正規分布の尤度補正による突発性雑音に頑健な音声認識*

◎山本仁　篠田浩一　嵯峨山茂樹　（東大・情報理工）

1　はじめに

近年の音声認識技術の進歩により，静かな環境においては精度の高い認識が可能になった．しかし，実環境では，雑音の存在や伝達特性などの影響で認識性能が劣化するという問題があり，中でも未知の非定常雑音への対処は重要な課題のひとつである [1]．

本稿では，短時間区間で加法性の非定常雑音である突発性雑音を対象とする．従来，事前に学習した雑音モデルを用いる方法 [2][3] が提案されているが，モデルの学習データに含まれない雑音に対しては対処が困難である．また，SS 法などの雑音除去法も非定常雑音への対処は難しい．本稿では，事前知識を用いない頑健な尤度計算法として正規分布の尤度補正 [4] を提案する．これは missing feature theory[5] の実現法のひとつである acoustic backing-off（AB）[6] を拡張した手法となっている．今回は，雑音重畳音声の大語彙連続音声認識実験により AB 法と比較を行い，効果を確認した．

2　正規分布の尤度補正

2.1　突発性雑音による認識誤り発生の機構

音響特徴の HMM の出力確率分布には正規分布が広く用いられているが、その対数尤度は 2 次関数となり分布平均からのずれの 2 乗に従い低下する．このことが突発性雑音による認識誤りの一因になると考えられる．図 1 左で，$/a/$ と $/o/$ の 2 つのモデルを考える．$/o/$ のサンプル y_s が観測されたとき，尤度は $/o/$ の方が大きく，$/a/$ よりも $/o/$ からの出力と考えられる．しかし，y_s が重畳雑音等の影響で y_o に変化したとき，尤度値はどちらも小さくなるが，対数尤度では図の通り $/a/$ と $/o/$ で大小が逆転するだけでなく，2 次曲線によってその間に大きな差を生じる．このように，尤度値自体は小さいにもかかわらず対数尤度に大きな差を許容することが，認識誤りの一因になると考えられる．

2.2　提案法：尤度の補正

以上の考察に基づき，尤度値が小さい場合には対数尤度に大きな差を与えない方法を考える．観測されたデータの分布 $N(\boldsymbol{y})$ に正の微小な定数 ϵ（補正定数と呼ぶ）を加えたものを $N'(\boldsymbol{y})$ とすると，その対数は，次式のように分布中心部では $\log N(\boldsymbol{y})$ に，分布裾部では $\log\epsilon$ に近づく（図 1 右）．

$$\log N'(\boldsymbol{y}) \;=\; \log\{N(\boldsymbol{y})+\epsilon\} \tag{1}$$

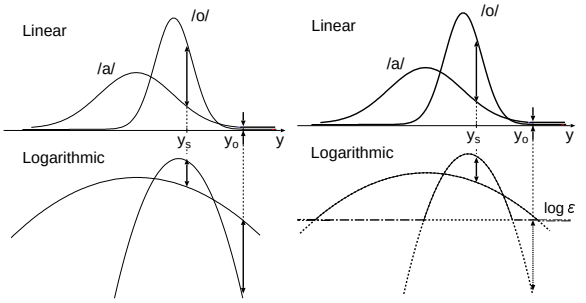

図 1: 正規分布の尤度補正の効果

$$\simeq \begin{cases} \log\epsilon & \text{if } |\boldsymbol{y}-\boldsymbol{\mu}| \gg 0 \\ \log N(\boldsymbol{y}) & \text{else} \end{cases} \tag{2}$$

この補正は多次元分布にも適用可能である．提案法により，雑音の事前知識や雑音部分の判定を用いずに分布から外れた入力に対する一様な尤度の付与が可能となり，認識性能の向上が期待できる．

なお，上記の $N'(y)$ は，混合分布に分岐確率を ϵ で分散 ∞ の分布を加えたものとの理解も可能である．

2.3　AB 法との比較

提案法に近い手法である acoustic backing-off（AB）[6][7] は，特徴次元ごとに，観測されたデータの分布 $p(y)$ と観測されないデータの分布 $p_0(y)$（矩形分布とする）の混合分布から尤度を求めるものである．λ をバックオフ係数，$R_{\rm M}$ と $R_{\rm m}$ をそれぞれ特徴量の学習データ内での最大値と最小値とし，次式で表される．

$$-\log p_{AB}(y) = -\log\{(1-\lambda)p(y)+\lambda p_0(y)\} \tag{3}$$

$$p_0(y) = \begin{cases} (R_{\rm M}-R_{\rm m})^{-1} & \text{if } R_{\rm m} \le y \le R_{\rm M} \\ 0 & \text{else} \end{cases} \tag{4}$$

AB 法は次元ごとにバックオフを行うため，スペクトルの変形が次元全般に分散する場合に効果が低いことが報告されている．また，変形が矩形領域を越える場合は $p_0(y)=0$ となり従来法にほぼ等価である．

提案法は多次元分布や混合分布にも適用できるため，AB 法のように処理を 1 次元ごとに行う必要はない．また，AB 法では学習データの範囲を予め調べる必要があるが，提案法ではその必要はない[1]．

3　認識実験による評価

3.1　実験条件

評価実験として大語彙連続音声認識実験を行った．認識システムには IPA の日本語ディクテーション基本ソフトウェア [8] を用いた．

---

*“Robust Speech Recognition based on the Compensation of Acoustic Likelihood from Gaussian distribution” by Hitoshi YAMAMOTO, Koichi SHINODA and Shigeki SAGAYAMA (Graduate School of Information Science and Technology, The University of Tokyo).

[1] $N'(y)$ はそのままでは確率分布と言えないが，積分範囲を事後的に音声と雑音重畳音声の覆う空間のみとすれば，その積分値は $N(y)$ のものとほぼ同一に見なせる．よって存在範囲が未知の雑音重畳音声が出現した場合にその尤度を $N'(y)$ から求めることに問題はないと考えられる．

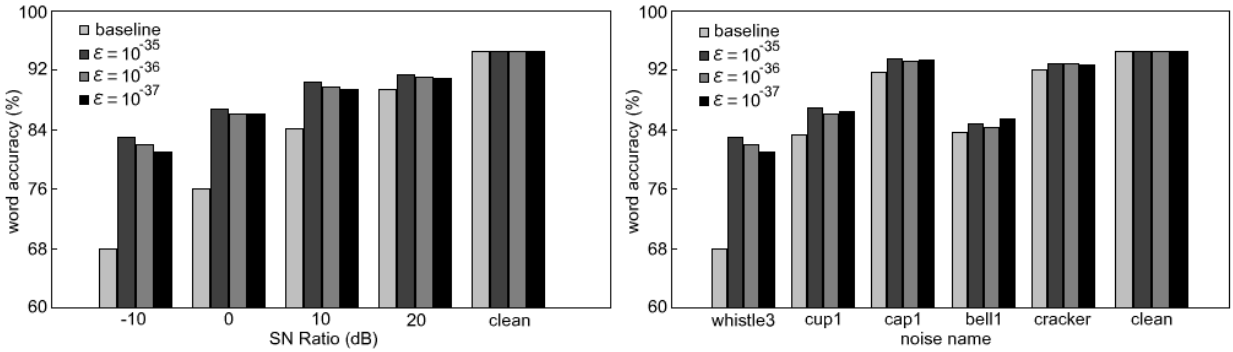

図 3: 左：雑音 whistle の SN 比を変化させたとき，右：雑音の種類を変化させたとき（WA, %）

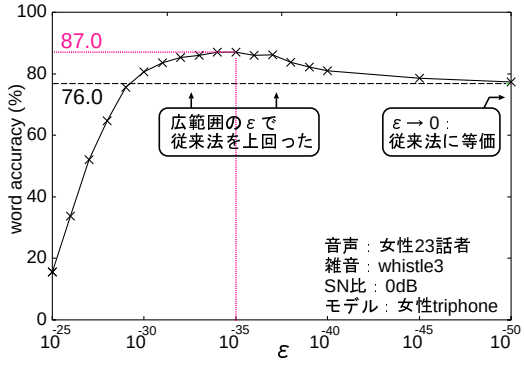

図 2: 提案法で補正定数を変化させたとき（WA, %）

評価データは計算機上で音声と雑音の波形を加算して作成した．音声データは IPA-98-TestSet から男女それぞれ 23 名の合計 200 文の新聞朗読音声を用いた．雑音データは RWCP 実環境音響音声データベース [9] の環境音のうち突発性雑音と考えられるものを 5 種類それぞれ数十サンプル用いた．音声と雑音の加算は次のように行った．まず，各サンプルから雑音を含む区間を 1 秒切り出し，それぞれに 1 秒の無音を挟んでつないだ．次に，この長い雑音データから各音声データと同時間分切り出し，雑音が音声部分に重なるか否かは考慮せずに加算した．

特徴量ベクトルは，12 次 MFCC と ΔMFCC，Δ 対数パワーの計 25 次元とした．音響モデルは，IPA の性別依存モデルから monophone と 2000 状態の triphone を用いた．混合数は 16 とした．言語モデルは IPA のモデルのうち 75month cutoff-1-1 を用いた．デコーダは Julius 3.1p2 の高速版を用い，音響尤度計算部に提案法および AB 法を実装した．

3.2 実験結果

前節の条件のもとで行った実験結果を示す．評価基準には各話者の単語認識精度（Word Accuracy:WA）の平均を用いる．提案法と AB 法の結果については，それぞれパラメタ（ϵ, R, λ）を数通り変化させて認識率が最も高かったものを示す．

提案法では幅広い ϵ で従来法（補正なし）を上回る認識性能が得られ，効果が最も大きいもので従来法の 76.0%に対して 87.0%の認識率を示した（図 2）．この効果はモデルの種類や話者，また SN 比や雑音の種類によらず確認された（図 3）．また，提案法で

表 1: 提案法と AB 法の認識率の比較（WA, %）

方法	適用時	パラメタ	tri phone	mono phone
従来法	—	—	76.0	70.0
AB 法	1 次元	$R = \pm 3.29\sigma$ $\lambda = 0.01$	75.4	70.2
提案法	1 次元	$\epsilon = 10^{-7.2}$	76.1	70.0
提案法	1 次元 多次元	$\epsilon = 10^{-7.2}$ $\epsilon = 10^{-35}$	87.1	75.6
提案法	多次元	$\epsilon = 10^{-35}$	87.0	75.9

は AB 法に比べて認識性能が高く，多次元での補正の効果が確認された（表 1）．

4 おわりに

実環境における音声認識の雑音への頑健性を高める手法として，正規分布から求まる音響尤度の補正を提案し，雑音重畳音声の大語彙連続音声認識によってその効果を確認した．また，AB 法との比較を行ない，提案法の優位性を確認した．

参考文献

[1] 中村哲, “実音響環境に頑健な音声認識を目指して,” 信学技報, SP 2002-12, 2002.

[2] 滝口哲也, 西村雅史, “フレーム単位でのモデル選択による突発性雑音下での音声認識,” 日本音響学会 2002 年春季講演論文集, pp.57-58, 2002.

[3] 伊田政樹, 中村哲, “雑音 DB を用いたモデル適応化 HMM の SN 比別マルチパスモデルによる雑音下音声認識,” 信学技報, SP 2001-92, 2001.

[4] 山本仁, 篠田浩一, 嵯峨山茂樹 “ガウス分布の補正による突発性雑音に頑健な音声認識,” 信学技報, SP 2002-45, 2002.

[5] R. P. Lippmann, B. A. Carlson, “Using missing feature theory to actively select features for robust speech recognition with interrupitations, filtering, and noise,” Proc. Eurospeech97, pp. 37-40, 1997.

[6] J. de Veth, B. Cranen, L. Boves, “Acoustic backing-off as an implementation of missing feature theory,” Speech Communication, Vol. 34, pp. 247-265, 2001.

[7] J. de Veth, B. Cranen, F. de Wet, L. Boves, “Acoustic pre-processing for optimal effectivity of missing feature theory,” Proc. Eurospeech99, pp. 65-68, 1999.

[8] 河原達也, 李晃伸, 小林哲則, 武田一哉, 峯松信明, 嵯峨山茂樹, 伊藤克亘, 伊藤彰則, 山本幹雄, 山田篤, 宇津呂武仁, 鹿野清宏, “日本語ディクテーション基本ソフトウェア (99 年度版) の性能評価,” 情処研報, SLP 31-2, 2000.

[9] S. Nakamura, K. Hiyane, F. Asano, T. Nishimura, T. Yamada, “Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition,” Proc. ICLRE, pp. 965-968, 2000.