SPLICE を用いた雑音抑圧手法の統合*

☆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なミスマッチがある場合，音声認識システムの性能は大きく低下してしまう．例えば背景雑音，録音機器の特性，部屋の残響の違いなどによって音響的ミスマッチが大きくなる．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となる．

ミスマッチを軽減するのに有効な手法には，VAD と二段階のウィーナーフィルタによって音声信号からから雑音成分を取り除く Advanced Front-End (AFE) [1, 2, 3] という手法がある．また雑音による音声の歪みをベクトルテーラー展開で近似する Vector Taylor Series (VTS) [4] や，雑音重畳音声からクリーンな音声への変換を区分的線形変換で近似する Stereo-based Piecewise Linear Compensation for Environments (SPLICE) [5] などの特徴量強調手法も挙げられる．他にも特徴量のヒストグラムを正規化することにより雑音の影響を取り除く Histogram Equalization (HEQ) [6, 7] などの正規化手法も有効である．

しかしこれらの手法の性能は雑音環境に依存しており，種々の雑音を広範囲の SNR に渡って抑圧することは非常に困難である．そのためより広範囲の雑音環境で認識性能を向上させるためには，これらの手法を統合することが必要であると考えられる．これまで複数の手法の統合は，逐次的に複数の雑音抑圧手法を適用した特徴量で認識する方法 [8, 9] と，各雑音抑圧手法の複数の仮説から信頼度による投票を行い最終的な認識結果を導く方法 [10] が提案されている．本稿では従来の方法とは異なる SPLICE を用いた雑音抑圧手法の統合を提案する．提案手法では各雑音抑圧手法を適用した特徴量を複数結合したベクトルから，SPLICE を用いてクリーンな特徴量を推定し音声認識を行う．提案法の有効性を評価するため AURORA-2 データベース [11] を用いて音声認識実験を行った．

2 従来の統合法

2.1 逐次的に雑音抑圧手法を適用 [8, 9]

この統合法では，雑音重畳音声に逐次的に雑音抑圧手法を適用して，より雑音にロバストな特徴量を抽出し音声認識を行う．複数の抑圧手法を適用する順番は多数考えることができ，その中から性能の高いものを選ぶことにより高品質な音声認識を行うことができる．

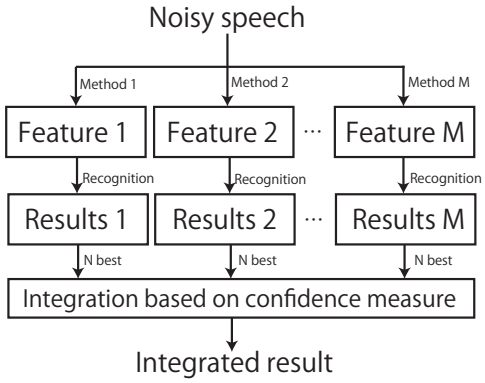

Fig. 1 Process flow of integration based on confidence measure

文献 [9] では AFE，SPLICE，HEQ をどのような順番で適用した特徴量が AURORA-2 データベースにおいて有効か検討し，AFE-SPLICE-HEQ の順に適用した特徴量が最も高い認識率が示した．

しかし他の雑音環境においても AFE-SPLICE-HEQ が最適な順番であるとは限らず，全ての順番の組み合わせについて認識を行うことは計算量の増加につながる．また逐次的に雑音抑圧手法を適用しているため，原理的に適用できない順序が生じてしまう．たとえば，AFE は音声波形を入力しケプストラムを出力するため，AFE と他のスペクトル領域での雑音抑圧手法をこの順序で適用することはできない．

2.2 信頼度を用いた統合

複数の手法を統合する方法として，複数の仮説の信頼度を計算して信頼度が高いものを統合結果として出力する方法が提案されている [10]．この統合法は Figure 1 のような流れ図で表される．

この統合法では M 個の雑音抑圧手法から得られる特徴量を用いてそれぞれ認識を行い，N-best を出力するので全体で MN 個の仮説を得る．次に雑音抑圧手法 m の順位 n である仮説 $U(m,n)$ のフレーム正規化対数尤度を $l(m,n)$ として，各仮説の信頼度 $S(m,n)$ を計算する．信頼度 $S(m,n)$ は以下のように定義さ

*Integration of Various Noise Suppression Methods based on SPLICE by T. Kai, M. Suzuki, N. Minematsu, K. Hirose (The University of Tokyo)

れる．

$$S(m,n) = s(m,n) \times \{s(m,1) - s(m,N)\} \quad (1)$$

$$s(m,n) = \frac{\exp\{l(m,n))\}}{\sum_{i=1}^{N} \exp\{l(m,i)\}} \quad (2)$$

この (2) 式の $s(m,n)$ は雑音抑圧手法の間で尤度を正規化したものに相当する．さらに (1) 式で，各雑音抑圧手法の中で順位による重み付けを行なっている．これにより，上位の仮説と下位の仮説の尤度の差が大きければ上位の仮説が重視され，逆に差が小さければ相対的に下位の仮説が重視されることになる．

MN 個の仮説の中には同じ認識結果を持つ仮説が複数存在することもあるので，ある認識結果 U の信頼度 $Score(U)$ を以下のように計算する．

$$Score(U) = \sum_{m,n\ :\ U(m,n)=U} S(m,n) \quad (3)$$

このように計算された信頼度 $Score(U)$ の中で信頼度が最も大きい認識結果が最終的な統合結果となる．これにより，ある手法では真の認識結果の順位が低くなっていても，他の手法で高順位であれば全体としての信頼度が大きくなり，真の認識結果を救うことができる．AURORA-2J データベースを用いた実験によって，特に multicondition training において大きな改善が得られたと報告されている．

しかし複数個の音響モデルで並列に音声認識を行わなければならないため計算コストが高いことが欠点と言える．また数字発声のような簡単なタスクであれば MN 個の仮説の中で，同じ認識結果が複数回登場する可能性が高い．しかしタスクが難しくなり語彙が大きくなると，MN 個の候補の中に同じ認識結果が登場しにくくなるため統合による性能向上が見られなくなると考えられる．

3　SPLICE による特徴量強調

クリーン音声の特徴量を $\boldsymbol{x}$，雑音重畳音声の特徴量を $\boldsymbol{y}$ とする．SPLICE は $\boldsymbol{y}$ から $\boldsymbol{x}$ への非線形な変換を以下のような区分的線形変換によって近似する．

$$\hat{\boldsymbol{x}} = \sum_{k} p(k|\boldsymbol{y}) \boldsymbol{A}_k \boldsymbol{y}' \quad (4)$$

この式の概念図を Figure 2 に示す．k は部分空間のインデックス，$\boldsymbol{A}_k$ は線形変換行列，$\boldsymbol{y}' = [1 \quad \boldsymbol{y}^\top]^\top$ を表す拡張特徴量ベクトルである．$\boldsymbol{A}_k$ は一定であるが，部分空間の重みである $p(k|\boldsymbol{y})$ が $\boldsymbol{y}$ に応じて変動するため，全体として非線形変換を表現することができる．

この式で特徴量強調をするためには，時間的に同期のとれたステレオデータ $\{\boldsymbol{x}_i\}, \{\boldsymbol{y}_i\}$ を学習データとして，どのように空間を分割するかを表す $p(k|\boldsymbol{y})$ と，線形変換 $\boldsymbol{A}_k$ を学習する必要がある．まず $\boldsymbol{y}_i$ の確率密度関数が GMM に従うと仮定して以下のように学習する．

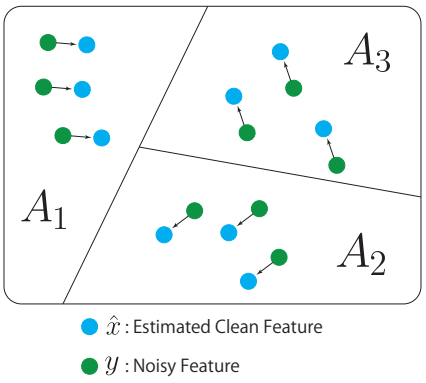

Fig. 2　Conceptual image of SPLICE

$$p(k) = \pi_k \quad (5)$$

$$p(\boldsymbol{y}_i|k) = \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (6)$$

ただし，$\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ はそれぞれ k 番目のインデックスに対応する正規分布の重み，平均，分散である．これにより $p(k|\boldsymbol{y}_i)$ は以下のように計算できる．

$$p(k|\boldsymbol{y}_i) = \frac{p(k)p(\boldsymbol{y}_i|k)}{p(\boldsymbol{y}_i)} \quad (7)$$

$$= \frac{\pi_k \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \quad (8)$$

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．

$$\boldsymbol{A}_k = \underset{\boldsymbol{A}_k}{\operatorname{argmin}} \sum_i p(k|\boldsymbol{y}_i) ||\boldsymbol{x}_i - \boldsymbol{A}_k \boldsymbol{y}_i'||^2 \quad (9)$$

これを解くと，

$$\boldsymbol{A}_k = \boldsymbol{X}\boldsymbol{P}_k\boldsymbol{Y}^\top(\boldsymbol{Y}\boldsymbol{P}_k\boldsymbol{Y}^\top)^{-1} \quad (10)$$

となる．ただし $\boldsymbol{X}, \boldsymbol{Y}$ は，それぞれ $\boldsymbol{x}_i, \boldsymbol{y}_i'$ を順番に並べた行列，$\boldsymbol{P}_k$ は $p(k|\boldsymbol{y}_i)$ を順番に並べたものを対角成分に持つ行列である．

このように $\boldsymbol{A}_k$ と $p(k|\boldsymbol{y})$ を事前に学習できれば，実際に特徴量を強調する際は式 (4) 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，高性能な特徴量強調が可能となる．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 の性能は低下してしまう．

4　SPLICE を用いた統合

通常の SPLICE は雑音重畳音声 $\boldsymbol{y}$ からクリーン音声 $\boldsymbol{x}$ の変換を近似する．しかし SPLICE に入力する特徴量に関して制約がないため，パラレルデータさえ

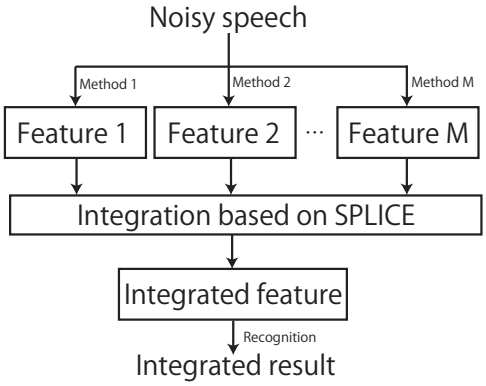

Fig. 3 Process flow of integration based on SPLICE

用意できれば $\boldsymbol{y}$ でなくとも良い．そのため，複数の特徴量を結合したベクトルを入力して，クリーン音声 $\boldsymbol{x}$ を推定する SPLICE を考えることができる．これを利用して，複数の特徴量から 1 つの統合された特徴量を導くことができる．実際の認識は最終的に統合された特徴量で 1 回行うのみなので，計算コストを低く抑えることができる．

今回提案する SPLICE を用いた統合は Figure 3 のような流れ図で表される．$\boldsymbol{y}$ に雑音抑圧手法 m をかけた特徴量を $\hat{\boldsymbol{y}}^{(m)}$，それらを結合したベクトルを $\boldsymbol{z}$ とすると，統合された特徴量 $\hat{\boldsymbol{x}}$ は以下のように求められます．

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

$$\text{where} \quad \boldsymbol{z}' = [1 \;\; \hat{\boldsymbol{y}}^{(1)\top} \;\; \hat{\boldsymbol{y}}^{(2)\top} \ldots \;\; \hat{\boldsymbol{y}}^{(M)\top}]^\top \tag{12}$$

$\boldsymbol{A}_k$ は前節と同様に，重み付き最小二乗誤差基準で学習すると

$$\boldsymbol{A}_k = \boldsymbol{X}\boldsymbol{P}_k\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{P}_k\boldsymbol{Z}^\top)^{-1} \tag{13}$$

となる．ただし，$\boldsymbol{X}, \boldsymbol{Z}$ はパラレル学習データである $\boldsymbol{x}_i, \boldsymbol{z}'_i$ を順番に並べた行列，$\boldsymbol{P}_k$ は $p(k|\boldsymbol{y}_i)$ を順番に並べたものを対角成分に持つ行列である．しかしこの $\boldsymbol{A}_k$ は（$\boldsymbol{x}$ の次元数）×（$\boldsymbol{z}'$ の次元数）の行列であり，前節の $\boldsymbol{A}_k$ と学習データ数は変わらないにも関わらずパラメータ数が非常に大きくなるため，過学習の問題が起きる．そこで重み付き最小二乗誤差基準に二次の正則化項を加えることにより過学習を避けて学習する．

$$\boldsymbol{A}_k = \underset{\boldsymbol{A}_k}{\operatorname{argmin}} \sum_i p(k|\boldsymbol{y}_i)||\boldsymbol{x}_i - \boldsymbol{A}_k\boldsymbol{y}'_i||^2 + \lambda \boldsymbol{1}^\top \boldsymbol{A}_k^\top \boldsymbol{A}_k \boldsymbol{1} \tag{14}$$

これを解くと $\boldsymbol{A}_k$ は以下のようになる．

$$\boldsymbol{A}_k = \boldsymbol{X}\boldsymbol{P}_k\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{P}_k\boldsymbol{Z}^\top - \lambda\boldsymbol{I})^{-1} \tag{15}$$

ただし，$\boldsymbol{1}$ は $\boldsymbol{z}'$ と同じ次元を持ち全ての要素が 1 のベクトル，$\boldsymbol{I}$ は単位行列，λ は正則化パラメータである．この学習により，GMM のインデックスの事後確率 $p(k|\boldsymbol{y}_i)$ に依存して，どの雑音抑圧手法をどんな重みで採用するのかが適切に学習できると期待される．

Table 1 Condition of recognition experiment

窓関数	ハミング窓
窓長	25 msec
シフト長	10 msec
HMM	16 状態，20 混合 GMM

5 音声認識実験

5.1 実験条件

SPLICE を用いた雑音抑圧手法の統合法の性能を確かめるため，AURORA-2 データベース [11] を用いて音声認識実験を行った．AURORA-2 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．実験条件を Table 1 に示す．データベースには学習データセットが 2 種類あり，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 training) と，クリーン音声と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 (multicondition training) について実験を行った．実験に用いる特徴量は，MFCC（0-12 次元）Δ，ΔΔ の計 39 次元に各雑音抑圧手法を適用したものになっている．

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用意されている．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．

実験に用いる雑音抑圧手法は以下の 4 つである．

- AFE
- VTS（対数パワースペクトル領域で特徴強調する）
- SPLICE（GMM の混合数は 1024）
- HEQ

これらの手法を，信頼度を用いた統合 (Confidence measure)，正則化なしの提案手法 (Proposed)，正則化ありの提案手法 (Proposed with regularized) でそれぞれ統合し，その認識精度の比較を行った．事前に予備実験を行い，信頼度を用いた統合における N-best の N は 20，提案手法（正則化あり）の正則化パラメータ λ は 10^{-3} に設定した．

5.2 認識結果

各評価セットにおける平均の単語正解精度を Figure 4, 5 に示す．参考のために，雑音抑圧手法を逐次的に適用した特徴量の中で最高性能を示した AFE-

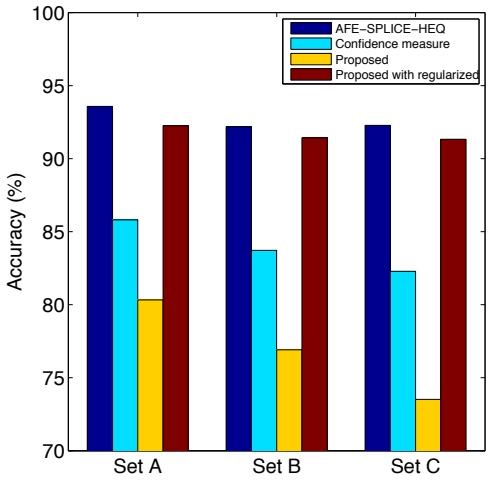

Fig. 4　Averages of accuracies for each training set (clean training)

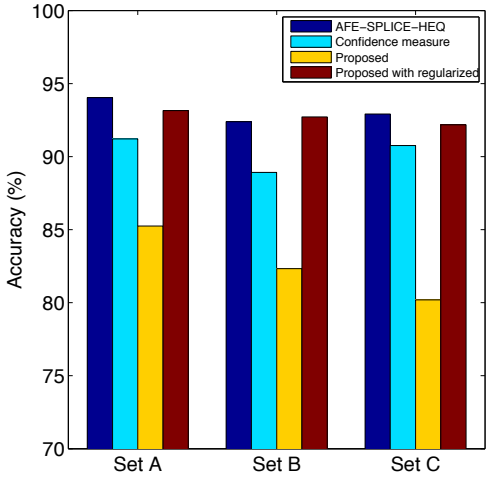

Fig. 5　Averages of accuracies for each training set (multicondition training)

SPLICE-HEQ の結果も載せている．Proposed の認識精度をみると，他の統合法と比べて性能が低く，学習データ不足による過学習が起こっていると考えられる．しかし Proposed with regularized の結果を見ると，正則化を加えることにより大幅に性能が向上しており，信頼度による統合よりも高い認識精度を示していることが分かる．一方で逐次適用した AFE-SPLICE-HEQ の性能を超えることはできていない．

6　まとめ

本稿では SPLICE を用いて複数の雑音抑圧手法を統合する手法を提案し，その有効性を AURORA-2 データベースを用いた音声認識実験で確かめた．正則化を施すことにより，過学習を避けつつ音声認識率の向上を達成し，信頼度による統合よりも高い性能を示すことがわかった．一方で従来の AFE-SPLICE-HEQ を有意に超える結果は得られなかった．

今後は統合する際の SPLICE に関する工夫が必要であると考えている．例えば今回は GMM を仮定して $p(k|\boldsymbol{y}_i)$ を計算したが，$p(k|\boldsymbol{y}_i)$ を識別的に学習する手法も提案されている．他にも SPLICE の入力を単なる特徴量の結合ではなく，前後数フレームの特徴量をまとめて結合することで高性能な特徴量強調が可能であることが報告されている [12]．また提案手法は AFE-SPLICE-HEQ の性能に及ばなかったが，他の雑音環境における提案手法の有効性を確かめるため，別のデータベースによる評価を行いたい．

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, *et. al*, in *Proc. ICSLP*, pp. 17–20, 2002.

[3] O. Kalinli, *et. al*, *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 8, pp. 1889–1901, 2010.

[4] J. C. Segura, *et. al*, in *Proc. EUROSPEECH*, pp. 221–224, Scandinavia, 2001.

[5] J. Droppo, *et. al*, in *Proc. ICSLP*, pp. 29–32, 2002.

[6] A. de la Torre, *et. al*, *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 355–366, May 2005.

[7] Y. Suh, *et. al*, *IEEE Signal Process. Lett.*, vol. 14, no. 4, pp. 287–290, 2007.

[8] 山田武志, 他, 情報処理学会研究報告, SLP-47-18, pp. 95–100, 2003.

[9] 甲斐常伸, 他, 電子情報通信学会技術研究報告, SP2012-28, vol. 112, no. 49, pp. 161–166, 2012.

[10] 岡田治郎, 他, 情報処理学会研究報告, SLP-49-19, pp. 109–114, 2003.

[11] H.-G. Hirsch,*et. al*, in *ISCA ITRW ASR2000*, pp. 181–188, Paris, France, 2000.

[12] M. Suzuki, *et. al*, “Feature Enhancement with Joint Use of Consecutive Corrupted and Noise Feature Vectors with Discriminative Region Weighting,” *IEEE Trans. Audio, Speech, Language Process.*, 2013. (submitted)