Eigen–SPLICEを用いた雑音環境下における音声認識の実験的検討*

☆千々岩圭吾, 鈴木雅之, 峯松信明, 広瀬啓吉 (東京大学)

1 はじめに

近年，スマートフォンやカーナビゲーションシステムの普及により，雑音環境下における音声認識の必要性が高まってきている．しかし，雑音によって歪められた音声は，クリーンな音声を用いて学習した音響モデルとのミスマッチを生じ，認識精度が著しく低下するという問題がある [1].

この問題に対応するため，独立成分分析 [2] やスペクトルサブトラクション [3] などの手法が多く提案されている．その中でも音声認識に用いる特徴量に対するもので，特徴量の統計的性質を表すモデルを用いて特徴量を変換する手法が着目されている．これは，音声特徴量の統計的な性質を混合正規分布 (GMM:Gaussian Mixture Model) などで事前に学習し，そのモデルを用いて歪んだ特徴量をクリーンな特徴量に変換しており，高精度な雑音抑制を実現する．代表的なものとしては，雑音付加音声の特徴量の分布を GMM で学習し，クリーン音声と雑音付加音声のパラレルデータから変換を学習する SPLICE(Stereo-based Piecewise LInear Compensation for Environments)[4]，クリーン音声の特徴量の GMM と，適応用の雑音付加音声特徴量からベクトルテーラー展開 (Vector Taylor Series:VTS) を用いて適応する手法 [5] などがある．これらは雑音環境下の音声認識用データベース AURORA-2[6] において高い性能を示している [5].

しかし，これらの特徴量変換による雑音抑制手法にもいくつかの問題がある．まず VTS を用いた手法については，計算の単純な対数スペクトルなどの特徴量においては計算量が小さくて済むが，MFCC などのより複雑な特徴量に関しては，多くの計算量が必要になるという問題がある．一方，SPLICE は変換関数を学習した環境と入力音声の雑音環境が似ていることを暗に仮定しており，未知の雑音環境下においては十分な性能を発揮することが保証されていない．SPLICE のこの問題点を解決するために，まず入力音声の雑音環境の種類を推定し，その推定結果の雑音環境に応じた GMM と区分的線形変換を用いて，雑音抑制する EMS(Evironmental Model Selection) の手法も提案されている．しかし，この手法も事前に学習した変換方法でしか変換できないため，未知の雑音環境下において高い性能を発揮することが保証されていない．

そこで今回は，特徴量の変換方法を入力の雑音環境に対して適応することで，未知の雑音環境下においても十分な性能を発揮させることを試みる．変換方法を適応するためには，本来であれば各コンポーネント毎に高次元の補正ベクトルのパラメータを推定する必要があり，多くの計算量を要してしまう．それを避けるため，提案手法では変換を表すベクトルに主成分分析を施すことで推定すべきパラメータを削減する．これにより，少量の適応データでも適切な特徴量変換を推定できるようにした．

本報告では，まず研究のベースとなった SPLICE について詳しく説明する．その上で，提案手法の手順を具体的に述べ，その有効性を実験的に示す．最後に，現状の問題点と今後の展望について説明する．

2 SPLICE

この節では，SPLICE[4] について詳しく説明する．SPLICE は，雑音付加音声の特徴量の統計的な性質を GMM でモデル化し，その GMM の各コンポーネント毎に特徴量を線形変換することで，雑音付加音声からクリーン音声への変換を実現する．これは，雑音付加音声の特徴量からクリーン音声の特徴量への非線形変換を，区分的線形変換によって近似的に実現している．以下，従来手法の各段階を詳しく説明する．

2.1 仮定

SPLICE は，まず雑音環境下の音声の特徴量ベクトルの分布が GMM で表現出来ると仮定している．その分布を EM(Expectation-Maximization) アルゴリズムによって学習する．

$$\begin{aligned} p(\boldsymbol{y}) &= \sum_s p(\boldsymbol{y}, s) = \sum_s p(\boldsymbol{y}|s)p(s), \text{但し} \\ p(\boldsymbol{y}|s) &= N(\boldsymbol{y}; \mu_s, \Sigma_s) \end{aligned} \tag{1}$$

ここで $\boldsymbol{y}, s, \mu, \Sigma$ はそれぞれ，雑音付加音声の特徴量ベクトル，GMM の各コンポーネントのインデックス，平均，分散を表す．

ここで，以下の仮定を設ける．雑音付加音声の特徴量ベクトルとそれが所属するコンポーネントが与えられたとき，その変換後のクリーン音声の確率分布も正規分布によって表され，その正規分布の平均は雑音付加音声のアフィン変換によって表されると仮定する．つまり次式で表される仮定を設ける．

$$p(\boldsymbol{x}|\boldsymbol{y}, s) = N(\boldsymbol{x}; \boldsymbol{A}_s\boldsymbol{y} + \boldsymbol{r}_s, \Gamma_s) \tag{2}$$

ここで，$\boldsymbol{x}, \boldsymbol{r}_s, \boldsymbol{A}_s, \Gamma_s$ はそれぞれ，クリーン音声の特徴量ベクトル，コンポーネント s における平均ベクトルの補正ベクトル，線形変換を表す行列，分散を表す．

*Evaluation of speech recognition in noisy environments using Eigen-SPLICE by CHIJIIWA, Keigo and SUZUKI, Masayuki, MINEMATSU, Nobuaki and HIROSE, Keikichi, (The University of Tokyo)

2.2 変換手法

以上のような仮定を設けることによって，ある入力の雑音付加音声の特徴量ベクトルが与えられたときの，出力のクリーン音声の特徴量は，下記の期待値として推定することができる．

$$\hat{\boldsymbol{x}} = E_x[\boldsymbol{x}|\boldsymbol{y}] = \sum_s p(s|\boldsymbol{y}) E_x[\boldsymbol{x}|\boldsymbol{y}, s] \tag{3}$$

ここで，(2) を用いると，

$$E_x[\boldsymbol{x}|\boldsymbol{y}, s] = \boldsymbol{A}_s \boldsymbol{y} + \boldsymbol{r}_s \tag{4}$$

となる．よって，雑音付加音声の特徴量から推定したクリーン音声の特徴量は，以下のようになる．

$$\hat{\boldsymbol{x}} = \sum_s p(s|\boldsymbol{y}) \left(\boldsymbol{A}_s \boldsymbol{y} + \boldsymbol{r}_s\right) \tag{5}$$

この変換は，入力が GMM のどのコンポーネントに所属しているかという事後確率を求め，その事後確率を重みとして各コンポーネントでの変換結果の重み付け和と解釈することが出来る．

2.3 学習

以上の特徴量の変換に必要なコンポーネント s における $\boldsymbol{r}_s, \boldsymbol{A}_s$，つまり $\boldsymbol{y}, s$ が分かったときの条件付き確率分布 $p(\boldsymbol{x}|\boldsymbol{y}, s)$ は最尤推定で，クリーン音声とそれに雑音が付加された雑音付加音声の対を用いて以下のように学習する．

$$\begin{aligned}
\boldsymbol{r}_s &= \frac{\sum_t p(s|\boldsymbol{y}_t)(\boldsymbol{x}_t - \boldsymbol{y}_t)}{\sum_t p(s|\boldsymbol{y}_t)}, \\
\boldsymbol{A}_s &= \frac{\sum_t p(s|\boldsymbol{y}_t)\left(\boldsymbol{x}_t - \boldsymbol{r}_s\right)\boldsymbol{y}_t^{\mathrm{T}}}{\sum_t p(s|\boldsymbol{y}_t)\boldsymbol{y}_t\boldsymbol{y}_t^{\mathrm{T}}}, \text{但し} \\
p(s|\boldsymbol{y}_t) &= \frac{p(\boldsymbol{y}_t|s)p(s)}{\sum_s p(\boldsymbol{y}_t|n)p(s)}
\end{aligned} \tag{6}$$

このとき t は特徴量系列のインデックスである．あるコンポーネントの補正関数は $\boldsymbol{r}_s$ は，雑音付加音声の特徴量がそのコンポーネントに所属する確率を重みとして，雑音付加音声とクリーン音声の差分の重み付け平均とすることが出来る．また，そのときの所属する事後確率は，ベイズの定理を用いて求めることが出来る．このときのクリーン音声と雑音付加音声の対，パラレルデータは実環境では口元のマイクで録音した歪が殆ど無い音声と離れたところで録音した歪んだ音声というようにして得ることが可能である．今回用いた AURORA–2 データベースではシミュレーションによって，クリーン音声に雑音を重畳している．

この SPLICE は用いる GMM の混合数を上げていくと，つまり特徴量空間をより細かく分割していくと，行列 $\boldsymbol{A}_s$ を単位行列にしても十分な効果が得られることが知られている [4]．ただし，本報告では混合数を抑えたため，$\boldsymbol{A}_s$ も非単位行列として推定した．

2.4 SPLICE の問題点

以上の SPLICE は定常雑音環境を暗に仮定しており，非定常雑音環境下においては十分な性能を発揮することが保証されていない．そこで，環境毎に GMM および区分的線形変換の方法を切り替える EMS が SPLICE の改善手法として提案されている [7]．これは，各々の学習環境毎に GMM を学習し，その GMM を用いて学習環境毎の区分的線形変換を学習する．そして変換の際には，入力系列 $\boldsymbol{y}_t$ がどの環境 e に依存しているかを推定し，

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(\boldsymbol{y}_t|e) \tag{7}$$

もっとも尤度の高い環境の GMM とその区分的線形変換関数を用いて変換する．このとき，環境の推定誤差を抑えるために，推定された環境の系列を時間方向でスムージングする．こうすることで，非定常な雑音環境においてもその雑音環境に適した変換を用いて雑音除去することが出来るようになる．

しかし，この EMS も事前に学習した変換方法でしか変換出来ないため，未知の雑音環境下において高い性能を発揮することが保証されていない．そこで，今回は入力環境に対して変換関数のパラメータを適応することを試みる．具体的には，少数の適応データで変換関数を適応できるように，変換関数を表すベクトルに主成分分析を施し，推定すべきパラメータを削減する．

3 提案手法

提案手法は，区分的線形変換のパラメータを入力環境に対して適応することで，未知雑音環境下でも性能を発揮することを目指す．また，少数のデータでも適応出来るように，変換を表すベクトルに主成分分析を施すことで，推定すべきパラメータを削減する．今回は簡単のため各コンポーネント毎の変換関数のうち，補正ベクトル $\boldsymbol{r}_s$ の項のみについて適応した．主成分分析によって推定すべきパラメータを削減する手法は，Eigen–MLLR[8] や固有声に基づいた性質変換 [9] などで広く用いられている．しかし，提案手法は確率分布のパラメータではなく変換関数のパラメータを推定するので，入力音声だけでなく雑音付加音声とクリーン音声のパラレルデータを必要とする点でこれらの手法と異なる．ところが，未知環境においてパラレルデータを得ることは難しい．そのため提案手法では，雑音付加音声から雑音のみの区間を抽出し，それを手持ちの学習データのクリーン音声に重畳することで擬似的にパラレルデータを作成する．

3.1 主成分の学習

まず，従来の SPLICE を用いて学習データの中の全ての環境に共通な変換関数の行列項として $\boldsymbol{A}_s^0$ を学習する．次に，学習データの中の特定の種類・SNR の雑音環境での変換関数の補正ベクトル項 $\boldsymbol{r}_s^i$ を次式のように求める．ただし，このとき添字 i は特定の種

類・SNR の雑音環境を表すインデックスである．

$$\hat{\boldsymbol{r}}_s^i = \underset{\boldsymbol{r}_s^i}{\operatorname{argmin}} \sum_t \sum_s p(s|\boldsymbol{y}_t^i) \left\{ \boldsymbol{x}_t^i - (\boldsymbol{A}_s^0 \boldsymbol{y}_t^i + \boldsymbol{r}_s^i) \right\}^2 \tag{8}$$

こうすることで学習データ中の特定の環境のための変換パレメータの $\boldsymbol{r}_s^i$ を得る．ただし，行列 $\boldsymbol{A}_s^0$ に関しては全ての環境で共通である．

次に，GMM の各コンポーネントの補正関数を全て連結することによって，特定の環境の変換関数を表すスーパーベクトル $\boldsymbol{SV}^i$ を得る．ただし，S は GMM の混合数である．

$$\boldsymbol{SV}^i = \{\hat{\boldsymbol{r}_1^i}, \cdots, \hat{\boldsymbol{r}_s^i}, \cdots, \hat{\boldsymbol{r}_S^i}\} \tag{9}$$

このスーパーベクトルは，学習データの中の全ての環境に関して各々学習する．そして，得られた複数のスーパーベクトルに対して主成分分析を施す．学習データの中の全ての環境の変換関数のスーバーベクトルの平均を表すバイアスベクトル $\boldsymbol{BV}$ と，その主成分を表すベクトル $\boldsymbol{PC}^m$ を得る．ただし，m は主成分のインデックスである．

$$\begin{aligned} \boldsymbol{BV} &= \{\boldsymbol{b}_1, \cdots, \boldsymbol{b}_s, \cdots, \boldsymbol{b}_S\} && (10) \\ \boldsymbol{PC}^1 &= \{\boldsymbol{c}_1^1, \cdots, \boldsymbol{c}_s^1, \cdots, \boldsymbol{c}_S^1\} \\ &\vdots \\ \boldsymbol{PC}^M &= \{\boldsymbol{c}_1^M, \cdots, \boldsymbol{c}_s^M, \cdots, \boldsymbol{c}_S^M\} && (11) \end{aligned}$$

これらのバイアスベクトルと主成分を用いて，ある環境での変換関数は以下のように表せる．

$$\begin{aligned} \hat{\boldsymbol{x}}_t &= \sum_s p(s|\boldsymbol{y}_t) \left(\boldsymbol{A}_s^0 \boldsymbol{y}_t + \boldsymbol{B}_s \boldsymbol{w} + \boldsymbol{b}_s \right), \text{但し} \\ \boldsymbol{B}_s &= \{\boldsymbol{c}_s^{1\mathrm{T}}, \cdots, \boldsymbol{c}_s^{M\mathrm{T}}\} && (12) \end{aligned}$$

ただし，ここで添字 $\boldsymbol{w}$ は主成分の重み付けを表す．また，添字 T は行列の転置を表す．

3.2 重みの推定

以上の操作によって，新たな雑音環境が現れたときに，適応すべき変換関数のパラメータが高次元の補正ベクトルから，低次元の重みベクトルになった．次にこの重み $\boldsymbol{w}$ の推定について述べる．この重みベクトルは，少数の未知環境下における雑音付加音声とクリーン音声のパラレルデータを用いて，最小誤差基準で下記のように推定される．

$$\hat{\boldsymbol{w}} = \underset{w}{\operatorname{argmin}} \sum_t \{\boldsymbol{x}_t - \sum_s p(s|\boldsymbol{y}_t) \left(\boldsymbol{A}_s \boldsymbol{y}_t + \boldsymbol{B}_s \boldsymbol{w} + \boldsymbol{b}_s \right)\}^2 \tag{13}$$

これを解くと，以下のような重み付けになる．

$$\begin{aligned} \hat{\boldsymbol{w}} &= \left(\sum_t \boldsymbol{M}_t^{\mathrm{T}} \boldsymbol{M}_t \right)^{-1} \left(\sum_t \boldsymbol{M}_t^{\mathrm{T}} \boldsymbol{E}_t \right), \text{但し} \\ \boldsymbol{M}_t &= \sum_s p(s|\boldsymbol{y}_t) \boldsymbol{B}_s \\ \boldsymbol{E}_t &= \boldsymbol{x}_t - \sum_s p(s|\boldsymbol{y}_t) \left(\boldsymbol{A}_s \boldsymbol{y}_t + \boldsymbol{b}_s, \right) && (14) \end{aligned}$$

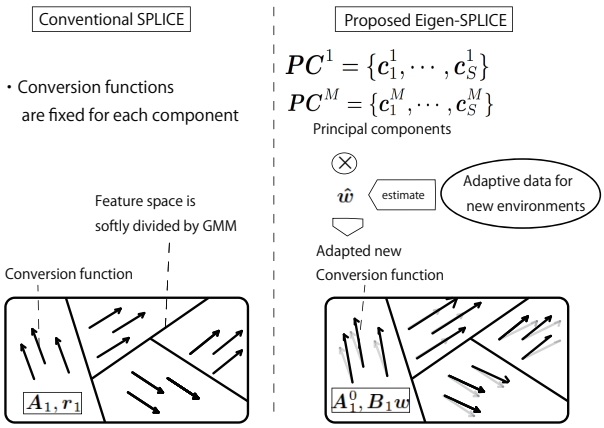

Fig. 1　Overview of Eigen-SPLICE

以上の手順によって，従来の SPLICE を少数の適応データを用いて，重みを推定することで未知の雑音環境下における変換関数を得られるようになった．この手法を Eigen–SPLICE と今後記述することにする (図 1 参照)．

3.3 擬似パラレルデータ

通常の雑音除去において，未知の雑音環境下において雑音付加音声とクリーン音声のパラレルデータを得ることは難しい．しかし未知の雑音付加音声は，学習用のクリーン音声に未知雑音を付加することで擬似的に作成可能である．具体的には，未知の雑音付加音声が得られたときに，雑音のみの区間を取り出して，それを学習データ中のクリーン音声に付加することで，擬似パラレルデータを得る．そして，この擬似パラレルデータを用いて未知環境における主成分の重み付けを推定する．

4 実験

以上のような提案手法の性能を評価するために，AURORA–2 データベースを用いて以下の実験を行った．

まず，AURORA–2 の train セット 4 タイプ (Subw., Babble, Car, Exhibit) × 4SNR(5,10,15,20[dB]) 計 16 雑音環境すべてを使って，雑音環境非依存の変換関数 $\boldsymbol{A}_s^0$ を学習する．次に $\boldsymbol{A}_s^0$ を用いて，(8) 式のように 16 雑音環境毎の $\boldsymbol{r}_s^i$ を学習する．その後，$\boldsymbol{r}_s$ スーパーベクトルを構成し，主成分分析を施し，主成分とバイアスベクトルを得る．

得られた主成分を用いて，AURORA–2 の testb セットに対して認識実験を行った．testb セットは train セットにはない種類の雑音 (Rest., Street, Airport, Sta.) が収録されている．つまり，testb セットは未知の雑音環境である．

今回は擬似パラレルデータを作る際に，入力雑音付加音声の始端と末端 250[ms] は雑音のみの区間という粗い仮定をして、雑音区間を切り出した．この切り出した雑音区間を，学習データの中からランダム

Table 1　Word recognition accuracy (a) without enhancement, baseline. (b) with the conventional enhancement. (c) with the proposed method.

(a)	Rest.	Street	Airport	Station	Avg.
20dB	90.16	94.63	86.88	88.18	89.96
15dB	71.31	84.59	66.39	69.01	72.83
10dB	44.77	59.90	40.42	42.77	46.97
5dB	10.51	31.08	11.08	15.82	17.12
0dB	-15.2	11.24	-6.26	0.00	-2.55
Avg.	40.31	56.29	39.70	43.16	44.86

(b)	Rest.	Street	Airport	Station	Avg.
20dB	99.36	98.36	99.02	98.86	98.90
15db	98.53	97.28	98.51	97.32	97.91
10dB	94.96	90.87	94.78	93.06	93.42
5dB	82.64	70.56	80.47	72.82	76.62
0dB	49.43	33.68	43.16	29.25	38.88
Avg.	84.99	78.14	83.19	78.26	81.14

(c)	Rest.	Street	Airport	Station	Avg.
20dB	99.32	98.37	99.08	99.07	98.96
15dB	98.46	97.34	98.60	97.87	98.07
10dB	95.95	92.38	96.99	94.60	94.98
5dB	85.88	77.60	85.98	80.04	82.38
0dB	58.03	45.56	56.07	41.84	50.38
Avg.	87.53	82.25	87.34	82.68	84.95

に 8 発声分選び出したクリーン音声に繰り返し重畳することで，入力発声毎に変換関数を適応する．また，用いる主成分の数は予備実験において高い性能を示した 6 とした．特徴量としては MFCC13 次元に Δ と $\Delta\Delta$ を加えた 39 次元 (HTK での MFCC_D_A_0)，雑音付加音声の特徴量の分布を表現する GMM の混合数は 64 混合とした．認識結果は testb セットのうち，SNR が 0[dB] から 20[dB] の Accuracy の平均を用いて評価した．認識用の HMM は train-set のクリーン音声のみから学習し，1 単語あたり 18 状態，1 状態あたり 20 混合の GMM を持つ単語 HMM を用いた．

実験結果を表 1 に示す．(a),(b),(c) はそれぞれ，雑音抑制なし，従来手法つまり単純な SPLICE による雑音抑制後，提案手法つまり Eigen–SPLICE による雑音抑制後の認識精度である．Eigen-SPLICE を用いて，各発声毎に変換関数に適応を施した場合，Street や Station などの未知雑音環境下において，認識精度が向上したことが分かる．これは，雑音の変動が主観的に大きいと感じる Street や Station などの雑音環境下においては，変換方法そのものを発声毎に切り替える提案手法が有効であるということが言える．testb セット全体においては，従来手法に比べ 20.2%の誤り訂正率を実現しており，提案手法が有効であるということが実験的に示された．

5　まとめ

従来手法の SPLICE では入力音声の雑音環境と変換関数を学習した環境が似ていることを暗に仮定しており，未知の雑音環境下においては十分な性能を発揮することが保証されていない．そこで本報告では未知の雑音環境に対して変換自体を適応することを試みた．変換方法を適応するためには，本来であれば多数のパラメータを推定する必要がある．それを避けるため，提案手法では変換を表すベクトルに主成分分析を施すことで推定すべきパラメータを削減した．これにより，適応に必要なデータを削減することができる．また，適応するのに必要なパラレルデータを，入力音声の雑音のみの区間を切り出すことによって擬似的に作成することを提案した．そして，提案手法が有効であることを実験的に示した．

以上のような本研究にも，未だいくつかの問題がある．まず，今回ベースラインとした従来の SPLICE の GMM の混合数を計算量を抑えるために低くしたため，本来の従来手法の性能を十分に発揮出来ていないということがある．今後は，より混合数を増やした状況においても，従来手法以上の性能を発揮できるように取り組んでいく．また，EMS などの他手法との性能比較もする予定である．

参考文献

[1] 猿渡洋, 信学技報, SP2010-103, pp.1–6, 2011.
[2] Hyvarinen A., *et al.*, “詳解 独立成分分析”, 東京電機大学出版局, 2005.
[3] Boll S., *IEEE Trans. Acoust. Speech, Signal prosecc.,* Vol.27, No.2, pp.113–120, 1979.
[4] Droppo J., *et al.*, *EUROSPEECH-2001*, pp.217–220, 2001.
[5] Stouten V., *KU Leuven, Ph.D. Thesis,* 2006.
[6] Pearce D., *et al.*, *Proc. of ICSLP ’00*, Vol.4, pp.29–32, 2000.
[7] Droppo J., *et al.*, *Proc. of ICASSP ’01*, Vol.1, pp.209–212, 2001.
[8] Chen K., *et al.*, *Proc. of ICSLP*, Vol.3 pp.742–754, 2000.
[9] 戸田智基, *et al.*, 信学技報, SP2006-39, pp.25–30, 2006.