コンディション変数の導入によるミスマッチがない場合にも頑健なステレオベース特徴量強調*

◎鈴木雅之, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

雑音混入による音声認識精度の低下を防ぐため，様々な特徴量強調法が提案されている．これらにより，雑音が混入した音声の認識精度を向上させることができるが，雑音が混入していない条件では，逆に精度を低下させてしまう場合がある．そのため，雑音が混入していない場合に性能が劣化しないような特徴量強調手法が望まれる．

本稿では，入力特徴量と音響モデルにミスマッチがあるかないかを表すコンディション変数を導入し，コンディションに依存させて特徴量強調を行うことで，ミスマッチがない条件にも頑健な手法を提案する．提案手法をステレオベース特徴量強調に導入した結果，ミスマッチのあるなしに関わらず，高い精度が実現出来ることが分かった．

2 提案手法

ある時刻フレームにおいて観測した特徴量を $\boldsymbol{y}$，それに対応するクリーン音声特徴量を $\boldsymbol{x}$ とおく．特徴量強調の目的は，$\boldsymbol{y}$ から，クリーン音声特徴量の推定値 $\hat{\boldsymbol{x}}$ を以下のように得ることである．

$$\hat{\boldsymbol{x}} = \mathbb{E}_{\boldsymbol{x}}[p(\boldsymbol{x}|\boldsymbol{y})] \tag{1}$$

ただし，$\mathbb{E}_{\boldsymbol{x}}[\cdot]$ は，$\cdot$ の $\boldsymbol{x}$ に関する期待値演算を表す．

提案手法では，ミスマッチがある場合 1，ない場合に 0 となるコンディション変数 c を導入し，(1) を以下のように展開する．

$$\begin{aligned}\hat{\boldsymbol{x}} =& p(c=1|\boldsymbol{y})\mathbb{E}_{\boldsymbol{x}}\left[p(\boldsymbol{x}|c=1,\boldsymbol{y})\right] \\ &+ (1-p(c=1|\boldsymbol{y}))\mathbb{E}_{\boldsymbol{x}}\left[p(\boldsymbol{x}|c=0,\boldsymbol{y})\right]\end{aligned} \tag{2}$$

ここで，$\mathbb{E}_{\boldsymbol{x}}\left[p(\boldsymbol{x}|c=0,\boldsymbol{y})\right]$ は，ミスマッチがない場合なので，$\boldsymbol{y}$ とおくのが自然である．一方，$\mathbb{E}_{\boldsymbol{x}}\left[p(\boldsymbol{x}|c=1,\boldsymbol{y})\right]$ は，ミスマッチがある場合なので，通常の特徴量強調結果を代入すればよい．特徴量強調手法には任意の手法を採用してよい．

$p(c=1|\boldsymbol{y})$ の計算のために，生成モデルを置く．クリーン音声特徴量を θ_0 をパラメタに持つ生成モデル，雑音を重畳した音声特徴量を θ_1 をパラメタに持つ生成モデルでモデル化する．本稿では，具体的に，特徴量が時間によらず GMM に従って生成されるとモデル化する．θ_0 と θ_1 は，予め学習データを用いて学習しておく．

T フレーム分の一発話 $\boldsymbol{y}_{1:T}$ が与えられたとき，その発話のコンディション変数が $c=1$ となる事後確率は，事前確率を $p(c=0)=p(c=1)=0.5$ と仮定すると，標準シグモイド関数 $\sigma(x)=(1+\exp(-x))^{-1}$ を用いて以下のように計算できる．

$$p(c=1|\boldsymbol{y}_{1:T}) = \sigma\left(\sum_{t=1}^{T}\left(\ln p(\boldsymbol{y}_t;\theta_1) - \ln p(\boldsymbol{y}_t;\theta_0)\right)\right) \tag{3}$$

ただし，$\boldsymbol{y}_t$ は時刻 t に観測した特徴量，$\ln p(\boldsymbol{y}_t;\theta_0)$ および $\ln p(\boldsymbol{y}_t;\theta_1)$ は，それぞれ θ_0，θ_1 に対応する GMM の $\boldsymbol{y}_t$ の対数尤度である．

本稿では，$p(c=1|\boldsymbol{y}_{1:T})$ が 0.5 より大きいか小さいか（シグモイド関数の中身の正負）によって，発話単位でコンディション変数を 1 か 0 に一意に定める．発話単位でコンディションを定めるメリットに，バックエンドを切り替えられる点がある．一般的に，クリーン音声の認識はクリーン音声で学習した音響モデルが最も精度が高く，雑音重畳音声の認識は雑音重畳音声に特徴量強調をかけた音声を用いて Noise Adaptive Training（NAT）した音響モデルの精度が高くなる．そこで，発話単位のコンディションが 0 の場合にはクリーン音声で学習した音響モデルを用い，発話単位のコンディションが 1 の場合には NAT で学習した音響モデルを用いることにする．これにより，すべての場合で高い精度が得られると考えられる．

3 DNA-CD

ミスマッチがない条件にも頑健な特徴量強調手法として，Dynamic Noise Adaptation with Condition Detection (DNA-CD) が提案されている [1]．DNA は，グラフィカルモデルを用いた特徴量強調手法である [2]．DNA-CD では，DNA 用のモデル θ_{DNA} と，そこから雑音モデルのみを削除した Null Mismatch モデル θ_{NM} の二つを用意する．そして時刻 t までに観測された特徴量 $\boldsymbol{y}_{1:t}$ を用い，時刻 t のコンディション変数 c_t の事後確率を以下のように計算する．

$$\begin{aligned}p(c_t=1|\boldsymbol{y}_{1:t}) =& \gamma\sigma\left(\ln p(\boldsymbol{y}_t;\theta_{\mathrm{DNA}}) - \ln p(\boldsymbol{y}_t;\theta_{\mathrm{NM}})\right) \\ &+ (1-\gamma)p(c_{t-1}=1|\boldsymbol{y}_{1:t-1})\end{aligned} \tag{4}$$

---

*Matched-condition robust stereo-based feature enhancement with condition detection by M.Suzuki, N.Minematsu, K.Hirose (The University of Tokyo), {suzuki, mine, hirose}@gavo.t.u-tokyo.ac.jp

ただし γ は 0 から 1 の値をとる定数で，事前に決定される．以上で求めたコンディション変数の事後確率を用い，以下のように特徴量強調を行う．

$$\hat{\boldsymbol{x}}_t = p(c_t = 1|\boldsymbol{y}_{1:t})\mathbb{E}_{\boldsymbol{x}_t}[p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t};\theta_{\mathrm{DNA}})] + (1 - p(c_t = 1|\boldsymbol{y}_{1:t}))\mathbb{E}_{\boldsymbol{x}_t}[p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t};\theta_{\mathrm{NM}})] \quad (5)$$

$\mathbb{E}_{\boldsymbol{x}_t}[p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t};\theta_{\mathrm{DNA}})]$ および $\mathbb{E}_{\boldsymbol{x}_t}[p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t};\theta_{\mathrm{NM}})]$ は，DNA のアルゴリズムを用いて推定される．

DNA-CD と提案手法の違いとして，(4) を用いてコンディション変数の事後確率を時間フレーム毎に求めていることがある．提案手法でも (4) のようなやり方を導入することが可能であるが，予備実験の結果，発話単位でコンディションを定めた場合の方が精度が高くなったため，今回は採用しなかった．

DNA-CD と比較した本稿の提案手法の新規性は，コンディション変数推定のためにモデルを別に用意することで，ステレオベース特徴量強調など任意の特徴量強調法においてコンディション変数の導入を可能にしたことにある．加えて，発話単位でコンディションを一意に定めるアプローチを採用することで，バックエンドの音響モデルを変更可能にしたことにも新規性がある．

4 実験

データベースとして AURORA2 を利用する [3]．AURORA2 は，雑音環境下における連続数字音声認識の評価を行うデータベースで，A セットが雑音環境クローズドテスト，B セットが雑音環境オープンテスト，C セットがチャネルノイズありのテストとなっている．

認識実験は以下の三つの条件で行った．clean 条件および multi 条件が従来手法，select 条件が提案手法である．

clean 条件 クリーン音声のみを用いて単語単位の HMM/GMM を最尤推定する．特徴量には PLP + Δ + ΔΔ を CMLLR し CMN したものを利用する．

multi 条件 SNR 5,10,15,20,∞ [dB] で雑音が付加された音声を用いて単語単位 HMM/GMM を NAT する．特徴量には，PLP + Δ + ΔΔ を [4] に倣って CMLLR し，[5] のステレオベース特徴量強調を行い，HEQ [6] で特徴量正規化をしたものを利用する．ステレオベース特徴量強調の学習データには，HMM/GMM の学習データをステレオデータにしたものを用いる．

select 条件 コンディション変数の推定結果に従い，clean 条件と multi 条件を選択して認識を行う．

Table 1　AURORA 2 における word accuracy の平均（%）．クリーンは clean1-4 の平均，A，B，C はそれぞれのテストセットにおける SNR 0-20 の平均を表す．

	クリーン	A	B	C
clean 条件	**99.67**	76.74	79.03	80.45
multi 条件	99.43	**94.34**	**93.55**	**93.75**
select 条件	**99.67**	**94.34**	93.54	**93.75**

コンディション変数の推定に用いる生成モデルには，32 混合の GMM を用い，特徴量には PLP + Δ + ΔΔ を CMN したものを利用した．θ_1 の学習には SNR 5,10,15,20 [dB] で雑音が付加された音声を，θ_0 の学習にはそれと同じ音声データで雑音が含まれていないものを利用した．

実験結果を Table 1 に示す．clean 条件はクリーン音声を認識する上では性能が高いが，雑音のミスマッチがあると精度が低くなってしまう．multi 条件は A，B，C セットを認識する上では性能が高いが，雑音のミスマッチがない条件では精度が低くなってしまう．提案手法である select 条件を用いると，ミスマッチのあるなしに関わらず，ほぼ最高性能を実現することができる．雑音環境オープンテストである B セットでも，multi 条件と select 条件の差はほぼないため，コンディション変数の推定は雑音環境に依らず適切に行えていることが分かる．

5 まとめ

ミスマッチがない条件で性能が低下しない特徴量強調を実現するために，発話単位でコンディション変数を推定し，バックエンドの音響モデルを切り替えながら音声認識を行う手法を提案した．提案手法をステレオベース特徴量強調に導入した結果，ミスマッチのあるなしに関わらず，高い認識精度が得られることが分かった．今後の課題は，DNA-CD との比較と，他の大規模なデータベースでの評価である．

参考文献

[1] S.J. Rennie, *et.al.*, Proc. ASRU, pp.137–140 (2011)

[2] S.J. Rennie, PhD thesis (2008)

[3] H.G. Hirsch and D. Pearce, Proc. ISCA ITRW ASR (2000)

[4] Y. Shinohara, *et.al.*, Proc. ICASSP, pp.4881–4884 (2008)

[5] 鈴木雅之 他，音講論（春） (2012)

[6] A. de la Torre, *et.al.*, IEEE TSAP, vol. 13, pp.355–366 (2005)