クリーン音声状態の識別に基づく特徴量強調*

◎鈴木雅之 (東大), 吉岡拓也, 渡部晋治 (NTT CS 研), 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音声認識の性能は背景雑音により大幅に低下することが知られており，雑音に頑健な音声認識は，多くの研究者の興味を集める研究分野である．特徴量強調は，雑音に頑健な音声認識を実現するための有力な手段の一つであり，様々な研究が進められている [1].

近年成功している特徴量強調の多くは，Gaussian mixture model (GMM) でモデル化された劣化音声の分布を用いて特徴量強調を行う．具体的には，劣化音声が観測されると，GMM インデックスの事後確率を求め，この事後確率に関連づけられた強調処理を行う．例えば vector Taylor series（VTS）を用いた特徴量強調 [2] では，事前に用意したクリーン音声の GMM と推定した雑音パラメータから，VTS 近似を用いることで劣化音声の GMM を合成し，これを利用する．劣化音声の GMM を雑音パラメータを用いて合成することで，GMM を処理対象の雑音環境に適合させることができる．そのため，精度が高くなるが，計算量も高くなってしまう問題がある．一方 SPLICE [3] では，多様な雑音を含んだ学習データを用いて劣化音声に関する GMM を用意して利用する．これにより計算量の問題は解決されるが，GMM の学習データが必ずしも処理対象の雑音環境に適合しないため，GMM による近似が不適切になる場合が多く，精度面で劣る．

これを鑑みて本稿では，劣化音声状態の事後確率の計算を，クリーン音声状態の事後確率の計算と近似し，これを識別的に推定する手法を提案する．これにより，計算量を低く保ったまま，所望の事後確率を高精度で推定することが可能になる．提案手法を AURORA2 データベースで評価した結果，特に，識別問題を解くのに用いる特徴量として，観測劣化音声特徴量と推定雑音特徴量を連結したベクトルをさらに近隣数フレーム分連結したものを用いた場合，精度向上が実現できた．

2 GMM を用いた特徴量強調

本節では，これまで提案された GMM を用いた特徴量強調を紹介する．具体的に，VTS を用いた特徴量強調，SPLICE，及び SPLICE を発展させた手法を紹介する．

2.1 VTS を用いた特徴量強調

クリーン音声特徴量を x，劣化音声特徴量を y，雑音特徴量を n とおく．これらの間の関係を，

$$y = x + g(x, n) \tag{1}$$

と表す．g は，ミスマッチ関数と呼ばれる．特徴量が対数メルフィルタバンク出力（FBANK）ドメインの場合，ミスマッチ関数 g は以下のように近似できる．

$$g(x, n) = \log(1 + \exp(n - x)) \tag{2}$$

MFCC ドメインの場合は，以下のように近似できる．

$$g(x, n) = D \log(1 + \exp(Cn - Cx)) \tag{3}$$

ただし D は DCT 行列，C はその逆行列である．

VTS を用いた特徴強調では，まず以下のように x の確率密度関数を GMM で学習する．

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k^x, \Sigma_k^x) \tag{4}$$

ただし，K は GMM の混合数，$\pi_k, \mu_k^x, \Sigma_k^x$ はそれぞれ k 番目のインデックスに対応する GMM の重み，正規分布の平均，分散である．次に，雑音特徴量 n の確率密度関数を正規分布によりモデル化する．

$$p(n) = \mathcal{N}(n; \mu^n, \Sigma^n) \tag{5}$$

ただし μ^n, Σ^n は推定された正規分布の平均，分散であり，観測データから推定される．次に，x と n の確率密度関数が上記の様に既知となった上で y の確率密度関数を GMM として近似するために，g を x と n に関して一次の VTS で近似する．

$$g(x^0, n^0) \approx A(x^0, n^0)x + B(x^0, n^0)n + c(x^0, n^0) \tag{6}$$

ただし，x^0，n^0 は VTS 近似を行う展開の中心，A，B は g をそれぞれ x，n で一回微分したもの，c は x と n に依存しない定数である．そして (6) を利用して，以下のように y の確率密度関数を GMM で近似する．

$$p(y) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y; \mu_k^y, \Sigma_k^y) \tag{7}$$

$$\mu_k^y = \mu_k^x + g(\mu_k^x, \mu^n) \tag{8}$$

$$\Sigma_k^y = A(\mu_k^x, \mu^n)\Sigma_k^y A(\mu_k^x, \mu^n)^T + B(\mu_k^x, \mu^n)\Sigma^n B(\mu_k^x, \mu^n)^T \tag{9}$$

---

*Feature enhancement based on clean speech state discrimination by M.Suzuki (Tokyo Univ.), T.Yoshioka, S.Watanabe (NTT), N.Minematsu, K.Hirose (Tokyo Univ.), Shinji Watanabe is now with Mitsubishi Electric Research Laboratories (MERL)

VTS を用いた特徴量強調では，このように推定された GMM を使って，以下のようにクリーン音声の推定値 $\hat{x}$ を近似する

$$\hat{x} = \sum_{k=1}^{K} p(k|y)\,(y - g(\mu_k^x, \mu^n)) \tag{10}$$

$$p(k|y) = \frac{\pi_k \mathcal{N}(y; \mu_k^y, \Sigma_k^y)}{\sum_k \pi_k \mathcal{N}(y; \mu_k^y, \Sigma_k^y)}. \tag{11}$$

一般的に GMM のインデックスの事後確率は，ある一つのインデックスがドミナントになることが多いため，(10) はさらに以下のように近似できる

$$\hat{x} = y - g(\mu_{k^*}^x, \mu^n) \tag{12}$$

$$k^* = \operatorname*{argmax}_k p(k|y). \tag{13}$$

VTS を用いた特徴量強調は効果が高いが，計算量が高いという問題がある．具体的には，雑音パラメータが変わるごとに，$\{\Sigma_k^y\}_{k=1\cdots K}$ の逆行列を計算する必要がある．ミスマッチ関数が (2) となる FBANK ドメインの場合，Σ_k^y は対角行列になるため計算量は問題にならないが，(3) となる MFCC ドメインの場合は全角になるため，計算量は非常に大きくなってしまう．そのため実用的には，FBANK ドメインを利用するか，もしくは雑音パラメータの更新を一発声ごと等に限定するかのどちらかが行われる．これにより，MFCC ドメインでフレーム毎に雑音推定値を更新する場合と比べて，精度が低下してしまう．

2.2 SPLICE

SPLICE では，y の確率密度関数を，予め用意した学習データを用いて GMM として学習する．

$$p(y) = \sum_{k=1}^{K} \pi_k \mathcal{N}(y; \mu_k^y, \Sigma_k^y) \tag{14}$$

そしてこの GMM のインデックスの事後確率 $p(k|y)$ を用い，以下のように $\hat{x}$ を求める．

$$\hat{x} = \sum_k p(k|y) A_k y' \tag{15}$$

ただし y' は $y' = [1\ y^T]^T$ なる拡張特徴量ベクトルである．これは VTS を用いた特徴強調と同様，さらに以下のように近似できる．

$$\hat{x} = A_{k^*} y'. \tag{16}$$

$\{A_k\}_{k=1\cdots K}$ は，例えば x と y の時間同期した学習データ $\{x_i, y_i\}_{i=1\cdots I}$ から重み付き最小二乗誤差基準で以下のように学習することができる

$$A_k = \operatorname*{argmin}_{A_k} \sum_{i=1}^{I} p(k|y_i) ||x_i - A_k y_i'||^2. \tag{17}$$

SPLICE では，y の GMM 及び A_k が予め学習されているため，音声強調時の計算コストは非常に低い．しかし，雑音特徴量に関するパラメータをなにも推定せず $p(y)$ を一つの GMM で近似するのは，VTS を用いた特徴強調で用いた (7) の GMM と比べてある程度の誤差が含まれるため，精度が十分でない可能性がある．この問題は，特に未知雑音環境下や非定常雑音環境下で顕著である．

2.3 NMN-SPLICE

SPLICE の問題を低減するヒューリスティックな手法として，noise mean normalization (NMN) が提案されている．NMN-SPLICE では，SPLICE の y の代わりに，そこから n の推定値 $\hat{n}$ を減算した，$y - \hat{n}$ を用い，その GMM を学習して用いる手法である

$$\hat{x} = \sum_k p(k|y - \hat{n}) A_k (y - \hat{n})'. \tag{18}$$

これはさらに以下のように近似できる

$$\hat{x} = A_{k^*}(y - n)' \tag{19}$$

$$k^* = \operatorname*{argmax}_k p(k|y - \hat{n}). \tag{20}$$

y よりも $y - \hat{n}$ の方がより GMM として近似しやすいと考えられるため，NMN-SPLICE は SPLICE の問題を少し緩和することができる．NMN-SPLICE は SPLICE より精度が高いことが知られているものの，なぜ $y - \hat{n}$ を用いると良いのかについては必ずしも明らかでない．

2.4 結合ベクトルと線形判別分析を用いた手法

NMN-SPLICE よりさらに高い精度を実現する手法として，y と $\hat{n}$ の結合ベクトルの適切な部分空間を線形判別分析（LDA）により求める手法を既に提案されている [4]

$$\hat{x} = \sum_k p(k|L[y^T \hat{n}^T]^T) A_k [1\ y^T \hat{n}^T]^T. \tag{21}$$

これはさらに以下のように近似できる

$$\hat{x} = A_{k^*}[1\ y^T \hat{n}^T]^T \tag{22}$$

$$k^* = \operatorname*{argmax}_k p(k|L[y^T \hat{n}^T]^T). \tag{23}$$

ここで L は以下のように求めた LDA を行う行列である．L の学習データには，x と y と $\hat{n}$ 時間同期したデータ $\{x_i, y_i, \hat{n}_i\}_{i=1\cdots I}$ を用いる．まず，$\{x_i\}_{i=1\cdots I}$ からクリーン音声 x の GMM を学習し，インデックスの事後確率 $\{p(k|x_i)\}_{i=1\cdots I}$ を求める．そして $\{[y_i^T \hat{n}_i^T]^T\}_{i=1\cdots I}$ を特徴量，$\{\{p(k|x_i)\}_{k=1\cdots K}\}_{i=1\cdots I}$ をラベルデータとして，LDA を用いて L を推定する．最後に $L[y^T \hat{n}^T]^T$ の GMM を学習すれば，$p(k|L[y^T \hat{n}^T]^T)$ が求められる．

3 提案手法

従来の GMM を用いた特徴量強調は，劣化音声状態を表す GMM のインデックスの事後確率が最も大きくなる k^* を推定し，k^* に対応付けられた変換を行うことで特徴強調を実現していた．本稿では，VTS を用いた特徴量強調で用いたように雑音に依存した y の GMM から得られる k^* を，識別モデルを利用して推定する手法を提案する．これにより，MFCC ドメインの VTS を用いた特徴強調のような全角の共分散行列を持つ GMM の事後確率を計算を避けることができ，計算量が実用的な範囲に収まる．また SPLICE と比較して，$p(y)$ を雑音に依存させずに GMM で近似する誤差がないため，特に未知雑音環境下や非定常雑音環境下における精度の向上が期待される．

3.1 k^* の推定

まず，以下のように x の GMM を学習する．

$$p(x) = \sum_k \pi_k \mathcal{N}(x; \mu_k^x, \Sigma_k^x) \tag{24}$$

$p(x)$ に真の雑音特徴量 n_i が重畳されると，$p(y_i)$ は以下のような混合分布として近似できる．

$$p(y_i) = \sum_k \pi_k f(\mu_k^x, \Sigma_k^x, n_i) \tag{25}$$

ただし，f は未知の確率密度関数である．ここで，この y の混合分布と x の GMM のインデックスは共通としているため，$\underset{k}{\mathrm{argmax}}\, p(k|y_i) \approx \underset{k}{\mathrm{argmax}}\, p(k|x_i)$ という近似がよく成り立つ．ここで，$p(k|x_i)$ は x_i と (24) から計算可能である．提案手法では，$\underset{k}{\mathrm{argmax}}\, p(k|x_i)$ を，$\underset{k}{\mathrm{argmax}}\, p(k|y_i)$ の近似値とし，これを任意の識別モデルを使って推定して k_i^* を求める．具体的には，識別モデルの特徴量を d として，

$$\hat{k}^* = \underset{k}{\mathrm{argmax}}\, p(k|d;\theta) \tag{26}$$

のように識別を行う．ただし θ は識別モデルのパラメータであり，$\{\underset{k}{\mathrm{argmax}}\, p(k|x_i), d_i, \}_{i=1\cdots I}$ から学習される．このようにクリーン音声状態 $p(k|x_i)$ をラベルとして教師あり学習を用いるのは，2.4 節で説明した著者らの従来手法 [4] と同等である．[4] との差分は，SVM などの任意の識別モデルを利用できるよう一般化した点にある．

提案法で用いる k_i^* を推定する識別モデルの特徴量には，任意の短時間特徴量を利用することができる．本稿では，[4] で利用している $[y_i^T \hat{n}_i^T]^T$ に加え，時刻 i の前後数フレーム分の特徴量を連結したものを特徴量として利用することを提案する．これにより，k_i^* を推定するための情報が増え，精度向上が見込める．

k_i^* を求める識別モデルを利用することは，HLDA を用いた特徴量変換 [5] や，TANDEM アプローチ [6] と関連が深いことを指摘しておく．HLDA や TANDEM アプローチでは，強制アライメントによって得られた HMM の状態ラベルを推定するように，HLDA あるいは MLP などを学習している．提案手法で利用しているラベル k_i^* も，GMM によるクリーン音声状態のラベルである．そのため，提案手法は HLDA や TANDEM アプローチと似た手法といえる．

3.2 k^* 推定後の変換

k^* を求めた後，VTS を用いた特徴量強調では (12) のようにミスマッチ関数の減算を，SPLICE 及びその変形では (16) のようにステレオデータから学習した線形変換を利用していた．提案手法では，そのどちらを利用することも可能である．今回は予備実験の結果から，SPLICE のように予め線形変換を学習して利用するものを採用することした．

4 実験

AURORA2 データベース [7] を用いて提案手法の評価を行った．AURORA2 データベースの評価セットのうち，評価には A セットと B セットを用いた．A セットが学習と評価で雑音環境のミスマッチがない場合，B セットがミスマッチがある場合である．C セットに関しては，今回はチャネルノイズを考慮していないため，評価は行わなかった．性能評価には，各セットにおける音声認識の単語誤り率の平均を用いる．なお各セットには，4 種類の雑音が 5 種類の SN 比（0〜20）で重畳されたサブセットが用意されている．音声認識に用いる HMM の各種パラメタは，すべて AURORA2 の complex backend に準拠し，クリーン音声を学習データとして HMM を学習した．特徴量には，MFCC とそのパワー（ケプストラムの 0 次元），およびその Δ，ΔΔ の合計 $(12+1)\times 3 = 39$ 次元を用い，それに CMN をかけたものを利用した．

k^* の推定には，SPLICE で用いられる y の GMM，NMN-SPLICE で用いられる $y-\hat{n}$ の GMM，[4] で用いられた $[y^T \hat{n}^T]^T$ を LDA したものの GMM，そして今回の提案手法である，$[y^T \hat{n}^T]^T$ の当該フレーム及び前後 4 フレームずつ計 9 フレームを連結した特徴量から多クラス線形 SVM [8] を用いて k^* を推定する手法，加えて特徴量は今回提案した $[y^T \hat{n}^T]^T \times 9$ フレームを用い，それに LDA をかけたものの GMM を用いる手法も比較した．なお，LDA を用いる場合の次元圧縮後の次元数は 39 とした．GMM を用いる場合には，k^* を一意に決定せず，k の事後確率による重み付け和を特徴量強調に利用した．GMM の混

Table 1　Average word error rates of AURORA2 (clean condition training)

k^* の推定	k^* 推定後の変換	備考	Set A	Set B
y の GMM	$A_{k^*}y$	SPLICE	10.73	12.51
y の GMM	$A_{k^*}[y^T \hat{n}^T]^T$		9.50	10.82
$y-\hat{n}$ の GMM	$A_{k^*}(y-\hat{n})$	NMN-SPLICE	10.16	10.25
$y-\hat{n}$ の GMM	$A_{k^*}y$		10.32	10.40
$y-\hat{n}$ の GMM	$A_{k^*}[y^T \hat{n}^T]^T$		9.30	9.52
$[y^T \hat{n}^T]^T$ の LDA の GMM	$A_{k^*}y$		8.92	9.33
$[y^T \hat{n}^T]^T$ の LDA の GMM	$A_{k^*}[y^T \hat{n}^T]^T$	[4] の手法	8.13	8.42
$[y^T \hat{n}^T]^T$ +4neighbors $\times$ 2 の SVM	$A_{k^*}y$	提案法	7.37	8.82
$[y^T \hat{n}^T]^T$ +4neighbors $\times$ 2 の SVM	$A_{k^*}[y^T \hat{n}^T]^T$	提案法	7.41	8.80
$[y^T \hat{n}^T]^T$ +4neighbors $\times$ 2 の LDA の GMM	$A_{k^*}y$	提案法	7.99	8.34
$[y^T \hat{n}^T]^T$ +4neighbors $\times$ 2 の LDA の GMM	$A_{k^*}[y^T \hat{n}^T]^T$	提案法	7.29	7.71

合数 K はすべて，1024 とした．また k^* を推定した後の変換には，y の線形変換，$[y^T \hat{n}^T]^T$ の線形変換の両方を実験した．

結果を Table 1 に示す．k^* を推定するための特徴量として，近隣数フレームを使うこと，k^* 推定後は $A_{k^*}y$ でなく $A_{k^*}[y^T \hat{n}^T]^T$ を使うことが，単語誤り率削減に効果的であることが分かった．識別モデルとして SVM を使う場合と，LDA で次元圧縮をかけた特徴量の GMM を利用するものでは，精度に大きな差は見られなかった．最も単語誤り率が低くなるのは，$[y^T \hat{n}^T]^T$ $+4 \times 2$ neighbors に LDA をかけた特徴量の GMM を用いて k^* を求め，$A_{k^*}[y^T \hat{n}^T]^T$ により音声強調した場合で，Set A で 7.29%，Set B で 7.71% となった．単語誤り率の平均削減率は，SPLICE から 35.46%，NMN-SPLICE から 26.51%，[4] から 9.37% となった．

5　まとめ

高速かつ高精度な音声強調を実現するため，劣化音声の混合分布の事後確率が最大になるインデックスを，クリーン音声状態の識別モデルを用いて推定する手法を提案した．さらに，その識別モデルの特徴量として，観測劣化音声特徴量と推定雑音特徴量を連結したベクトルをさらに近隣数フレーム分連結したものを用いることも提案した．

実験の結果，識別モデルとして SVM を用いる場合と，既に提案されている LDA と GMM を組み合わせて用いる手法では同等程度の精度となったが，特徴量として近隣数フレームを連結したものを用いることで，従来手法を超える精度を実現できた．

参考文献

[1] J. Droppo and A. Acero, “Environmental robustness,” in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., pp. 653–679. Springer, 2008.

[2] V. Stouten, “Robust Automatic Speech Recognition in Time-varying Environments,” *PhD thesis*, 2006.

[3] J. Droppo, Li Deng., and A. Acero, “Evaluation of SPLICE on the Aurora 2 and 3 Tasks,” in *ICSLP*, 2002, pp. 29–32.

[4] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, and K. Hirose, “Mfcc enhancement using joint corrupted and noise feature space for highly non-stationary noise environments,” in *ICASSP*, 2012 (to appear).

[5] Nagendra Kumar and Andreas G. Andreou, “Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition,” *Speech Communication*, vol. 26, no. 4, pp. 283 – 297, 1998.

[6] H. Hermansky, D.P.W. Ellis, and S. Sharma, “Tandem connectionist feature extraction for conventional hmm systems,” in *ICASSP*, 2000, vol. 3, pp. 1635 –1638 vol.3.

[7] H.G. Hirsch and D. Pearce, “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *ISCA ITRW ASR*, 2000.

[8] Kai-Wei Chang and Dan Roth, “Selective block minimization for faster convergence of limited memory large-scale linear models,” in *KDD*, New York, NY, USA, 2011, KDD ’11, pp. 699–707, ACM.