特徴量正規化と SPLICE を組み合わせた雑環境下音声認識*

☆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，実際に認識を行う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．背景雑音，録音機器の音響特性，チャネル歪みなどの様々な原因によって環境間のミスマッチは大きくなってしまう．雑音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手法の 1 つに, 特徴量正規化があげられる．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 [1, 2]．CMN は特徴量ベクトルからその平均値を差し引くことにより，特徴量ベクトルの平均値を 0 に正規化する．これにより定常なチャネル歪みの影響を効率的に取り除くことができる．MVN は特徴ベクトルの平均と同時に分散も正規化することにより，背景雑音の影響を軽減することができる．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 [3, 4]．HEQ は特徴量のヒストグラムが正規分布の形状になるように変換を施す．この変換は非線形になるので，HEQ は雑音による非線形な歪みを軽減することができると考えられる．

特徴量正規化とは異なるアプローチとして特徴量強調がある．特徴量強調の代表的な手法として Stereo-based PIecewise Linear Compensation for Environments (SPLICE) [5] が挙げられる．クリーン音声に雑音が重畳すると，一般に特徴量は非線形な変化をうける．SPLICE では雑音重畳音声からクリーン音声への非線形な変換を，線形変換の重み付け和によって近似することにより特徴量から雑音の影響を取り除く．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．各線形変換と雑音重畳音声の GMM 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，非線形変換の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．

特徴量正規化，特徴量強調はそれぞれ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．そこで本稿では特徴量正規化と SPLICE の組み合わせた手法を提案し，より雑音に頑健な音声認識を実現する．SPLICE をかけた後に特徴強調を施すことにより，SPLICE では取り除ききれなかった雑音を軽減することが期待できる．また SPLICE はどのようなドメインの特徴量でも強調することができる．そこで，正規化を施した特徴量を SPLICE の入力とする手法も提案する．正規化した特徴量はもとの特徴量よりミスマッチが軽減されているので，SPLICE による強調がより効果的に働くと考えられる．

2 特徴量正規化・強調とその組み合わせ

この節ではこれまで提案された雑音に頑健な特徴量を得る手法を紹介する．特に，CMN と HEQ の特徴量正規化手法と特徴強調の 1 つである SPLICE を紹介する．また特徴量正規化と SPLICE を組み合わせる提案手法に関しても詳しく説明する．

2.1 特徴量正規化

CMN はケプストラムからその平均値を差し引くことにより平均値を 0 に正規化する．ケプストラムは対数スペクトル領域の特徴量なので，チャネル歪みによる影響が軽減される．正規化を施した特徴量 $\hat{\boldsymbol{x}}$ は以下のように表される．

$$\hat{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{\mu} \tag{1}$$

ただし $\boldsymbol{\mu}$ は正規化前の特徴量 $\boldsymbol{x}$ の平均値である．$\hat{\boldsymbol{x}}$ の平均値は 0 になっており，CMN は特徴量の 1 次統計量を正規化する手法だと言える．MVN はこれに加えて 2 次統計量まで正規化する手法である．CMN は非常に単純であるが雑音に頑健な音声認識を実現する手法である．

HEQ では $\boldsymbol{x}$ から $\hat{\boldsymbol{x}}$ への変換 F を求める必要がある．F は以下のように計算される．

$$\hat{\boldsymbol{x}} = F(\boldsymbol{x}) = C_{\text{normal}}^{-1}(C(\boldsymbol{x})) \tag{2}$$

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．この変換により正規化した特徴量 $\hat{\boldsymbol{x}}$ のヒストグラムは標準正規分布となる．言い換えると HEQ は特徴量のすべての統計量を正規化する手法だと言える．こ

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

の点で HEQ は CMN や MVN の拡張と考えることができる. 変換 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.

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線形変換の重み付け和によって近似する．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.

$$\hat{\boldsymbol{x}} = \sum_k p(k|\boldsymbol{y}) \boldsymbol{A}_k \boldsymbol{y}' \tag{3}$$

ただし，$\boldsymbol{A}_k$ は線形変換，$\boldsymbol{y}'$ は $[1 \ \ \boldsymbol{y}^\top]^\top$ 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.

$$p(\boldsymbol{y}) = \sum_k \pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{4}$$

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．次に線形変換 $\boldsymbol{A}_k$ 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.

$$\boldsymbol{A}_k = \underset{\boldsymbol{A}_k}{\operatorname{argmin}} \sum_i p(k|\boldsymbol{y}_i) ||\boldsymbol{x}_i - \boldsymbol{A}_k \boldsymbol{y}_i'||^2 \tag{5}$$

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 (3) を計算するだけでよい．これにより計算コストは小さいが雑音に頑健な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 HEQ と SPLICE を組み合わせた特徴量

以上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．しかしすべての場合におい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提案する.

特徴量正規化と SPLICE 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 (SPLICE-CMN) が報告されている [5]．そこで SPLICE の後に HEQ を施す特徴量 (SPLICE-HEQ) を提案する．この特徴量 $\hat{\hat{x}}$ は以下のようにあらわされる.

$$\hat{\boldsymbol{x}} = \sum_k p(k|\boldsymbol{y}) \boldsymbol{A}_k \boldsymbol{y}' \tag{6}$$

$$\hat{\hat{\boldsymbol{x}}} = C_{\text{normal}}^{-1}(C(\hat{\boldsymbol{x}})) \tag{7}$$

ただし $\hat{\boldsymbol{x}}$ は SPLICE で強調した特徴である．HEQ を施すことにより 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と考えられる.

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 [6]．そこであらかじめ HEQ をかけた雑音重畳音声の特徴量を SPLICE で強調することを提案する (HEQ-SPLICE)．まずは雑音重畳音声の特徴量 $\boldsymbol{y}$ を HEQ で正規化し $\hat{\boldsymbol{y}}$ を得る.

$$\hat{\boldsymbol{y}} = C_{\text{normal}}^{-1}(C(\boldsymbol{y})) \tag{8}$$

次に $\hat{\boldsymbol{y}}$ を SPLICE で強調するために，$\hat{\boldsymbol{y}}$ の確率密度関数を GMM として学習する.

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

SPLICE に用いる線形変換 $\boldsymbol{A}_k$ は正規化されたパラレルデータ $\{\hat{\boldsymbol{x}}_i, \hat{\boldsymbol{y}}_i\}$ を用いて以下の式で学習できる.

$$\boldsymbol{A}_k = \underset{\boldsymbol{A}_k}{\operatorname{argmin}} \sum_i p(k|\hat{\boldsymbol{y}}_i) ||\hat{\boldsymbol{x}}_i - \boldsymbol{A}_k \hat{\boldsymbol{y}}_i'||^2 \ . \tag{10}$$

以上のように学習された GMM と $\boldsymbol{A}_k$ を用いて最終的に得られる特徴量 $\hat{\hat{\boldsymbol{x}}}$ は以下のようになる.

$$\hat{\hat{\boldsymbol{x}}} = \sum_k p(k|\hat{\boldsymbol{y}}) \boldsymbol{A}_k \hat{\boldsymbol{y}}' \tag{11}$$

正規化され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.

今回提案した 2 つの手法 HEQ-SPLICE，SPLICE-HEQ を組み合わせ，前にも後ろにも HEQ をかけた特徴量にすることもできる (HEQ-SPLICE-HEQ).

3 実験

様々な組み合わせによる性能を比較するため Aurora-2 データベース [7] を用いて実験を行った．このデータベースは背景雑音とチャネル歪みによるミスマッチのある英語連続数字発声で構成されている．データベースには音声認識システムの性能を比較するため，3 つのテストセットがある．Set A は学習セットと同じ背景雑音が重畳された音声，Set B は学習セットと異なる背景雑音が重畳された音声，Set C は A，B とは異なるチャネル歪みが加えられた音声を含んでいる．音響モデルは 16 状態の単語 HMM で，各状態は対角 20 混合 GMM の出力確率を持っている．クリーン音声のみを学習データとして学習した音響モデル (clean acoustic models) とクリーン音声と

Table 1　Summary of word accuracies for HEQ using clean acoustic models

HEQ	Set A				Set B				Set C		
	N1	N2	N3	N4	N1	N2	N3	N4	N1	N2	Average
CLEAN	99.66	99.70	99.46	99.63	99.66	99.70	99.46	99.63	99.69	99.64	99.62
SNR20	98.04	98.61	98.81	97.59	98.96	97.97	98.81	98.33	98.19	98.22	98.35
SNR15	95.64	96.49	97.05	94.91	96.90	96.10	97.29	96.02	95.46	95.98	96.18
SNR10	88.70	91.05	91.83	87.84	92.05	90.39	93.20	91.08	89.35	90.02	90.55
SNR5	74.95	73.88	75.22	73.28	77.22	75.24	78.17	76.18	73.87	76.39	75.44
SNR0	46.48	43.05	46.47	47.49	50.02	45.95	51.21	47.27	45.96	46.52	47.04
SNR-5	20.11	16.60	18.91	22.52	20.39	18.23	21.92	18.61	19.50	18.26	19.51
Average	80.76	80.62	81.88	80.22	83.03	81.13	83.74	81.78	80.57	81.43	81.52

Table 2　Summary of word accuracies for SPLICE using clean acoustic models

SPLICE	Set A				Set B				Set C		
	N1	N2	N3	N4	N1	N2	N3	N4	N1	N2	Average
CLEAN	99.48	99.40	99.37	99.48	99.48	99.40	99.37	99.48	99.57	99.49	99.45
SNR20	98.89	99.06	99.14	98.61	99.14	98.58	98.93	98.89	98.68	97.76	98.77
SNR15	97.64	98.55	98.45	97.78	98.53	97.25	98.06	97.62	97.18	95.56	97.66
SNR10	95.27	96.16	96.03	94.57	95.67	91.02	94.09	91.89	91.93	88.72	93.54
SNR5	87.96	81.80	82.64	83.74	83.39	67.74	77.57	71.03	75.04	68.20	77.91
SNR0	63.28	42.78	46.73	57.11	53.67	32.50	41.75	26.54	40.87	35.25	44.05
SNR-5	28.89	13.48	14.76	24.00	18.70	12.06	12.76	7.56	15.75	15.30	16.33
Average	88.61	83.67	84.60	86.36	86.08	77.42	82.08	77.19	80.74	77.10	82.39

Table 3　Summary of word accuracies for HEQ-SPLICE-HEQ using clean acoustic models

HEQ-SPLICE-HEQ	Set A				Set B				Set C		
	N1	N2	N3	N4	N1	N2	N3	N4	N1	N2	Average
CLEAN	99.66	99.70	99.46	99.63	99.66	99.70	99.46	99.63	99.69	99.64	99.62
SNR20	98.71	99.00	99.02	98.89	98.96	98.61	99.11	99.04	98.50	98.40	98.82
SNR15	97.30	98.28	98.51	97.66	98.19	97.97	98.90	98.46	97.45	97.94	98.07
SNR10	95.12	96.04	96.48	94.69	96.19	95.28	97.23	95.83	95.27	94.92	95.71
SNR5	88.24	88.03	89.20	86.55	89.19	86.88	90.58	87.90	88.46	86.40	88.14
SNR0	66.56	60.13	61.29	68.93	67.15	60.85	69.79	63.13	66.29	61.79	64.59
SNR-5	29.94	22.82	21.47	36.62	29.94	25.03	29.91	26.29	29.54	25.03	27.66
Average	89.19	88.30	88.90	89.34	89.94	87.92	91.12	88.87	89.19	87.89	89.07

雑音重畳音声の両方を学習データとして学習した音響モデル (multi-conditions models) の 2 つの音響モデルを用意しそれぞれについて実験を行った．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 の計 39 次元をもちいた．特徴量正規化と SPLICE の組み合わせは次の 6 種類を用いた．1) SPLICE のみ，2) HEQ のみ，3) SPLICE の後に CMN，4) SPLICE の後に HEQ，5) SPLICE の前に HEQ，6) SPLICE の前後に HEQ．SPLICE で用いる GMM は 1024 混合で学習している．また特徴量正規化は 1 発声ごとに施している．

Figure 1 に clean acoustic models を用いた認識の認識結果の平均を示す．また Table 1，2，3 に HEQ，SPLICE，HEQ-SPLICE-HEQ の認識結果を示す．Figure 1 から SPLICE は Set A においては高い認識率を示すものの，Set B と Set C では認識率が低下している．対照的に HEQ は全セットで平均的な認識率を示している．また SPLICE-CMN よりも SPLICE-HEQ の方が全セットで認識率が向上しており，CMN を単純に HEQ に置き換えるだけでも認識率が向上することがわかる．SPLICE-HEQ や HEQ-SPLICE をみると, SPLICE と HEQ を組み合わせて適用することにより大きく認識率が上がっている．SPLICE-HEQ と HEQ-SPLICE との比較では SPLICE-HEQ の方が認識率が高く，SPLICE の後にかけた HEQ の方が認識率の向上に大きく貢献している．しかし，SPLICE の前にかけた HEQ は Set B，Set C の認識率向上に寄与していることがわかる．HEQ-SPLICE-HEQ の組み合わせが最も高い認識率を示し，SPLICE のみと比べて 41%，SPLICE-CMN に比べて 25% 単語誤り率が改善された．

Figure 2 に multi acoustic models を用いた認識の認識結果を示す．全体的な傾向は clean acoustic models での認識と似ており，SPLICE-CMN と SPLICE-

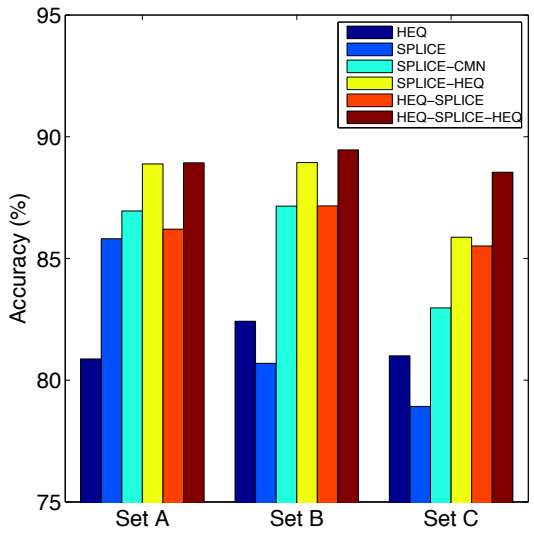

Fig. 1　Average word accuracies of Aurora-2 recognition results using clean acoustic models

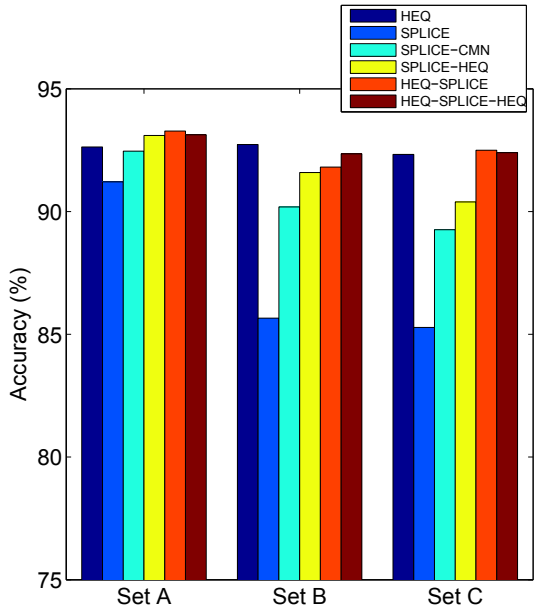

Fig. 2　Average word accuracies of Aurora-2 recognition results using multi-conditions acoustic models

HEQ を比較すると SPLICE-HEQ が全セットにおいて認識率が向上している．しかし HEQ 単独での認識率が十分高く，HEQ の前に SPLICE をかける SPLICE-HEQ ではむしろ認識率が低下している．HEQ-SPLICE-HEQ が全組み合わせの中で最高性能を示しているものの，HEQ 単独との差は小さいものであった．

4　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，SPLICE で強調した後に HEQ で正規化した特徴量，また HEQ で正規化した特徴量を SPLICE で強調した特徴量を提案した．実験結果より HEQ と SPLICE を組み合わせた特徴量は clean acoustic model，multi acoustic model のどちらにおいても従来より高い認識率を示した．

今後は雑音に対する頑健性を向上させる別の手法との比較が必要である．例えばウィーナーフィルタによる雑音除去などを施した advanced front-end [8] との比較があげられる．また Aurora-3 や Aurora-4 といった他のデータベースを用いて実験し，同様の傾向が得られるかを確かめたい．

参考文献

[1] Jankowski, Vo, and Lippmann, “A comparison of signal processing front ends for automatic word recognition,” *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 286–293, 1995.

[2] Viikki and Laurila, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[3] Torre, Peinado, Segura, Perez-Cordoba, Benitez, and Rubio, “Histogram equalization of speech representation for robust speech recognition,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 355–366, May 2005.

[4] Suh, Ji, and Kim, “Probabilistic Class Histogram Equalization for Robust Speech Recognition,” *IEEE Signal Process. Lett.*, vol. 14, no. 4, pp. 287–290, 2007.

[5] Droppo, Deng, and Acero, “Evaluation of SPLICE on the Aurora 2 and 3 tasks,” in *Proc. ICSLP*, 2002, pp. 29–32.

[6] Suzuki, Yoshioka, Watanabe, Minematsu, and Hirose, “Framewise MFCC Enhancement in Observation and Noise Feature Space,” in *Proc. ICASSP*, 2012.

[7] Hirsch, Pearce, Eurolab, Gmbh, and Labs, “The Aurora Experimental Framework For The Performance Evaluation of Speech Recognition Systems Under Noisy Conditions,” in *ISCA ITRW ASR2000*, Paris, France, 2000, pp. 181–188.

[8] Macho, Mauuary, Noé, Cheng, Ealey, Jouvet, Kelleher, Pearce, Saadoun, R, and Ag, “Evaluation of a noise-robust DSR front-end on Aurora databases,” in *Proc. ICSLP*, 2002, pp. 1–4.