モデル適応と音響尤度補正を併用した雑音に頑健な音声認識*

◎山本仁　西本卓也　篠田浩一　嵯峨山茂樹　（東大・情報理工）

1　はじめに

実環境での音声認識において，周囲雑音の存在は認識性能の劣化をもたらす [1]．加法性雑音の場合，周囲雑音を事前に学習し，モデルを適応化することで認識性能を向上させることができる．しかし，事前に学習しきれない雑音成分や，事前に学習されない実環境の未知雑音の存在には，従来のモデル適応手法のみでは対処が困難である．一方，非定常で未知という性質をもつ突発的な雑音への頑健性を高める方法として，音響尤度の補正法が提案され [4, 5]，クリーン音声に突発的な雑音が重畳した状況では効果が確認されている．

本稿では，モデル適応と音響尤度補正を併用した音声認識手法を提案する．モデル適応手法によって周囲雑音に対処し，音響尤度補正によって突発的な雑音に対処する．雑音重畳音声の大語彙連続音声認識実験による評価でその効果を確認した．

2　モデル適応

モデル適応手法として，今回はヤコビ適応法 [2] を考える．ヤコビ適応法は，適応処理が高速で，周囲雑音だけでなく，伝達特性や話者性（声道長）も含めた同時適応も可能 [3] という特徴を持つ．本稿では雑音のみを対象とし，図 1 に示す枠組みを構築する．この枠組みにより，時々刻々と変化する周囲雑音に適応した，頑健性の高い音声認識が可能になると考えられる．

ヤコビ適応法の原理を示す．事前に仮定した環境の音声を $\boldsymbol{Y}^c_{ref}$ 雑音を $\boldsymbol{N}^c_{ref}$，発話時の環境の音声を $\boldsymbol{Y}^c_{tar}$ 雑音を $\boldsymbol{N}^c_{tar}$ とする．また，$\boldsymbol{Y}^c$ と $\boldsymbol{N}^c$ はケプストラム領域，$\boldsymbol{Y}^s$ と $\boldsymbol{N}^s$ はスペクトル領域を表すとする．ヤコビ適応法では，雑音の変化にともなうモデルの非線形変換を次式のように線形近似し，ケプストラム領域で適応処理を行なう．

$$\boldsymbol{Y}^c_{tar} \simeq \boldsymbol{Y}^c_{ref} + \boldsymbol{J}_N(\boldsymbol{N}^c_{tar} - \boldsymbol{N}^c_{ref}) \tag{1}$$

$\boldsymbol{J}_N$ はヤコビ行列であり，次式から事前処理として求めておくことができる．

$$\boldsymbol{J}_N = \frac{\partial \boldsymbol{Y}^c}{\partial \log \boldsymbol{Y}^s}\frac{\partial \log \boldsymbol{Y}^s}{\partial \boldsymbol{Y}^s}\frac{\partial \boldsymbol{Y}^s}{\partial \boldsymbol{N}^s}\frac{\partial \boldsymbol{N}^s}{\partial \log \boldsymbol{N}^s}\frac{\partial \log \boldsymbol{N}^s}{\partial \boldsymbol{N}^s} \tag{2}$$

$$= \boldsymbol{C}\frac{\boldsymbol{N}^s}{\boldsymbol{Y}^s}\boldsymbol{C}^{-1} \quad (\boldsymbol{C}：\text{コサイン変換行列}) \tag{3}$$

*“Robust Speech Recognition Using Model Adaptation and Acoustic Likelihood Compensation” by Hitoshi YAMAMOTO, Takuya NISHIMOTO, Koichi SHINODA and Shigeki SAGAYAMA (The University of Tokyo).

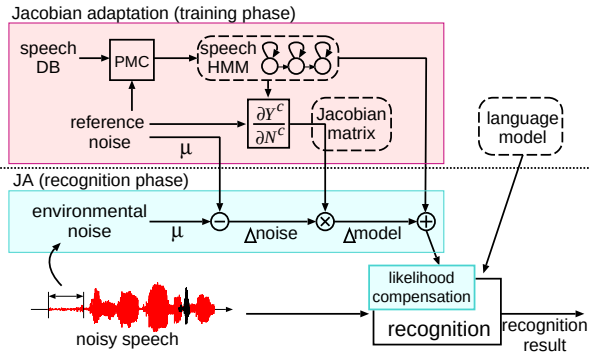

図 1: 提案手法の枠組み

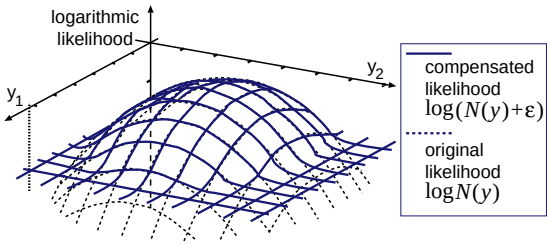

図 2: 音響尤度補正の効果（2 次元の場合）

3　音響尤度補正

音響尤度補正は，音響的にモデルから外れた入力（outlier）の尤度値を底上げして認識計算への影響を抑える方法である [4]．この方法は，acoustic backing-off（AB）[6] を多次元に拡張したものとも解釈できるが，AB 法と比較し，雑音の影響による特徴量の変化が次元全般にわたるような場合でも効果がある点で優れている [5]．

HMM の出力確率分布 $N(\boldsymbol{y})$ に正の微小な定数 ϵ（補正定数）を加えたものを $N'(\boldsymbol{y})$ とすると，その対数 $\log N'(\boldsymbol{y})$ 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 2 参照）．

$$\log N'(\boldsymbol{y}) = \log\{N(\boldsymbol{y}) + \epsilon\} \tag{4}$$

$$\simeq \begin{cases} \log \epsilon & \text{if } |\boldsymbol{y} - \boldsymbol{\mu}| \gg 0 \\ \log N(\boldsymbol{y}) & \text{else} \end{cases} \tag{5}$$

この補正計算により，outlier に対して生じるモデル間の過大な尤度差を抑制できる．認識計算時に補正を行なうことで，雑音の事前知識や判定機構を必要とせずに突発的な雑音の影響を低減することができる．

4　評価実験

4.1　実験条件

評価実験として雑音重畳音声の大語彙連続音声認識実験を行った．認識システムには連続音声認識コンソーシアム（CSRC）のソフトウェア [8] を用いた．音

響特徴量は，13 次の MFCC（0 次含む）と ΔMFCC の計 26 次元とした．クリーン音声の音響モデルには，新聞記事読み上げコーパス（JNAS）の約 4 万文の音声を学習データとし，3000 状態 16 混合の男女別性別依存の triphone モデルを作成した．これと電子協騒音データベースの展示会場雑音から，PMC[7] によって 20dB のモデルを作成し，ヤコビ適応の初期モデルとした．雑音のモデルは 1 状態で 1 混合のガウス分布とした．ヤコビ適応は MFCC の分布平均にのみ施した．タスクを新聞朗読音声認識とし，言語モデルは，CSRC のモデルから語彙サイズ 20k で 111 か月分の新聞記事で学習したものを用いた．デコーダには Julius 3.3p2 の高速版を用い，これに音響尤度補正を実装した．

評価データには，JNAS から選ばれた IPA-98TestSet に計算機上で雑音を重畳したものを用いた．周囲雑音として展示会場雑音を 3 通りの SN 比（15, 20, 25dB）で音声データの全区間に加算した．また，突発性雑音には RWCP 実環境音響音声データベースから whistle3 雑音（笛の音）を用い，音声区間かどうかに関わらず 1 秒間隔に SN 比–15dB で加算した．音声のパワーは音声区間（Viterbi alignment の結果）の平均値，展示会場雑音のパワーは全区間の平均値，whistle3 雑音のパワーは全区間での最大値とした．

4.2 実験結果

前節の条件で行なった実験結果を表 1・表 2 に示す．性別と各 SN 比について，クリーンのモデルの場合（base），ヤコビ適応を行なった場合（JA），さらに音響尤度補正を併用した場合（both），である．認識率は単語正解精度（word accuracy: WA）の男女別の平均とし，各表には認識率の向上が平均して最大であった補正定数 ϵ での結果を示す．まず，表 1 は展示会場雑音のみを重畳した場合である．ヤコビ適応で認識性能が大きく向上したが，尤度補正との併用では向上しなかった．これは展示会場雑音には尤度補正の対象となる outlier が少なかったためと考えられる．次に，表 2 は展示会場雑音に加え，突発的な whistle3 雑音を重畳した場合である．突発的な雑音の有無に関してヤコビ適応の効果の差は小さかったが，尤度補正との併用では認識率が向上した．これより，周囲雑音の存在する環境においても，突発性雑音に対して尤度補正の効果があることが示された．

図 3 に補正定数 ϵ と認識率の関係を示す．補正定数 ϵ は 10^{-33}〜10^{-38} とした．展示会場雑音のみのときに比べ，whistle3 雑音が加わった場合は ϵ の値によって認識率に差が大きかった．雑音条件によって適した ϵ の値が異なると考えられる．

5 おわりに

実環境の雑音に頑健な音声認識手法として，モデル適応と音響尤度補正の併用を試みた．雑音重畳音声の大語彙連続音声認識実験により，認識性能の向上を確認した．なお今回は，学習時・認識時ともに乗法性歪みを考慮しなかった．乗法性歪みを考慮した方法（e.g. [9]）を用いることによって，ベースの認識性能向上と，モデルの分布の分散が小さくなることによる尤度補正の効果の増大とが期待される．また今後，他の雑音条件への適用や，雑音条件の変動に頑健な尤度補正法の検討も必要である．

表 1: 雑音: 展示会場（$\epsilon=10^{-37}$, WA（%））

	male			female		
SNR	base	JA	both	base	JA	both
15dB	48.8	77.9	78.4	52.0	81.3	80.3
20dB	70.2	83.7	84.6	67.9	84.8	84.6
25dB	80.3	86.3	85.1	78.9	85.2	86.0

表 2: 雑音: 展示会場+whistle3（$\epsilon=10^{-35}$, WA（%））

	male			female		
SNR	base	JA	both	base	JA	both
15dB	33.3	59.5	66.3	37.6	61.6	66.7
20dB	55.0	68.2	73.2	54.3	67.6	75.0
25dB	65.5	70.3	75.1	64.7	72.0	76.8

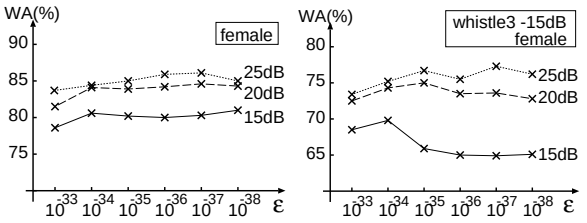

図 3: 補正定数 ϵ と認識率の関係

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] Shigeki Sagayama, Yoshikazu Yamaguchi, Satoshi Takahashi, and Jun-ichi Takahashi, “Jacobian approach to fast acoustic model adaptation,” Proc. ICASSP97, pp. 835-838, 1997.

[3] Hiroshi Shimodaira, Nobuyoshi Sakai, Mitsuru Nakai, and Shigeki Sagayama, “Jacobian joint adaptation to noise, channel and vocal tract length,” Proc. ICASSP2002, pp. 197-200, 2002.

[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] 山本仁, 篠田浩一, 嵯峨山茂樹, “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,” 音講論, 1-9-10, Sep 2002.

[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] M. J. F. Gales, S. Young, “An improved approach to the hidden Markov model decomposition of speech and noise,” Proc. ICASSP92, pp. 233-236, 1992.

[8] 河原達也他, “連続音声認識コンソーシアム 2001 年度版ソフトウェアの概要,” 情処研報, SLP 43-3, 2002.

[9] 庄境誠, 中村哲, 鹿野清宏, “ケプストラム平均正規化と HMM 合成法に基づくモデル適応化法 E-CMN/PMC と自動車内音声認識への適用,” 信学論 D-II, vol. J-80-D-II, No. 10, pp. 2636-2644, 1997.