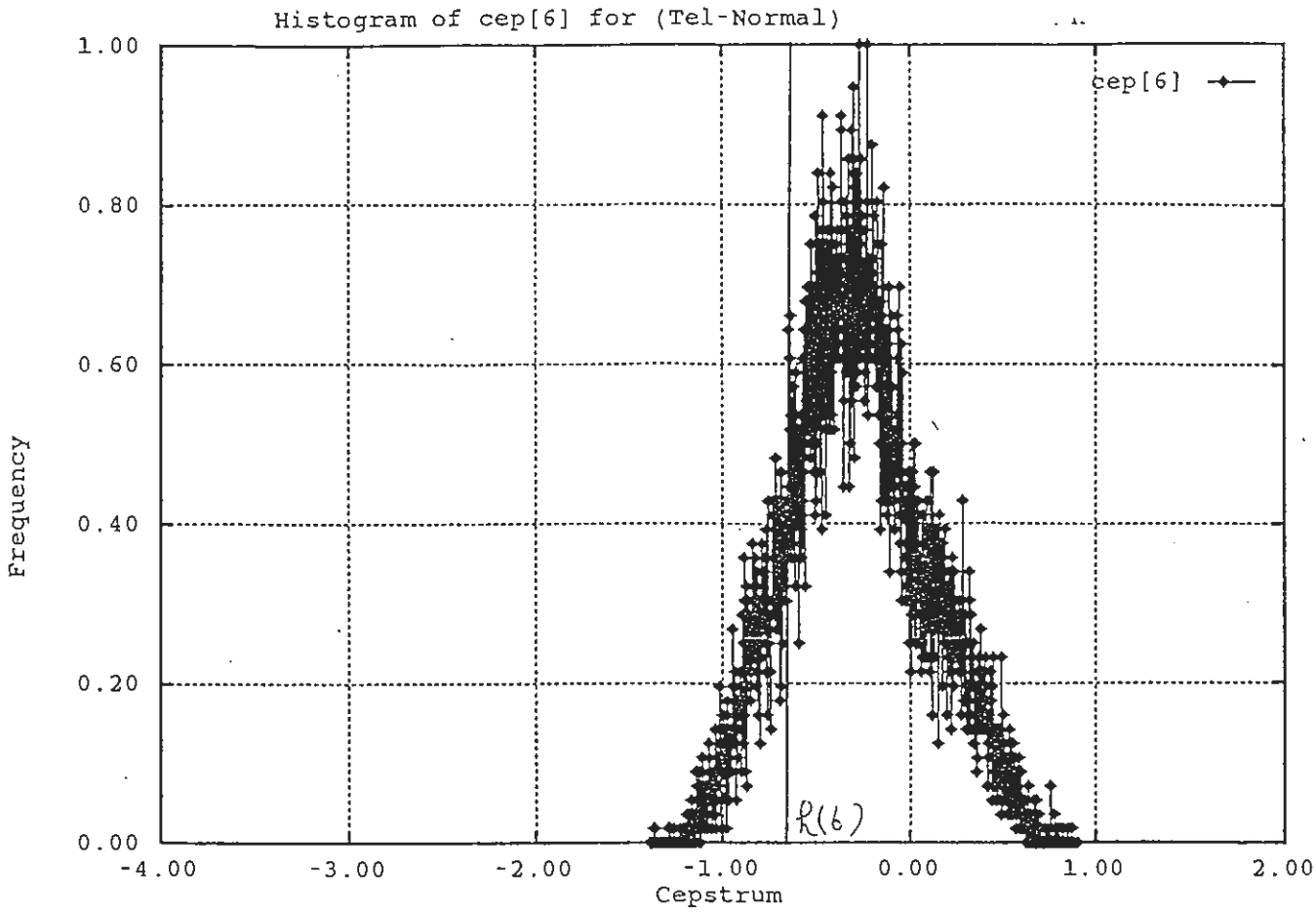


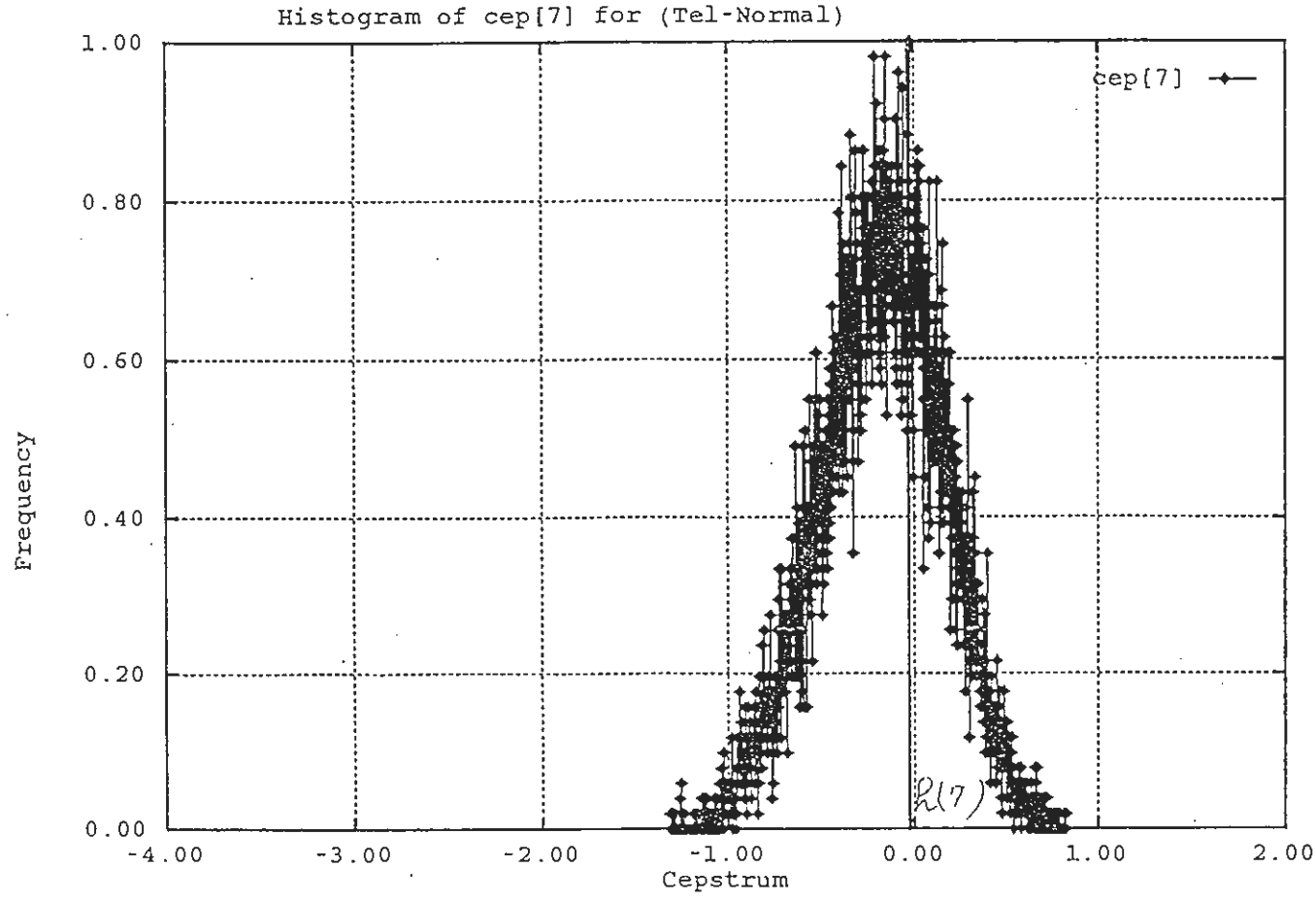

Histogram of cep[2] for (Tel-Normal)
cep[2]
Frequency
1.00
0.80
0.60
0.40
0.20
0.00
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum
h(2)

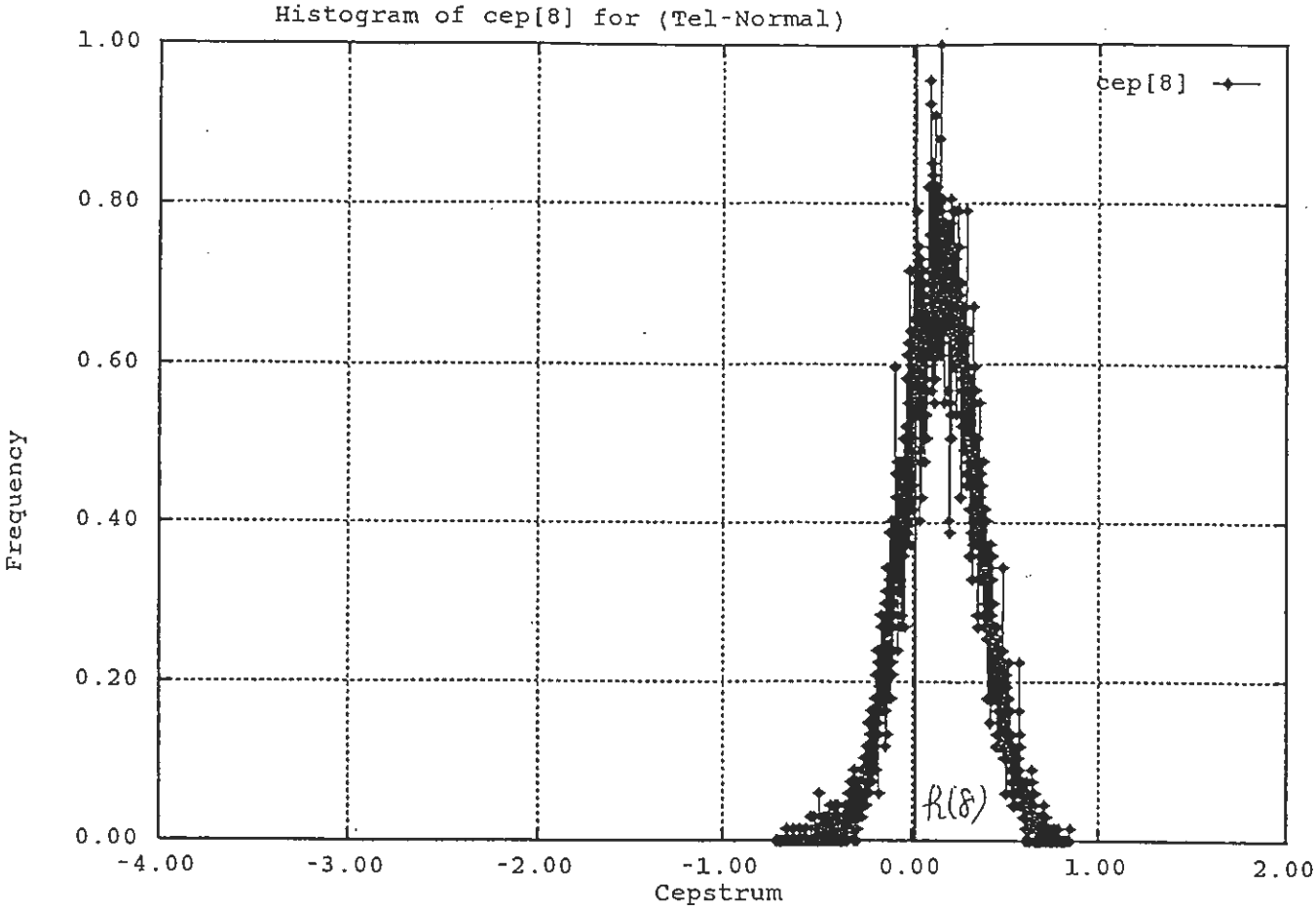

Histogram of cep[3] for (Tel-Normal)
cep[3]
Frequency
1.00
0.80
0.60
0.40
0.20
0.00
-2.00
-1.50
-1.00
-0.50
0.00
0.50
1.00
1.50
2.00
Cepstrum
h(3)

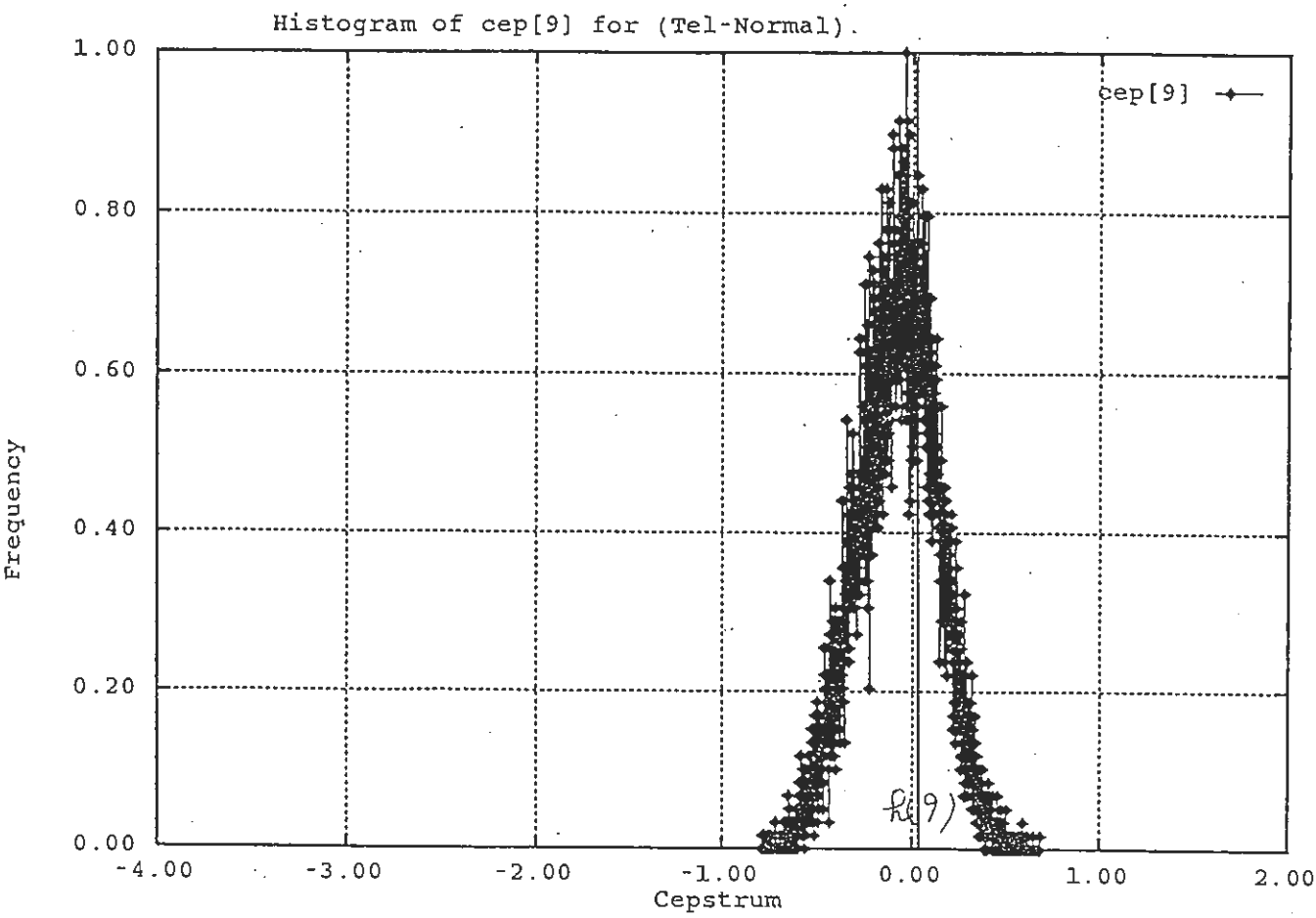

Histogram of cep[4] for (Tel-Normal)
cep[4]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum
h(4)

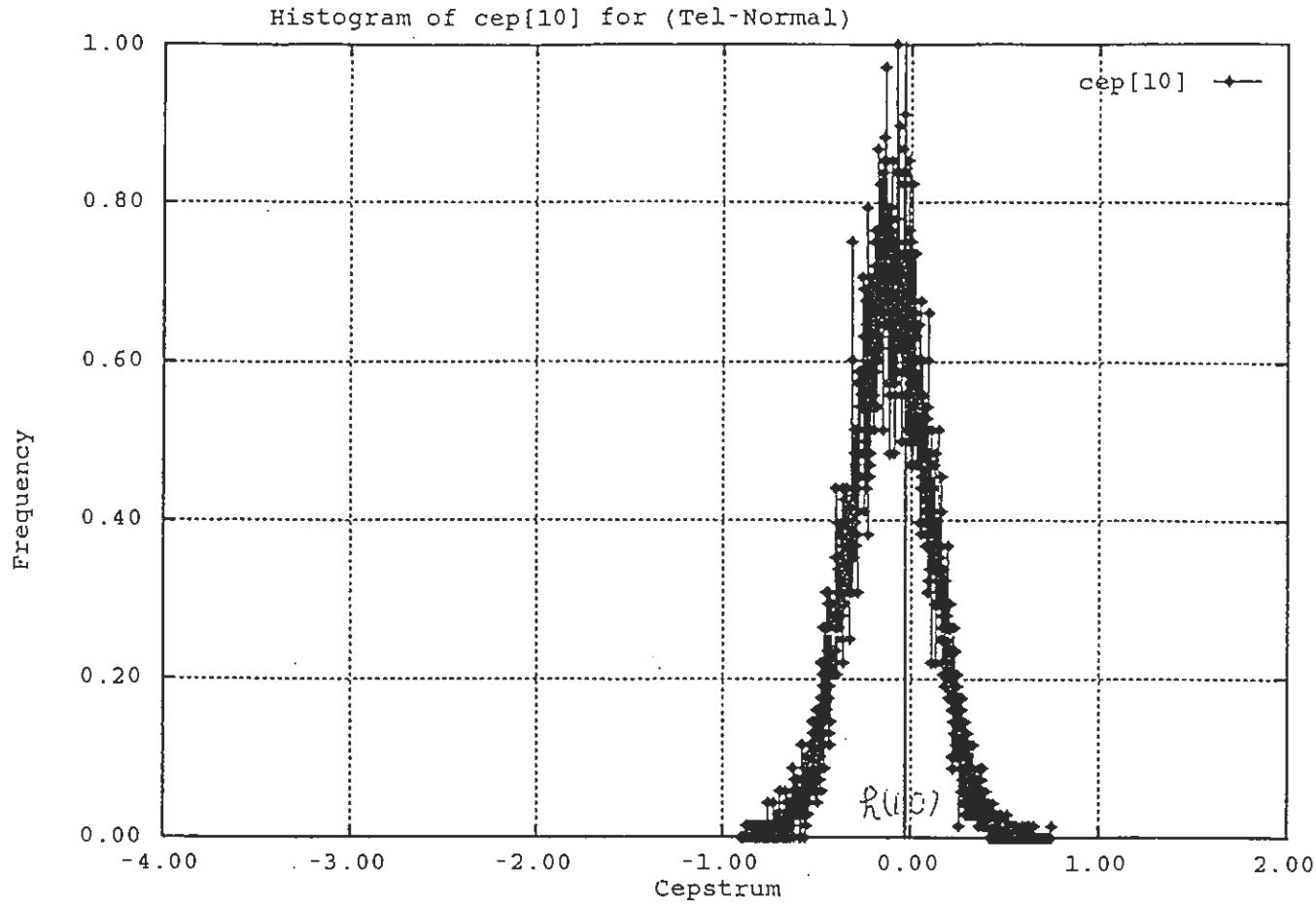

Histogram of cep[5] for (Tel-Normal)
cep[5]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum
h(5)

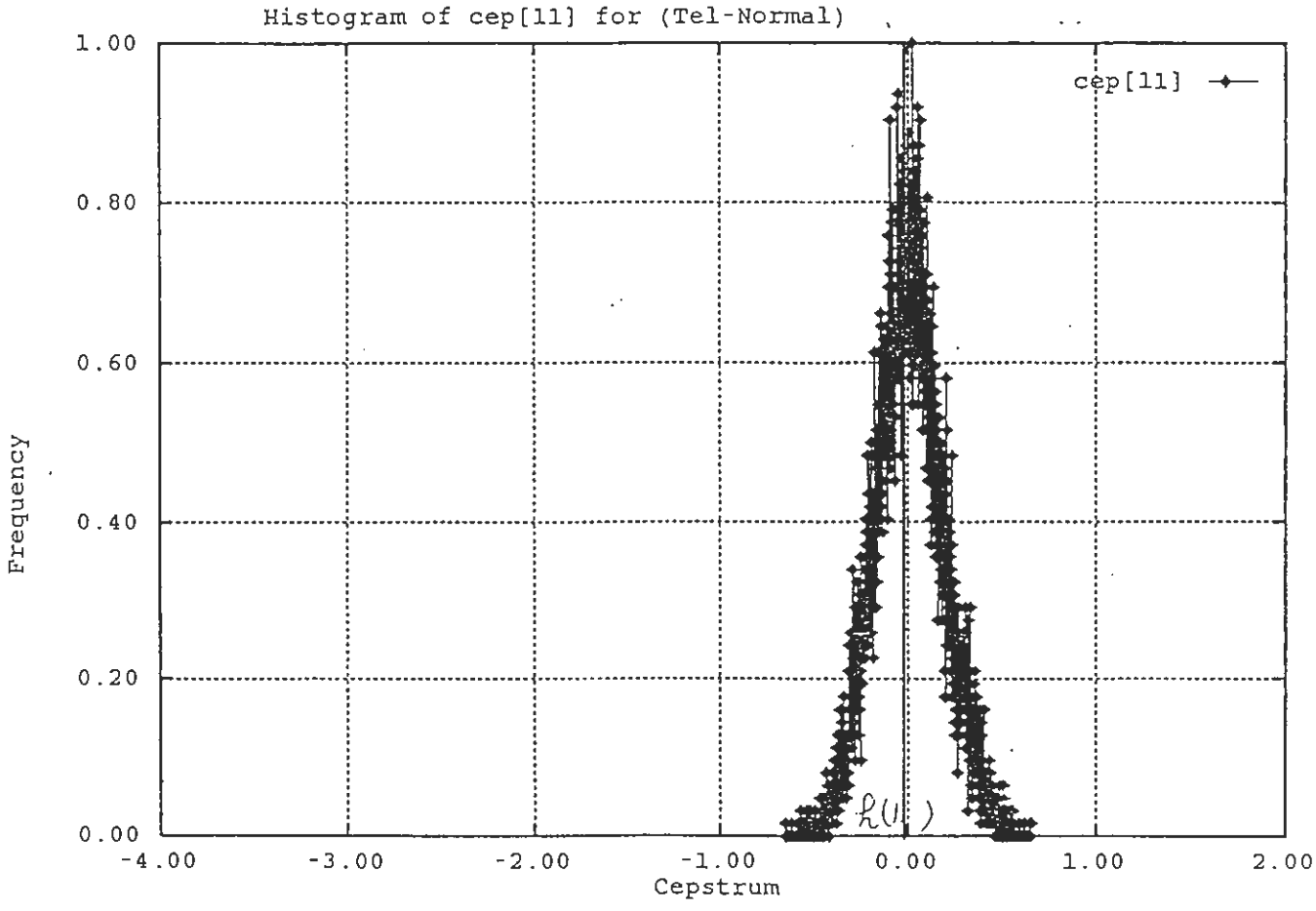

Histogram of cep[6] for (Tel-Normal)
cep[6]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum
R(6)

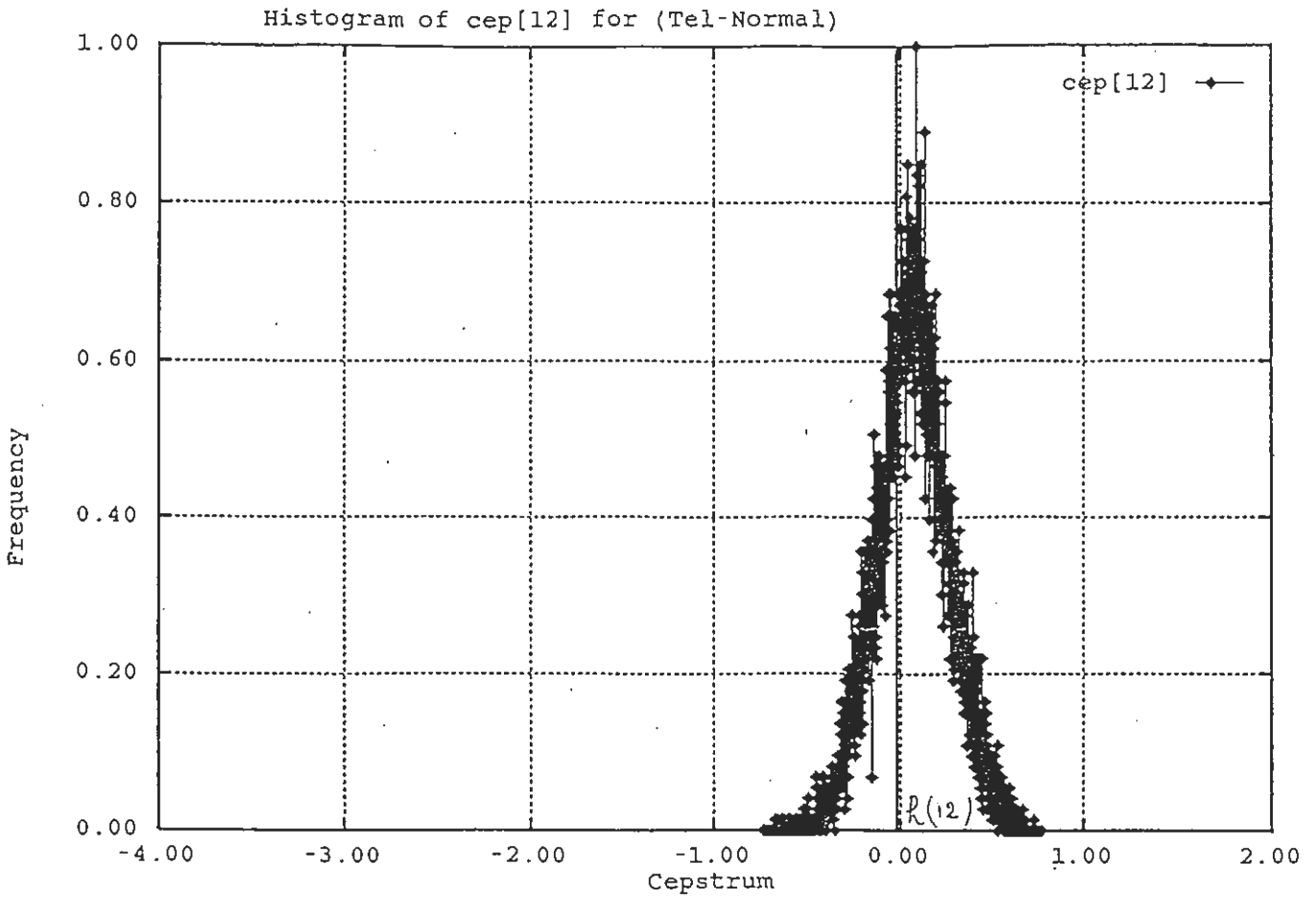

Histogram of cep[7] for (Tel-Normal)
cep[7]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum
R(7)

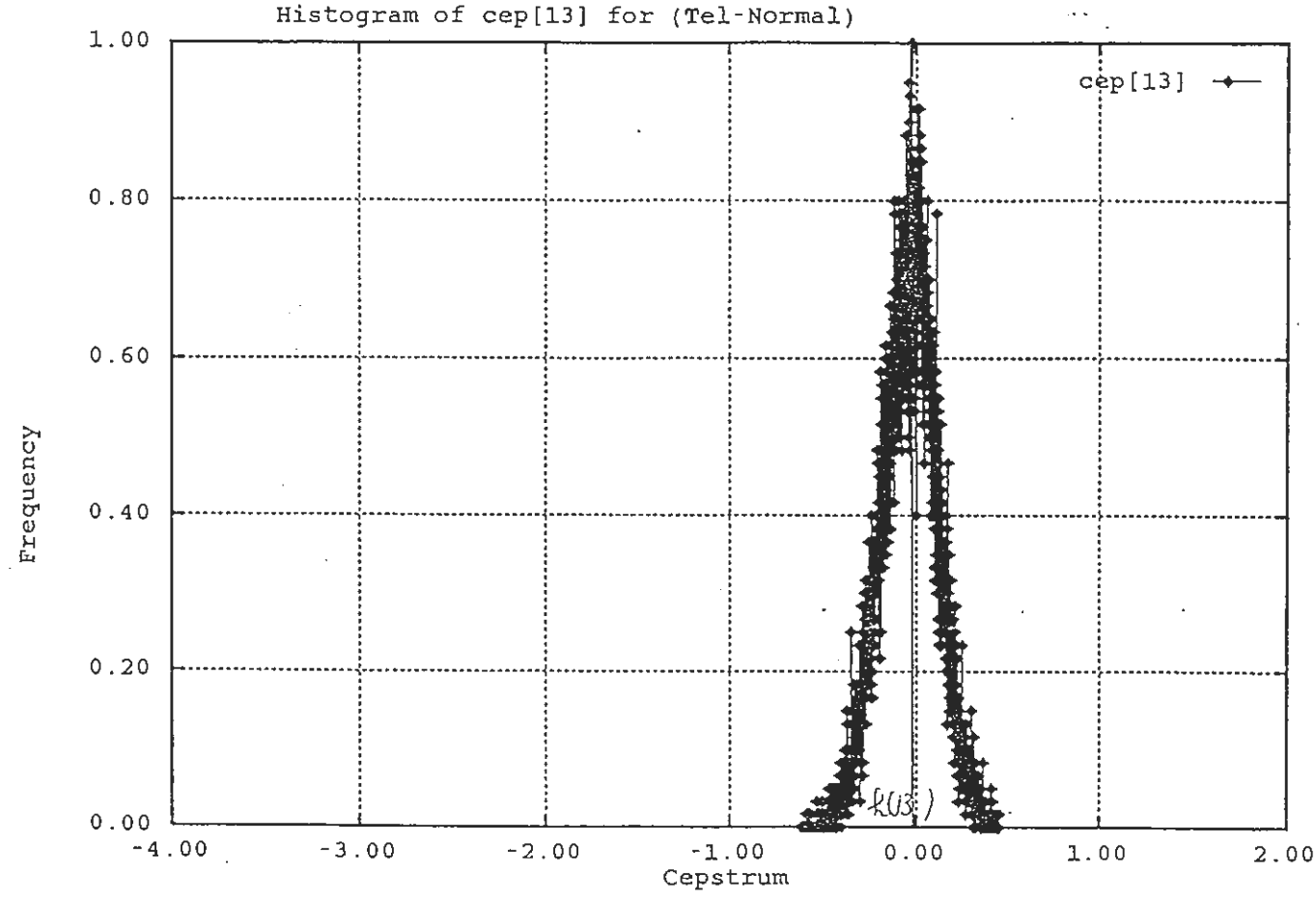

Histogram of cep[8] for (Tel-Normal)
cep[8]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum
R(8)

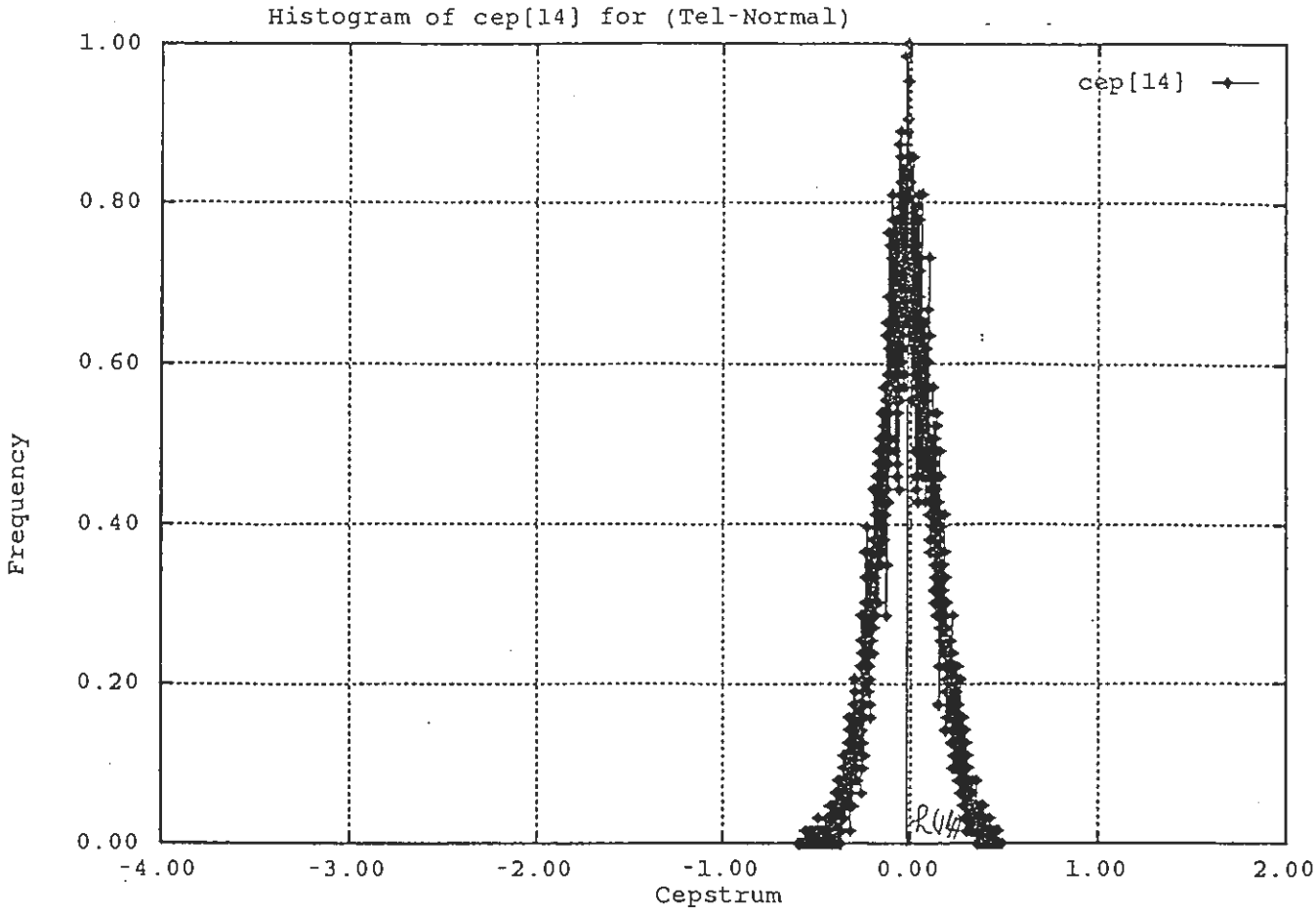

Histogram of cep[9] for (Tel-Normal)
cep[9]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum
R(9)

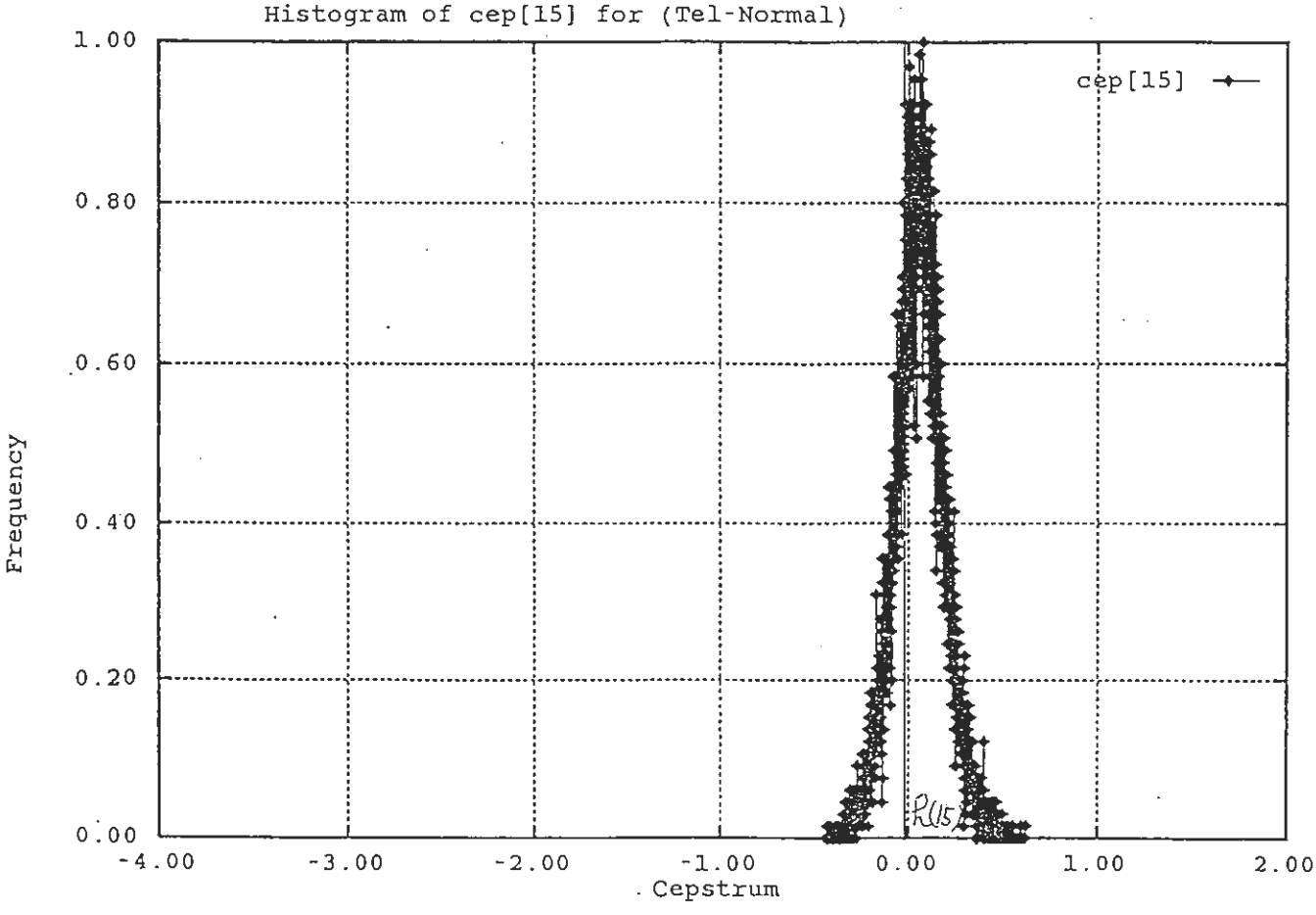

Histogram of cep[10] for (Tel-Normal)
cep[10]
Frequency
1.00
0.80
0.60
0.40
0.20
0.00
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum

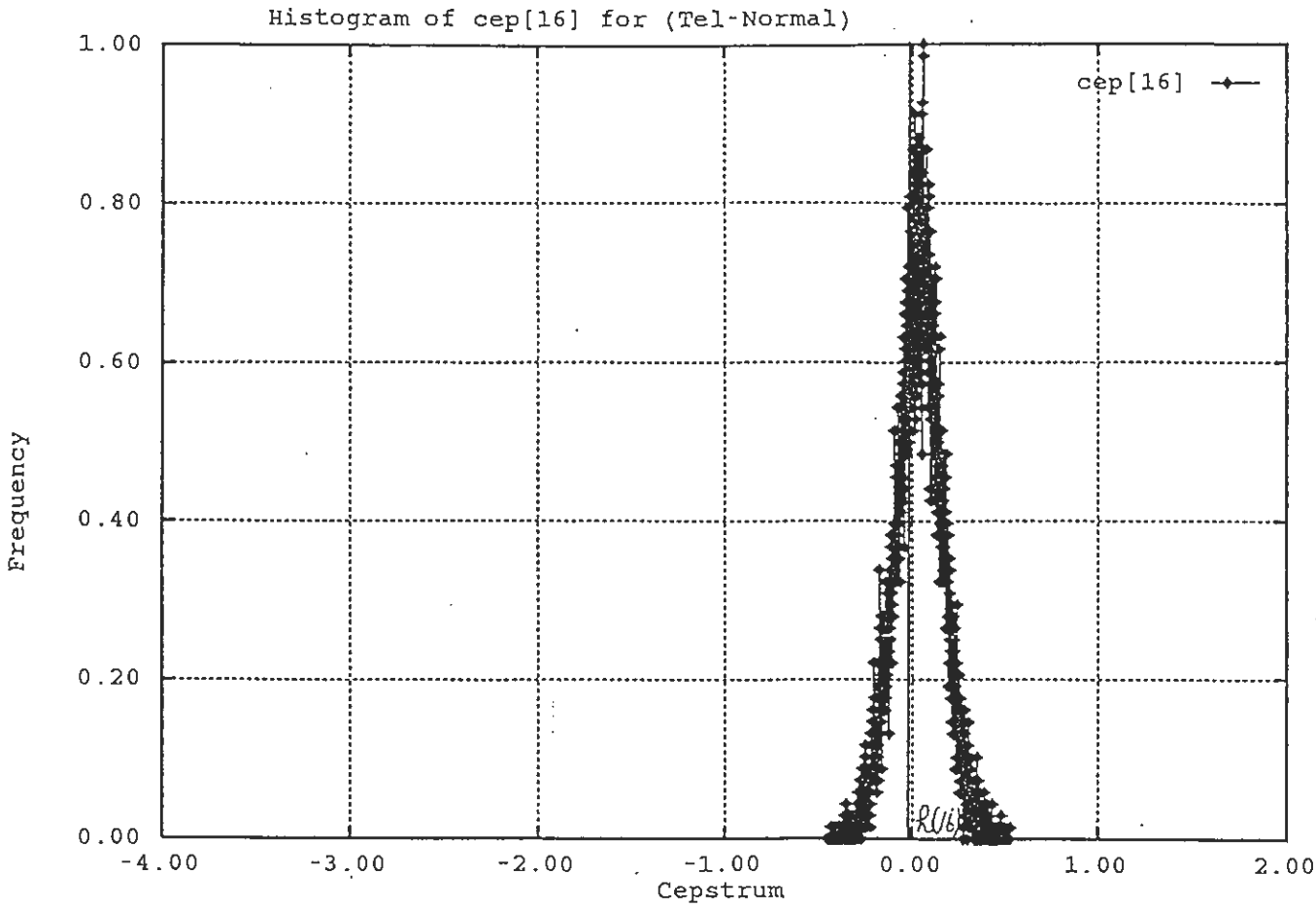

Histogram of cep[11] for (Tel-Normal)
cep[11]
Frequency
1.00
0.80
0.60
0.40
0.20
0.00
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum


Histogram of cep[12] for (Tel-Normal)
cep[12]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum


Histogram of cep[13] for (Tel-Normal)
cep[13]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum


Histogram of cep[14] for (Tel-Normal)
cep[14]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum


Histogram of cep[15] for (Tel-Normal)
cep[15]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum


Histogram of cep[16] for (Tel-Normal)
cep[16]
1.00
0.80
0.60
0.40
0.20
0.00
Frequency
-4.00
-3.00
-2.00
-1.00
0.00
1.00
2.00
Cepstrum

C　LPC 分析における相関係数の計算には要注意

音声情報処理研究室の人ならおそらく誰もが使用しているはずの 34 次元データファイル。これはもともと嵯峨山室長が作成したデータファイルであるが、その分析過程において一つ注意しなければならないことがある。それは、相関係数を計算する時の演算精度である。

嵯峨山室長が作成した数多くのプログラムの内で相関係数を計算する /NFS/atr-fs/pub1/common/src/saga/analys/mlabcorr.c がある。その中に、サブルーチンモジュールとして sigcor0 がある。そのモジュールの中で、相関係数を求めているパラメータ “sqsum” と “c” が float 型で宣言されている。結論としては、これらの変数は double 型で宣言しておく必要がある。そうでないと、実際の 34 次元データファイルに非常に膨大な値が入っていたりすることがある。また、その異常な値の入った箇所について調べたところ、無音以外の音素でもそれが生じていることが分かった。

今回のことは、たまたま電話音声波形ファイルを LPC 分析することによって 34 次元データファイルを作成し、次元毎の最大値と最小値を求めた時に見つけたことである。これまで嵯峨山室長が作成したデータファイルに関しては、今回の演算精度の面では全く問題がなかったようである。