Deep Learningに基づくクリーン音声状態識別による雑音環境下音声認識*

☆柏木陽佑, 齋藤大輔, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

近年，クリーン環境では自動音声認識の精度が大きく向上し実用段階に入ってきている．しかし，ノイジー環境では未だにその性能は十分だとは言えない．学習時と利用時において様々な環境ミスマッチが生じるが，耐雑音性を高めることは非常に重要な課題である [1].

特徴量強調は耐雑音性を確保するためのフロントエンド処理であり，SPLICE [2] や Denoising AutoEncoder (DAE) [3]，その他にも各種手法 [4–6] が提案されている．これらは，ノイジー音声特徴量からクリーン音声特徴量を推定することによりミスマッチを低減する．

SPLICE は 2 つの段階に分けることができる．まず，観測ノイジー音声特徴量空間をガウス混合モデル (Gaussian Mixture Model; GMM) でモデル化し，入力フレームに対する，GMM コンポーネントの事後確率を計算する．次に，計算した事後確率を重みとして線形変換の足し合わせでクリーン音声特徴量を推定する．しかし，ノイジー音声特徴量領域の分割とクリーン音声特徴量領域における分割は一致しないことが多く，局所領域での線形性の仮定が正しいとは限らない．また，ノイジー音声特徴量領域の分割は雑音の種類に大きく依存することも問題である．

そこで，クリーン音声特徴量領域で分割を行う手法として REgularized piecewise linear mapping with DIscriminative region weighting And Long-span features (REDIAL) が提案されている [7–9]．REDIAL は GMM と線形判別分析 (Linear Discriminant Analysis; LDA) によりノイジー音声特徴量からクリーン音声状態を識別的に推定する．REDIAL は特に HMM を雑音音声込みで学習する multi-condition で高い認識性能を示すことが実験的に示されている．しかし，LDA は線形変換であるため，ノイジー音声特徴量とクリーン音声状態の複雑な関係を適切にモデル化できているとは言えない．

一方，DAE はニューラルネットによって観測ノイジー音声特徴量からクリーン音声特徴量を非線形かつ直接的に推定する．DAE を多層にした Deep Denoising AutoEncoder (DDAE) は特にノイズクローズド環境で高い精度でクリーン音声を推定することが報告されている．しかし，ノイズオープン環境では，過学習の影響で大きく性能が低下する．

そこで，本提案手法は Deep Neural Network (DNN) [10] をクリーン音声状態の事後確率推定にのみ用い，区分的線形変換によってクリーン音声特徴量を推定する．まず，クリーン音声特徴量空間を GMM でモデル化し，クリーン音声状態ラベルを得る．クリーン音声とノイジー音声は時間対応の取れているパラレルデータであるため，このクリーン音声状態ラベルを DNN により観測ノイジー音声特徴量より推定を行い，クリーン音声状態に対する事後確率を得る．その後，REDIAL と同様に，観測ノイジー音声特徴量から重み付きの線形変換によりクリーン音声特徴量を推定する．

本稿の構成を示す．まず，2 節で提案手法の定式化を行う．3 節で従来手法との比較を行い，4 節で実験を示す．最後に 5 節でまとめる．

2 定式化

特徴量強調は雑音環境下での音声認識精度を向上させるため観測されたノイジー音声特徴量から対応するクリーン音声特徴量を推定する．本研究では，観測特徴量が得られた際のクリーン音声特徴量状態に対する事後確率を DNN を用いて推定し，クリーン音声特徴量を推定した事後確率を重み付けとした区分的線形変換により求める．

DNN を採用する理由は，従来の区分的線形変換手法は常に変換が線形で表現できるような領域に分割できているとは限らないためである．例えば，SPLICE はノイジー音声特徴量空間を GMM でモデル化し，ノイジー音声特徴量の状態に対する事後確率を重みとして用いる．しかし，この重みの代わりに，クリーン音声特徴量空間を GMM でモデル化して正解のクリーン音声特徴量状態を用いて推定した事後確率を利用すると，高い精度でクリーン音声特徴量が推定することができる．オラクルの結果と比較して SPLICE の結果が大きく低下するのは，クリーン音声特徴量状態とノイジー音声特徴量状態のミスマッチの影響だと考えられる．REDIAL は LDA によってこの影響を取り除くことを行っているが，LDA はあくまで線形の写像なため限界がある．それに対して DNN は非線形変換であるため，より正確にクリーン音声特徴量状態を推定することができると考えられる．時刻 t における N 次元のクリーン音声特徴量 $\boldsymbol{x}_t$ とノイジー音声特徴量 $\boldsymbol{y}_t$ のパラレルデータ $\{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}$ を考える．Fig. 1 に学習段階の流れを示す．まず，クリーン音声特徴量の確率密度関数 $p(\boldsymbol{x})$ を GMM で学習する．

$$p(\boldsymbol{x}_t) = \sum_k p(k)p(\boldsymbol{x}_t|k) \quad (1)$$

$$p(\boldsymbol{x}_t|k) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{\boldsymbol{x}}, \boldsymbol{\sigma}_k^{\boldsymbol{x}}) \quad (2)$$

$$p(k) = \pi_k^{\boldsymbol{x}} \quad (3)$$

これを用いて，GMM のコンポーネントインデクスに対する事後確率を，

$$p(k|\boldsymbol{x}_t) = \frac{p(\boldsymbol{x}_t|k)p(k)}{\sum_{k'} p(\boldsymbol{x}_t|k')p(k')} \quad (4)$$

と表すことができる．

一方，評価段階ではクリーン音声特徴量 $\boldsymbol{x}_t$ は観測できないため観測されたノイジー音声特徴量 $\boldsymbol{y}_t$ から

*Piecewise linear transformation based on discriminative approach using deep learning and its use for noise robust automatic speech recognition by Y.Kashiwagi D.Saito, N.Minematsu, and K.Hirose (The University of Tokyo)

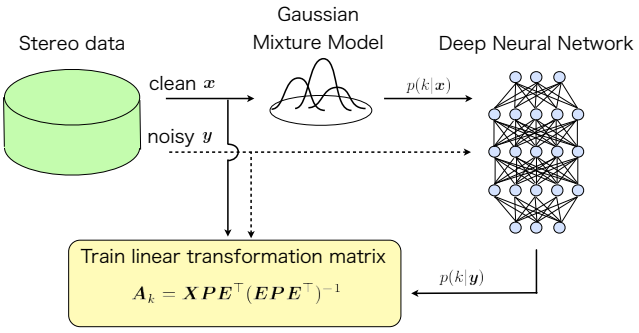

Fig. 1 The training phrase of our proposed method.

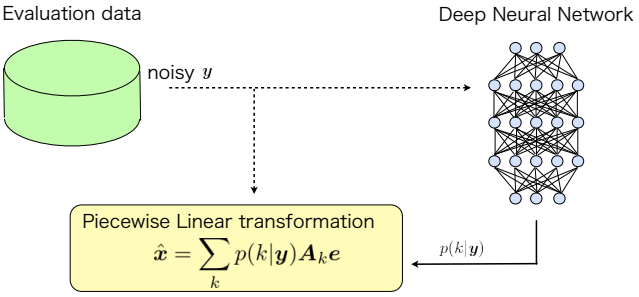

Fig. 2 The testing phase of our proposed method.

クリーン音声状態に対する事後確率 $p(k|\boldsymbol{y}_t)$ を推定する必要がある (Fig. 2)．そこで，識別モデルとして DNN をクリーン音声特徴量状態と観測特徴量を用いて学習する．

$$p(k|\boldsymbol{y}_t) \simeq p(k|\boldsymbol{d}_t) = \mathrm{softmax}_k(\boldsymbol{V}\boldsymbol{h}^{(n)}(\boldsymbol{d}_t) + \boldsymbol{c}) \quad (5)$$

$$\boldsymbol{h}^{(n)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(n)}\boldsymbol{h}^{(n-1)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(n)}) \quad (6)$$

$$\boldsymbol{h}^{(1)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(1)}\boldsymbol{d}_t + \boldsymbol{b}^{(1)}) \quad (7)$$

ここで，$\boldsymbol{d}_t$ は前後フレームを連結した入力特徴量ベクトルであり，

$$\boldsymbol{d}_t = [\boldsymbol{y}_{t-s}^\top, \ldots, \boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_t^\top, \boldsymbol{y}_{t+1}^\top, \ldots, \boldsymbol{y}_{t+s}^\top]^\top \quad (8)$$

となる．また，$\boldsymbol{\sigma}$ はベクターシグモイド関数であり，$\boldsymbol{V}$，$\boldsymbol{W}^{(n)}$ と $\boldsymbol{c}$，$\boldsymbol{b}^{(n)}$ はニューラルネットの重みとバイアスのパラメータ，$\boldsymbol{h}^{(n)}(\boldsymbol{y})$ は n 番目の隠れ層の出力ベクトルである．なお，事前学習として各層の初期値を Restricted Boltzmann Machine (RBM) で求める [10]．以上により，観測特徴量からクリーン音声特徴量状態インデクスに対する事後確率 $p(k|\boldsymbol{y}_t)$ を推定することが可能となる．

評価段階では，クリーン音声特徴量 $\boldsymbol{x}_t$ を事後確率 $p(k|\boldsymbol{y}_t)$ を重みとして用いる区分的線形変換によって推定する．

$$\hat{\boldsymbol{x}}_t = \sum_k p(k|\boldsymbol{y}_t)\boldsymbol{A}_k\boldsymbol{e}_t \quad (9)$$

ここで，$\boldsymbol{A}_k$ は，コンポーネント k における線形変換行列であり，$\boldsymbol{e}_t$ は前後フレームを連結した拡張行列

Table 1 Performance gap between ordinary SPLICE and the oracle SPLICE. (Word Error Rate (%))

	SPLICE	SPLICE (Oracle)
clean	0.57	0.69
SNR 20	1.08	0.76
SNR 15	1.99	0.70
SNR 10	4.65	0.77
SNR 5	16.76	0.93
SNR 0	49.96	0.95
SNR -5	81.46	1.13
Average	14.89	0.82

である．

$$\boldsymbol{e}_t = [1, \boldsymbol{y}_{t-u}^\top, \ldots, \boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_t^\top, \boldsymbol{y}_{t+1}^\top, \ldots, \boldsymbol{y}_{t+u}^\top]^\top \quad (10)$$

なお，$\boldsymbol{A}_k$ は重み付き最小二乗誤差基準で学習する．

$$\hat{\boldsymbol{A}}_k = \underset{\boldsymbol{A}_k}{\mathrm{argmin}} \sum_j p(k|\boldsymbol{y}_j)||\boldsymbol{x}_j - \boldsymbol{A}_k\boldsymbol{e}_j||^2 \quad (11)$$

これは解析解を得ることができる．$\boldsymbol{A}_k$ は，

$$\hat{\boldsymbol{A}}_k = \boldsymbol{X}\boldsymbol{P}\boldsymbol{E}^\top(\boldsymbol{E}\boldsymbol{P}\boldsymbol{E}^\top)^{-1} \quad (12)$$

として計算することができる．ここで，$\boldsymbol{X} \in \mathcal{R}^{N\times T}$ と $\boldsymbol{E} \in \mathcal{R}^{(N(2u+1)+1)\times T}$ はそれぞれ出力と入力特徴量の拡張ベクトルを並べたデータ行列，$\boldsymbol{P} \in \mathcal{R}^{T\times T}$ は $p(k|\boldsymbol{y}_t)$ を対角に並べた行列である．

3 従来の特徴量強調手法との比較

本節では，提案手法と区分的線形変換手法である SPLICE とその拡張である REDIAL との比較を述べる．また，ニューラルネットを用いた代表的な雑音抑圧手法である DAE との比較も行う．

3.1 SPLICE

SPLICE は特徴量強調手法の 1 つであり，区分的線形変換によってノイジー音声特徴量からクリーン音声特徴量を推定する．

$$\hat{\boldsymbol{x}}_t = \sum_i p(i|\boldsymbol{y}_t)\boldsymbol{A}_i \begin{bmatrix} 1 \\ \boldsymbol{y}_t \end{bmatrix} \quad (13)$$

SPLICE では，ノイジー音声特徴量の確率密度関数 $p(\boldsymbol{y})$ を GMM でモデル化する．

$$p(\boldsymbol{y}_t) = \sum_i \pi_i^{\boldsymbol{y}}\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_i^{\boldsymbol{y}}, \boldsymbol{\sigma}_i^{\boldsymbol{y}}) \quad (14)$$

これを用いて，ノイジー音声特徴量状態に対する事後確率 $p(i|\boldsymbol{y}_t)$ を

$$p(i|\boldsymbol{y}_t) = \frac{p(\boldsymbol{y}_t|i)p(i)}{\sum_{i'} p(\boldsymbol{y}_t|i')p(i')} \quad (15)$$

として計算する．式 (14) のように，SPLICE ではノイジー音声のみを用いて領域分割の学習を行う．しか

し，これは学習データの雑音環境に過学習してしまうことが予想され，クリーン音声特徴量空間でモデル化を行うべきだと考えられる．

そこで，ノイジー音声特徴量領域とクリーン音声特徴量領域のそれぞれで GMM によるモデル化と重み付けを行った場合の比較を行った．Table 1 は Aurora 2 のテストセット A（ノイズクローズド環境）を用いた認識結果である．ここで，クリーン音声特徴量領域で重み付けを行ったオラクルの SPLICE は，

$$\hat{\boldsymbol{x}}_t = \sum_{i^*} p(i^*|\boldsymbol{x}_t)\boldsymbol{A}_{i^*}\begin{bmatrix}1\\ \boldsymbol{y}_t\end{bmatrix} \tag{16}$$

$$p(i^*|\boldsymbol{x}_t) = \frac{p(\boldsymbol{x}_t|i^*)p(i^*)}{\sum_{i^{*\prime}} p(\boldsymbol{x}_t|i^{*\prime})p(i^{*\prime})} \tag{17}$$

としてクリーン音声状態 i^* に対する事後確率 $p(i^*|\boldsymbol{x}_t)$ を計算する．オラクルの結果は雑音の大きな環境でも頑健にクリーン音声を推定することができており，これは事後確率 $p(i^*|\boldsymbol{y}_t)$ を精確に推定することができていることに加え，クリーン音声ドメインにおける分割がより線形変換を仮定することができるためだと考えられる．なお，NMN-SPLICE は雑音の平均成分を除去することで分割をクリーン音声特徴量領域に近づくことが期待されている [2]．

3.2 REDIAL

REDIAL は SPLICE を拡張したものであり，クリーン音声特徴量に近づけた空間で分割を行う．学習段階では，まずクリーン音声の状態をラベルとした LDA により観測ノイジー特徴量を次元圧縮し，ノイジー特徴量をクリーン音声状態の識別が容易な特徴量へ次元圧縮する．LDA の次元圧縮行列 $\boldsymbol{L}$ を対応するクリーン音声特徴量の状態インデクスをソフトラベルとして学習する．次に，LDA により次元圧縮を行ったベクトル $\boldsymbol{v}_j = \boldsymbol{L}\boldsymbol{d}_j$ を用いて K^* 混合の GMM を学習する．

$$p(\boldsymbol{v}_t) = \sum_{k^*=1}^{K^*} \pi_{k^*}^{\boldsymbol{v}}\mathcal{N}(\boldsymbol{v}_t;\boldsymbol{\mu}_{k^*}^{\boldsymbol{v}},\boldsymbol{\sigma}_{k^*}^{\boldsymbol{v}}) \tag{18}$$

入力特徴量が得られた際のクリーン音声特徴量状態インデクスに対する事後確率 $p(k^*|\boldsymbol{y}_t)$ を

$$p(k^*|\boldsymbol{y}_t) \simeq p(k^*|\boldsymbol{v}_t) = \frac{p(\boldsymbol{v}_t|k^*)p(k^*)}{\sum_{k^{*\prime}} p(\boldsymbol{v}_t|k^{*\prime})p(k^{*\prime})} \tag{19}$$

として計算する．最終的に得られた事後確率を重みとして用いた区分的線形変換によりクリーン音声特徴量を

$$\hat{\boldsymbol{x}}_t = \sum_{k^*} p(k^*|\boldsymbol{y}_t)\boldsymbol{A}_{k^*}\boldsymbol{e}_t \tag{20}$$

$$\boldsymbol{e}_t = [1, \boldsymbol{y}_{t-u}^\top, \ldots, \boldsymbol{y}_{t-1}^\top, \boldsymbol{y}_t^\top, \boldsymbol{y}_{t+1}^\top, \ldots, \boldsymbol{y}_{t+u}^\top]^\top \tag{21}$$

として推定する．なお，$\boldsymbol{A}_{k^*}$ は非常に大きな行列になるため，学習の際に正則化を導入する．

提案手法と REDIAL は，ノイジー音声特徴量からクリーン音声特徴量状態を推定する部分である．しかし，LDA は線形変換であるため，クリーン音声特徴量状態とノイジー音声特徴量との間の複雑な関係を適切にモデル化できているとは言えない．

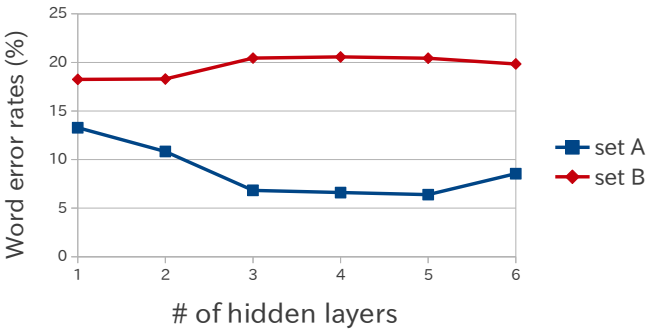

Fig. 3 The performance of DDAE with different numbers of hidden layers in Aurora 2 dataset.

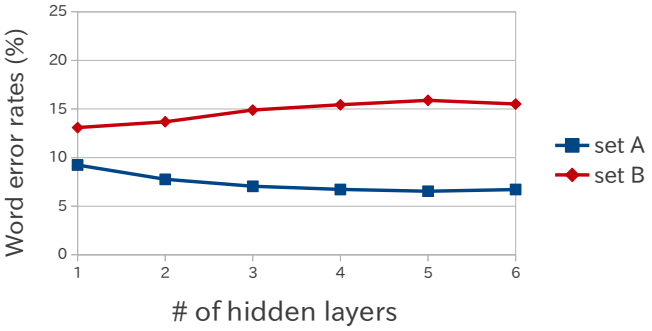

Fig. 4 The performance of the proposed method with different numbers of hidden layers in Aurora 2 dataset.

3.3 DAE

DAE はニューラルネットを用いてノイジー音声特徴量からクリーン音声特徴量を直接的に推定する手法である．DDAE は DAE を多層にしたものであり，クリーン音声特徴量を

$$\hat{\boldsymbol{x}}_t = \boldsymbol{U}\boldsymbol{h}^{(n)}(\boldsymbol{d}_t) + \boldsymbol{c} \tag{22}$$

$$\boldsymbol{h}^{(n)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(n)}\boldsymbol{h}^{(n-1)}(\boldsymbol{d}_t) + \boldsymbol{b}^{(n)}) \tag{23}$$

$$\boldsymbol{h}^{(1)}(\boldsymbol{d}_t) = \boldsymbol{\sigma}(\boldsymbol{W}^{(1)}\boldsymbol{d}_t + \boldsymbol{b}^{(1)}) \tag{24}$$

として推定する．ここで，$\boldsymbol{U}$ は重み行列である．本稿では，RBM により各層の初期値を決め，最小二乗誤差基準によるバックプロパゲーションにより学習した DDAE との比較を行う．Fig. 3 は層の数を変化した際の Aurora 2 における単語認識誤り率をプロットしたものである．セット A とセット B はそれぞれノイズクローズド環境とノイズオープン環境であり，各隠れ層のノード数は 1024 で統一している．結果によると，ノイズクローズド環境では層の数を増やすと誤り率が減るが，ノイズオープン環境ではほぼ変化しない．これはニューラルネットが過学習しているためだと考えられる．なお，近年リカレント構造を持った複雑なニューラルネットを用いたオートエンコーダも提案されている [11]．

4 実験

提案手法の有効性を Aurora 2 による連続数字読み上げ認識実験によって示す．データはいくつかの環境の雑音が重畳されたノイジー音声とそれに対応

Table 2　Performance comparison among our proposal and conventional methods (word error rates %).

	clean condition (WER. %)				multi condition (WER. %)			
	set A	set B	set C	average	set A	set B	set C	average
Baseline	48.93	55.80	39.23	47.98	10.57	11.89	14.33	12.27
SPLICE	14.89	19.31	21.59	18.60	9.20	14.50	15.22	12.97
REDIAL	16.70	20.59	21.14	19.48	8.98	**13.26**	**12.45**	11.56
DDAE	**6.39**	20.44	17.20	14.68	5.97	18.50	14.67	13.04
PROPOSED	7.04	**14.93**	**15.54**	**12.51**	**5.64**	15.20	13.29	**11.38**

するクリーン音声が含まれており，パラレルデータを利用することができる．認識に用いる音響モデルは HMM を利用し，クリーン音声のみで学習したもの (clean condition) と雑音音声込みで学習したもの (multi condition) の二通りで比較した．

評価データは 3 種類 (A,B,C) があり，セット A は学習時と同じノイズクローズド環境，セット B は学習時と異なるノイズオープン環境，セット C はチャネルオープン環境である．特徴量として MFCC とパワー，その 1 次，2 次微分を用いた．DNN の学習は KALDI [13] を利用し，入力に用いる特徴量は全て前後 3 フレームの計 7 フレームに統一した．また，各層のノード数は 1024 に設定し，推定するクリーン音声特徴量の状態は 1024 に設定した．DNN のバックプロパゲーションにおいては，学習データ 8,440 発話のうち 844 発話を開発セットとした．また，REDIAL と同様に線形変換の学習では正則化を導入している．

まず，DNN の層の数による認識率の違いを Fig. 4 に示す．層の数を増やすとノイズクローズド環境ではエラーレートが低下するが，ノイズオープン環境では DDAE と同様に効果が表れない．これは DDAE と同様にニューラルネットが過学習しているためだと考えられる．

次に，従来手法との比較を行う．比較を行うのは SPLICE，REDIAL，DDAE である．SPLICE のノイジー音声状態数は 1024 に設定した．REDIAL はクリーン音声状態数は 1024 に設定し，LDA による次元圧縮後の特徴量ベクトルは 64 次元にした．また，DDAE は中間層の数を 5 とした．

Table 2 に結果を示す．clean condition では提案手法は非常に良い結果が得られている．一方，multi condition では REDIAL とそれほど差が出ていない．これは LDA の線形変換という制約が逆に GMM の形状を保持できるためだと考えられる．また，DDAE と比較した場合，特にノイズオープン環境で大きな差が生じている．これは DDAE が過学習してしまっているためであり，GMM で分散を考慮したクラスラベルの識別を行う提案手法が過学習の影響を低下することができているからだと考えられる．

5　まとめ

本稿では，パラレルデータを用いた新しい特徴量強調手法を提案した．提案手法は，クリーン音声特徴量を GMM でクラスタリングを行い，DNN によってクリーン音声状態をノイジー音声特徴量から推定する．その後，この事後確率を重みとして区分的線形変換によりクリーン音声特徴量を推定する．Aurora 2 を用いた連続数字読み上げ認識タスクによる評価で提案手法は特にノイズオープン環境で DDAE より低いエラーレートを示すことができた．今後，提案手法を初期値として全体をニューラルネット形式で表現して全体を最適化することが考えられる．

参考文献

[1] Gales, Mark JF, “Model-based approaches to handling uncertainty,” Robust Speech Recognition of Uncertain or Missing Data-Theory and Applications. Springer, Berlin, Germany, pp. 101–125, 2011.

[2] Droppo, Jasha and Deng, Li and Acero, Alex, “Evaluation of SPLICE on the Aurora 2 and 3 tasks,” International Conference on Spoken Language Processing, pp. 29–32, 2002.

[3] Vincent, Pascal and Larochelle, Hugo and Bengio, Yoshua and Manzagol, Pierre-Antoine, “Extracting and composing robust features with denoising autoencoders,” Proceedings of the 25th international conference on Machine learning, pp. 1096–1103, 2008.

[4] Li, Jinyu and Seltzer, Michael L and Gong, Yifan, “Improvements to VTS feature enhancement,” Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4677–4680, 2012.

[5] Afify, Mohamed and Cui, Xiaodong and Gao, Yuqing, “Stereo-based stochastic mapping for robust speech recognition,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 17, no. 7, pp. 1325–1334, 2009.

[6] Gemmeke, Jort F and Virtanen, Tuomas and Hurmalainen, Antti, “Exemplar-based sparse representations for noise robust automatic speech recognition,” Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 7, pp. 2067–2080, 2011.

[7] Senior, Andrew and Cho, Youngmin and Weston, Jason “Learning improved linear transforms for speech recognition.,” Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 1957–1960, 2012.

[8] Suzuki, Masayuki and Yoshioka, Takuya and Watanabe, Shinji and Minematsu, Nobuaki and Hirose, Keikichi “MFCC enhancement using joint corrupted and noise feature space for highly non-stationary noise environments,” Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, pp. 4109–4112, 2012.

[9] Suzuki Masayuki, Yoshioka Takuya, Watanabe Shinji, Minematsu Nobuaki, and Hirose Keikichi “Feature Enhancement With Joint Use of Consecutive Corrupted and Noise Feature Vectors With Discriminative Region Weighting,” IEEE TASLP, (to appear).

[10] Hinton, Geoffrey E and Osindero, Simon and Teh, Yee-Whye “A fast learning algorithm for deep belief nets,” Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.

[11] Maas, Andrew L and Le, Quoc V and O’Neil, Tyler M and Vinyals, Oriol and Nguyen, Patrick and Ng, Andrew Y, “Recurrent Neural Networks for Noise Reduction in Robust ASR,” INTERSPEECH, 2012.

[12] http://htk.eng.cam.ac.uk/

[13] Povey, Daniel and Ghoshal, Arnab and Boulianne, Gilles and Burget, Lukas and Glembek, Ondrej and Goel, Nagendra and Hannemann, Mirko and Motlicek, Petr and Qian, Yanmin and Schwarz, Petr and others “The kaldi speech recognition toolkit,” IEEE 2011 workshop on automatic speech recognition and understanding, 2011.