雑音抑圧・特徴量強調・特徴量正規化を組み合わせた雑音に頑健な大語彙音声認識*

☆甲斐常伸, 鈴木雅之, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

音響モデルを学習する環境と，実際に認識システムを動作させる環境の間に音響的ミスマッチがある場合，音声認識システムの性能は多くの場合低下してしまう．例えば背景雑音，録音機器の特性の違いは，このような音響的ミスマッチの要因となる．環境の違いに頑健な音声認識システムを実現するためには，これらのミスマッチを軽減する手法が必要である．

ミスマッチを軽減する 1 つのアプローチとして雑音抑圧がある．このアプローチは特徴量に含まれる雑音を抑圧し，雑音の影響を軽減する手法である．例えば，パワースペクトル領域でウィーナーフィルタを用いて雑音抑圧する手法がある．この手法は特徴量中の雑音成分を推定し，その雑音を打ち消すような線形フィルタを設計する．Advanced Front-End (AFE) [1] と呼ばれる音声認識システムを評価するために標準化されている特徴量抽出法では，2 段階のウィーナーフィルタが使われている．AFE は雑音に頑健な特徴量としてよく用いられており，先行研究においても高い認識性能を示している [2, 3]．

また別のアプローチとして，Mel-Frequency Cepstral Coefficient (MFCC) 領域で雑音成分を軽減し，特徴を強調する特徴量強調というアプローチもある．Stereo-based Piecewise LInear Compensation for Environments (SPLICE) [4] が代表的な手法として挙げられる．SPLICE では雑音による非線形な歪みを区分的線形変換によって近似し，雑音の影響が軽減された特徴量を推定する．線形変換の重み付けは雑音重畳音声の Gaussian Mixture Model (GMM) を用いて計算される．各線形変換とその重み付けは事前に学習データを用いて学習しておくことができるので，実際に特徴強調する際の計算コストは小さい一方で，高い認識率を得ることができる．しかし学習環境とは異なる環境の音声を強調する場合，区分的線形変換による近似が不正確になるため認識率が低下してしまう問題がある．

雑音が重畳した特徴量からクリーンな特徴量を直接推定するのではなく，特徴量に正規化をかけることにより，雑音の影響を軽減する特徴量正規化のアプローチも有効である．Cepstral Mean Normalization (CMN) や Mean and Variance Normalization (MVN) などが特徴量正規化の代表的な手法である [5, 6]．これらの手法は特徴量統計量を正規化することにより，雑音によって歪んだ特徴量を補正することができる．また，Histogram EQualization (HEQ) と呼ばれる正規化手法はもともと画像処理の分野で頻繁に用いられており，近年音声認識の分野においても有効な手法であることが知られるようになった [7, 8]．HEQ は特徴量の分布の形状を正規化する手法であり，CMN や MVN のよりも更に高次の統計量を正規化するという意味で自然な拡張と捉えることができる．

上述の雑音抑圧，特徴量強調，特徴量正規化はそれぞれ雑音によるミスマッチを軽減する効果があるが，今までこれらの手法を組み合わせた手法はあまり研究されていない．AFE，SPLICE，HEQ を組み合わせた特徴量を用いることにより，さらに雑音に頑健な音声認識が可能であると考えられる．我々は既に連続数字音声認識においては，AFE，SPLICE，HEQ の順に適用した特徴量がもっとも雑音に頑健であることを報告した [9]．本研究では大語彙連続音声認識においても，AFE，SPLICE，HEQ の順に適用した特徴量が音響的ミスマッチの軽減に有効であることがわかった．

2 音響的ミスマッチを軽減する手法

この節では本研究で用いる音響的ミスマッチを軽減する手法を紹介する．具体的には，パワースペクトル領域で雑音抑圧を施す AFE，MFCC 領域での特徴量強調の 1 つである SPLICE，特徴量正規化手法の HEQ について説明する．

2.1 Advanced front-end

Advanced front-end (AFE) は音声認識システムの評価を目的として，欧州電気通信標準化機構 (European Telecommunications Standards Institute; ETSI) によって標準化されている特徴量抽出法である [1]．AFE はクライアント側で特徴抽出，圧縮，量子化を施し，そのデータを受け取ったサーバ側で特徴量のデコード，解凍，音声認識を行う分散音声認識を想定した特徴量である．本研究では AFE の雑音抑圧に関する部分を利用するので，以下では雑音抑圧部について詳しく説明する．

雑音除去部のブロック図を Figure 1 に示す．雑音除去は 2 段階のウィーナーフィルタを通すことにより実現される．まずは音声波形から 25msec の窓長で短時間フーリエ変換しスペクトル $S_{in}(f,t)$ を求める．

*Combination of Noise reduction, Feature Enhancement and Feature Normalization for Noise Robust Speech Recognition by T. Kai, M. Suzuki, N. Minematsu, K. Hirose (The University of Tokyo)

$S_{in}(f,t)$ を時間方向に平滑化したものを瞬時パワースペクトル $S_{PSD}(f,t)$ とする．また，ウィーナーフィルタの設計に必要な雑音のパワー $S_N(f,t)$ は $S_{PSD}(f,t)$ と VAD のフラグによって計算することができる．

1 段目のウィーナーフィルタの特性 $H(f,t)$ は以下のようにして計算することができる．

$$H(f,t) = \frac{\eta(f,t)}{1+\eta(f,t)} \tag{1}$$

$$\text{where} \quad \eta(f,t) = \frac{S_{den}(f,t)}{S_N(f,t)} \tag{2}$$

η は SNR を表しており，$S_{den}(f,t)$ は音声の部分のパワーを表しているが直接これを求めることは難しい．そこで，雑音を除去した 1 フレーム前のスペクトル $S_{den3}(f,t-1)$ を用いて $S_{den}(f,t)$ は以下のように計算する．

$$\begin{aligned} S_{den}(f,t) =& \beta S_{den3}(f,t-1) \\ &+ (1-\beta)\max\{S_{PSD}(f,t) - S_N(f,t), 0\} \end{aligned} \tag{3}$$

$H(f,t)$ を求めることができれば，1 段目のウィーナーフィルタを適用したスペクトル S_{den2} は以下のように計算できる．

$$S_{den2}(f,t) = H(f,t)S_{PSD}(f,t) \tag{4}$$

2 段目のウィーナーフィルタの特性 $H_2(f,t)$ は以下のようにして計算できる．

$$H_2(f,t) = \frac{\eta_2(f,t)}{1+\eta_2(f,t)} \tag{5}$$

$$\text{where} \quad \eta_2(f,t) = \max\left\{\frac{S_{den2}(f,t)}{S_N(f,t)}, \eta_{th}\right\} \tag{6}$$

ただし，$\eta_{th} = 0.079432823$ でありこれは SNR が -22dB の場合に対応している．そして $H_2(f,t)$ が求まれば $S_{den3}(f,t)$ は以下のように計算する．

$$S_{den3}(f,t) = H_2(f,t)S_{PSD}(f,t) \tag{7}$$

これを繰り返すことにより次のフレームのフィルタ特性を順番に計算することができる．

ウィーナーフィルタの特性が求まれば，23 チャネルのメルフィルタバンクによってメル尺度に変換し，メル逆コサイン変換することによりウィーナーフィルタのインパルス応答が求まる．このインパルス応答を用いて，入力された音声波形から雑音を軽減した音声波形を求めることができる．

2.2 SPLICE

クリーン音声の特徴量を $\boldsymbol{x}$，雑音重畳音声の特徴量を $\boldsymbol{y}$ とおく．SPLICE は $\boldsymbol{y}$ から $\boldsymbol{x}$ への非線形な変換を区分的線形変換によって近似する．クリーン音声の特徴量の推定値 $\hat{\boldsymbol{x}}$ は以下のように求められる．

$$\hat{\boldsymbol{x}} = \sum_k p(k|\boldsymbol{y})\boldsymbol{A}_k\boldsymbol{y}' \tag{8}$$

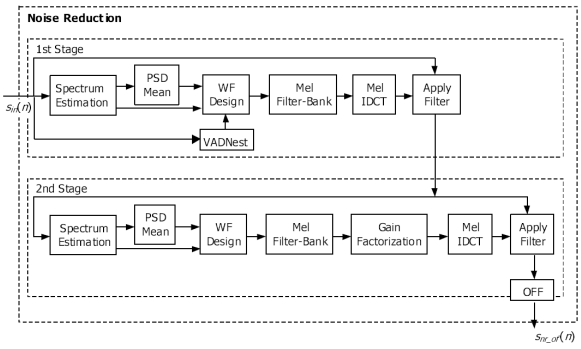

Fig. 1　Block scheme of noise reduction in AFE

ただし，$\boldsymbol{A}_k$ は線形変換行列，$\boldsymbol{y}'$ は $[1 \quad \boldsymbol{y}^\top]^\top$ で表される拡張特徴量ベクトルである．$p(k|\boldsymbol{y})$ はあらかじめ学習されている雑音重畳音声の GMM から計算される．k は GMM を構成する各正規分布のインデックスである．

SPLICE の学習では，まず雑音重畳音声の特徴量 $\boldsymbol{y}$ の確率密度関数を GMM として以下のように学習する．

$$p(\boldsymbol{y}) = \sum_k \pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{9}$$

ただし，π_k,$\boldsymbol{\mu}_k$,$\boldsymbol{\Sigma}_k$ 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．これにより $p(k|\boldsymbol{y})$ は以下のように計算できる．

$$p(k|\boldsymbol{y}) = \frac{p(k)p(\boldsymbol{y}|k)}{p(\boldsymbol{y})} \tag{10}$$

$$= \frac{\pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \tag{11}$$

次に線形変換行列 $\boldsymbol{A}_k$ は，重み付き最小二乗誤差基準で以下のように学習できる．

$$\boldsymbol{A}_k = \underset{\boldsymbol{A}_k}{\text{argmin}} \sum_i p(k|\boldsymbol{y}_i)||\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}_i'||^2 \tag{12}$$

この学習にはステレオデータ，つまりクリーン音声の特徴量 $\boldsymbol{x}_i$ とその音声に雑音を重畳させた音声の特徴量 $\boldsymbol{y}_i$ が必要になる．このように線形変換 $\boldsymbol{A}_k$ と $\boldsymbol{y}$ の GMM は事前に学習しているので，実際に特徴強調する際は式 (8) を計算するだけでよい．これにより計算コストは小さい一方で，雑音が低減された特徴量を得ることができる．ただし学習データにない未知の雑音環境下や非定常雑音環境下では SPLICE の性能は落ちてしまう．

2.3 Histogram equalization

CMN や MVN は 1 次，2 次統計量を正規化する正規化手法であるが，HEQ はより高次の統計量を正規化する．すなわち特徴量ベクトルの確率密度関数が標準正規分布に従うように正規化する [7, 8]．特徴量を $\boldsymbol{x}$

とした場合，正規化後の特徴量 $\hat{\boldsymbol{x}}$ は下記のように計算できる．

$$\hat{\boldsymbol{x}} = F(\boldsymbol{x}) = C_{\text{normal}}^{-1}(C(\boldsymbol{x})) \tag{13}$$

ただし C は $\boldsymbol{x}$ の累積密度関数，C_{normal}^{-1} は平均 0, 分散 1 である標準正規分布の累積密度関数の逆関数である．変換関数 F により $\hat{\boldsymbol{x}}$ は標準正規分布に従うように変換される．変換関数 F は非線形であるため，HEQ は雑音による非線形な歪みを取り除くことができる．

3 音響的ミスマッチを軽減する手法を組み合わせた特徴量

上述の手法によりある程度雑音に対して頑健な音声認識が実現できる．しかしすべての雑音環境に対してそれらの手法が有効に働くわけではなく，それぞれの手法を適用したとしても軽減しきれないミスマッチが残ってしまう．そこで本稿ではこれらの手法を組み合わせることによりミスマッチをさらに軽減した，より雑音に対して頑健な特徴量を考えることができる．

AFE は音声波形を入力し雑音抑圧を行って MFCC を出力するため，組み合わせ手法の一番最初に適用することになる．AFE は HEQ や SPLICE と違って入力音声中の雑音を推定して雑音抑圧に利用するため，雑音環境に左右されない認識率の向上が期待できる．ここではクリーンな音声から AFE によって抽出される特徴量を $\boldsymbol{x}^{(\text{AFE})}$，雑音重畳音声から AFE によって抽出される特徴量を $\boldsymbol{y}^{(\text{AFE})}$ とする．

さらに SPLICE は任意の特徴量を入力として適切に強調することができる [10]．そこであらかじめ AFE をかけた雑音重畳音声の特徴量を SPLICE で強調することができる．$\boldsymbol{y}^{(\text{AFE})}$ を SPLICE で強調するために，$\boldsymbol{y}^{(\text{AFE})}$ の確率密度関数を GMM として学習する．

$$p(\boldsymbol{y}^{(\text{AFE})}) = \sum_k \pi_k \mathcal{N}(\boldsymbol{y}^{(\text{AFE})}; \boldsymbol{\mu}_k \boldsymbol{\Sigma}_k) \tag{14}$$

SPLICE に用いる線形変換行列 $\boldsymbol{A}_k$ は AFE をかけたパラレルデータ $\{\boldsymbol{x}_i^{(\text{AFE})}, \boldsymbol{y}_i^{(\text{AFE})}\}$ を用いて以下の式で学習できる．

$$\boldsymbol{A}_k = \underset{\boldsymbol{A}_k}{\text{argmin}} \sum_i p(k|\boldsymbol{y}^{(\hat{\text{AFE}})}{}_i)||\boldsymbol{x}^{(\hat{\text{AFE}})}{}_i - \boldsymbol{A}_k \boldsymbol{y}^{(\hat{\text{AFE}})}{}_i'||^2 \tag{15}$$

$\boldsymbol{y}^{(\text{AFE})}$ を SPLICE で特徴強調したものを $\boldsymbol{y}^{(\text{AFE,SPLICE})}$ とすると，学習された GMM と $\boldsymbol{A}_k$ を用いて $\boldsymbol{y}^{(\text{AFE,SPLICE})}$ は以下のように推定できる．

$$\boldsymbol{y}^{(\text{AFE,SPLICE})} = \sum_k p(k|\boldsymbol{y}^{(\text{AFE})}) \boldsymbol{A}_k \boldsymbol{y}'^{(\text{AFE})} \tag{16}$$

AFE をかけた特徴量はもとの特徴量よりミスマッチが少ないため，SPLICE による強調がより有効に働くことが期待される．

さらに SPLICE と特徴量正規化を組み合わせた先行研究として，SPLICE をかけた後に CMN をかける手法が報告されている [4]．それと同様に CMN より高い精度が実現できることが報告されている HEQ を，SPLICE の後に適用する．この特徴量 $\boldsymbol{y}^{(\text{AFE,SPLICE,HEQ})}$ は以下のようにあらわされる．

$$\boldsymbol{y}^{(\text{AFE,SPLICE,HEQ})} = C_{\text{normal}}^{-1}(C(\boldsymbol{y}^{(\text{AFE,SPLICE})}) \tag{17}$$

ただし C は $\boldsymbol{y}^{(\text{AFE,SPLICE})}$ の累積密度関数である．HEQ を適用することにより SPLICE では取り除ききれなかった非線形な歪みを軽減できる．また HEQ によって最終的に特徴量は正規分布に従うことになる．このような特徴量は HMM としてモデル化しやすくなるため，認識率の向上に貢献すると考えられる．このようにして AFE，SPLICE，HEQ を順に適用した特徴量を，本稿では AFE-SPLICE-HEQ と呼ぶ．

4 大語彙連続音声認識実験

音響的ミスマッチを軽減する手法を組み合わせた特徴量の性能を確認するため，Aurora-4 データベース [11] を用いて音声認識実験を行った．Aurora-4 データベースは大語彙音声認識タスクでの特徴量の性能や雑音に対する頑健性を比較するために作られたデータベースである．大語彙音声認識タスクは，wall street journal に基づいた音声になっている．

4.1 Aurora-4 データベースでの学習，評価

データベースには音響モデルを学習するための学習セットが用意されている．この学習セットはクリーンな音声のみを用いて音響モデルを学習セットであり，83 名の話者による Sennheiser microphone で収録されたクリーンな音声 7,138 個を使って学習する．また純粋に特徴量の性能のみを比較しやすくするため，音響モデルの学習条件などは [11] の文献と出来るだけそろえて Table 1 のような条件で学習した．また学習するパラメータの数を少なくするためトライフォンの状態共有を行うが，状態共有した後の状態数は 3000 状態程度になるように状態共有の条件を調整した．特徴量としては MFCC+Energy+Δ+ΔΔ（MFCC_E_D_A, $13 \times 3 = 39$ 次元）と，AFE，SPLICE，HEQ をそれぞれ単独で適用した特徴量と，AFE-SPLICE-HEQ の計 5 種類を比較した．SPLICE に用いる GMM の混合数は 512 とし，HEQ は各発声ごとに正規化を行った．

データベースには認識実験を行うための評価セットが 14 種類用意されている．Set1〜7 は Sennheiser microphone で収録された音声で，Set2〜7 の各セットには car，babble，restaurant，street，airport，train station の 6 種類の雑音が SNR5dB から 15dB の範囲で付加されており，加算性雑音の違いによるミスマッチがある評価セットである．Set8〜14 は Sennheiser

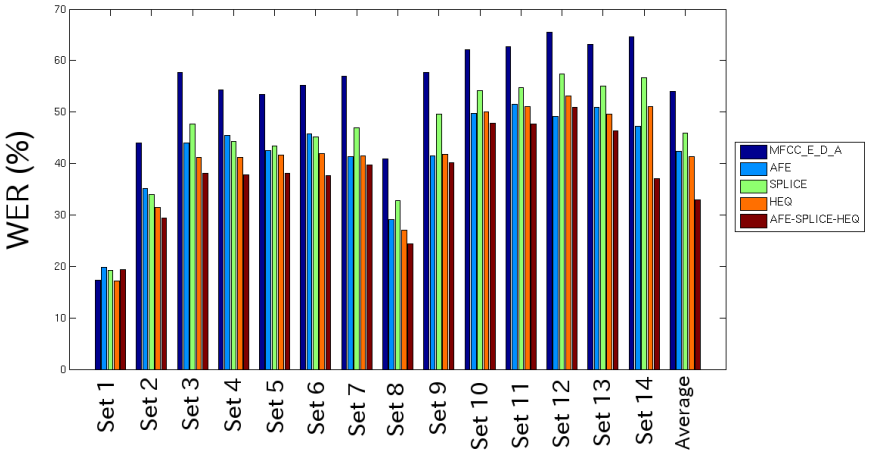

Fig. 2　Word error rate of each test sets

Table 1　A condition of back-end

HMM	cross-word triphone model
状態数	left-to-right, 3 状態
窓長	25msec
シフト長	10msec
出力確率	対角 16 混合 GMM
N-gram	standard WSJ 5k backoff bigram
発音辞書	CMU dictionary (v0.6)

microphone とは異なる音響特性を持っている second microphone で収録された音声で，乗算性雑音の違いによるミスマッチがある．Set9〜14 には 6 種類の雑音が同様に付加されている．各評価セットは 8 人の話者による 330 発声が含まれている．

4.2　認識結果

各評価セットとその平均の Word Error Rate (WER) を Figure 2 に示す．MFCC_E_D_A の結果を見ると，Set 1 に比べて加算性雑音のある Set2〜7 は WER が大きくなっており，さらに乗算性雑音の加わる Set8〜14 はさらに WER が大きくなる．ミスマッチを軽減する手法を単独で適用した特徴量は Set 1 以外の評価セットでは WER が低下している．特に HEQ は 3 つの手法の中では平均して削減率が大きく，多くの評価セットで小さい WER を示している．しかし AFE と HEQ の結果を比較してみると，Set8〜14 においては HEQ よりも AFE の WER の方が小さい場合もあり，AFE は乗算性雑音があるような環境でもミスマッチを有効に軽減できることがわかる．AFE-SPLICE-HEQ の認識結果を見るともともと音響モデルと評価データのミスマッチがない Set 1 では WER が大きくなっているものの，Set2〜14 では軒並み WER が大きく低下している．各手法がそれぞれ別の種類の雑音環境のミスマッチを軽減しており，組み合わせることによりさらに広範囲の雑音環境に対して頑健になっていることがわかる．

5　まとめ

本稿では雑音に対してより頑健な音声認識を実現するために，既存手法の AFE，SPLICE，HEQ を組み合わせた特徴量の性能を調査した．大語彙連続音声認識タスクの実験を行った結果，AFE，SPLICE，HEQ を順に適用した特徴量は，連続数字音声認識タスクの場合と同様に大きな認識結果の改善が見られた．AFE-SPLICE-HEQ の特徴量は MFCC_E_D_A の特徴量に比べて WER が 37% 改善した．

参考文献

[1] ETSI, “ES 202 050 v1.1.5, Speech Processing, Transmission and Quality Aspects (STQ), Distributed speech recognition, Advanced front-end feature extraction algorithm, Compression algorithms,” Tech. Rep., 2007.
[2] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, in *International Conference on Spoken Language Processing*, pp. 1–4, 2002.
[3] O. Kalinli, M. L. Seltzer, J. Droppo, and A. Acero, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1889–1901, 2010.
[4] J. Droppo, L. Deng, and A. Acero, in *International Conference on Spoken Language Processing*, pp. 29–32, 2002.
[5] C. R. Jankowski, H. D. H. Vo, and R. P. Lippmann, *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 286–293, 1995.
[6] O. Viikki and K. Laurila, *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.
[7] A. de la Torre, A. Peinado, J. Segura, J. Perez-Cordoba, M. Benitez, and A. Rubio, *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, May 2005.
[8] Y. Suh, M. Ji, and H. Kim, *IEEE Signal Processing Letters*, vol. 14, no. 4, pp. 287–290, 2007.
[9] 甲斐常伸, 鈴木雅之, 峯松信明, 広瀬啓吉, 信学技報, SP2012-28, vol. 112, no. 47, pp. 161–166, 2012.
[10] M. Suzuki, T. Yoshioka, S. Watanabe, N. Minematsu, and K. Hirose, in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 4109–4112, 2012.
[11] N. Parihar and J. Picone, “DSR Front End LVCSR Evaluation,” Aurora Working Group, Tech. Rep., 2002.