主成分分析を用いたGMMに基づく耐雑音音声認識フロントエンドの高精度化*

◎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，スマートフォンなどの普及により，雑音環境下の音声認識の機会が増えている．ところが，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，認識精度を著しく低下させる．そのため，これまで様々な領域において多くの耐雑音手法が提案されている．例えば，単純な処理で良い効果を得られるスペクトルサブトラクションなどがある [1]．それらの中でも，統計モデルに基づいた特徴量変換は高い性能を示し，特に注目されている．これらの手法は，音声特徴量の統計的性質を混合正規分布 (GMM:Gaussian Mixture Model) などでモデル化し，それを用いて特徴量変換する．代表的なものとして，雑音付加音声の特徴量の確率分布をGMMで学習し，クリーン音声と雑音付加音声の対から変換関数を学習するSPLICE(Stereo-based Piecewise LInear Compensation for Environments)[3]，クリーン音声特徴量の確率分布をモデル化したGMMと，推定された雑音特徴量からベクトルテーラー展開 (Vector Taylor Series: VTS) を用いて特徴量を変換する手法 [4] などがある．これらの手法は，実験においてその高い性能を検証されている．

しかし，VTSを用いた手法やSPLICEにもいくつかの問題がある．まずVTSを用いた手法は，対数スペクトルなどの特徴量に対して導入した場合は計算量が低く抑えられるが，MFCCなどに導入した場合では多くの計算量が必要になり現実的でないという問題がある．一方SPLICEは，変換関数を学習した雑音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という問題がある．

これらの問題を解決するために著者らは，主成分分析に基づいたSPLICEの改善手法，Eigen–SPLICEを既に提案している [5]．これは，変換関数を発声毎に適応することで，未知の雑音環境下においても高い性能を発揮するものである．しかし，このEigen–SPLICEは適応の際に，入力音声の始端前と終端後は雑音のみの区間であるという仮定を設け，擬似的なパラレルデータを作成する必要があった．そこで今回は，声質変換の分野で広く用いられている，擬似パラレルデータ無しで変換関数の適応が可能な，Joint VectorのGMMを用いた手法を，雑音環境下における特徴量強調法として導入することを試みる．

2 Joint GMM を用いた特徴量変換

この節では，本研究のベースとなったJoint GMMを用いた特徴量変換手法 (以下J-GMM法と記述)[7]について説明する．この手法は，まず学習データ中の入力特徴量と出力特徴量を連結することでJoint Vectorを得る．次に，得られたJoint Vectorの確率分布をGMMを用いてモデル化する．これをJoint GMMと呼び，その平均や分散共分散行列などのパラメータを用いて，与えられた入力特徴量を所望の特徴量に変換する．この細かい手順について雑音環境下における音声特徴量強調の枠組みとして，以下述べていく．

2.1 パラレル学習

まず，Joint GMMの学習について述べる．ここでは，t フレーム目 (最大 T まで) の入力の雑音付加音声の特徴量を $\boldsymbol{y}_t$，出力のクリーン音声特徴量を $\boldsymbol{x}_t$ と記述する．まず，クリーン音声特徴量とそれに雑音が付加された雑音付加音声の特徴量を対応するフレーム毎に連結し，Joint Vector $\left[\boldsymbol{x}_t^{\mathsf{T}}, \boldsymbol{y}_t^{\mathsf{T}}\right]^{\mathsf{T}}$ を作成する．但し，$^{\mathsf{T}}$ は転置を表す．次に得られたJoint Vectorを用いて，結合確率密度 $p\left(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{\lambda}\right)$ をGMMとして学習する．

$$\boldsymbol{\lambda} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}} \prod_{t=1}^{T} p\left(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{\lambda}\right). \tag{1}$$

ここで，$\boldsymbol{\lambda}$ はGMMのパラメータを表し，学習にはEMアルゴリズムを用いる．GMMとしてモデル化しているので，結合確率密度は以下のように表現される．

$$p\left(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{\lambda}\right) = \sum_{s=1}^{S} \alpha_s \mathcal{N}\left(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right),$$
$$\boldsymbol{\mu}_s = \begin{bmatrix} \boldsymbol{\mu}_s^x \\ \boldsymbol{\mu}_s^y \end{bmatrix}, \ \boldsymbol{\Sigma}_s = \begin{bmatrix} \boldsymbol{\Sigma}_s^{xx} & \boldsymbol{\Sigma}_s^{xy} \\ \boldsymbol{\Sigma}_s^{yx} & \boldsymbol{\Sigma}_s^{yy} \end{bmatrix}. \tag{2}$$

ここで $\mathcal{N}\left(\boldsymbol{x}_t, \boldsymbol{y}_t; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s\right)$ は，平均・分散共分散がそれぞれ $\boldsymbol{\mu}_s$・$\boldsymbol{\Sigma}_s$ となる正規分布であり，s は正規分布のインデックスを表している．また，α_s は s 番目の正規分布の重みを表す．

---

*Principal component analysis based improvement of speech recognition front-end for noisy environments by CHIJIIWA Keigo, SUZUKI Masayuki, MINEMATSU Nobuaki and HIROSE Keikichi, (The University of Tokyo)

2.2 最尤変換

J–GMM 法では，下記の尤度関数の最大化基準に基づき，変換を施す．

$$p\left(\boldsymbol{x}_t|\boldsymbol{y}_t,\boldsymbol{\lambda}\right)=\sum_{s=1}^{S}p(s|\boldsymbol{y}_t,\boldsymbol{\lambda})p(\boldsymbol{x}_t|\boldsymbol{y}_t,s,\boldsymbol{\lambda}), \tag{3}$$

t フレーム目における $p(s|\boldsymbol{y}_t,\boldsymbol{\lambda})$ および $p(\boldsymbol{x}_t|\boldsymbol{y}_t,s,\boldsymbol{\lambda})$ は以下のようになる．

$$p(s|\boldsymbol{y}_t,\boldsymbol{\lambda})=\frac{\alpha_s\mathcal{N}\left(\boldsymbol{y}_t;\boldsymbol{\mu}_s^y,\boldsymbol{\Sigma}_s^{yy}\right)}{\sum_{i=1}^{S}\alpha_i\mathcal{N}\left(\boldsymbol{y}_t;\boldsymbol{\mu}_i^y,\boldsymbol{\Sigma}_i^{yy}\right)}, \tag{4}$$

$$p(\boldsymbol{x}_t|\boldsymbol{y}_t,s,\boldsymbol{\lambda})=\mathcal{N}\left(\boldsymbol{x}_t;\boldsymbol{E}_t(m_s),\boldsymbol{D}(s)\right), \tag{5}$$

ただし，

$$\boldsymbol{E}_t(s) = \boldsymbol{\mu}_s^x+\boldsymbol{\Sigma}_s^{xy}{\boldsymbol{\Sigma}_s^{yy}}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{\mu}_s^y\right), \tag{6}$$

$$\boldsymbol{D}(s) = \boldsymbol{\Sigma}_s^{xx}-\boldsymbol{\Sigma}_s^{xy}{\boldsymbol{\Sigma}_s^{yy}}^{-1}\boldsymbol{\Sigma}_s^{yx}. \tag{7}$$

所望のクリーン音声の特徴量は以下のように求められる．

$$\hat{\boldsymbol{x}}_t=\underset{\boldsymbol{x}}{\operatorname{argmax}}\sum_{s=1}^{S}p(s|\boldsymbol{y}_t,\boldsymbol{\lambda})p(\boldsymbol{x}_t|\boldsymbol{y}_t,s,\boldsymbol{\lambda}). \tag{8}$$

これから出力される特徴量は，

$$\hat{\boldsymbol{x}}_t=\sum_{s=1}^{S}p(s|\boldsymbol{y}_t,\boldsymbol{\lambda})\left\{\boldsymbol{\mu}_s^x+\boldsymbol{\Sigma}_s^{xy}{\boldsymbol{\Sigma}_s^{yy}}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{\mu}_s^y\right)\right\} \tag{9}$$

のように推定される．

3 主成分分析を用いた改良

以上のような J-GMM 法を主成分分析を用いて，変換方法を適応できるように改善する手法について説明する．これは多対一の声質変換を可能にする固有声に基づいた声質変換手法 (Eigenvoice Conversion: EVC)[6] を雑音環境下の音声特徴量強調に導入したものである．

この手法は，まず全学習データを使って特定の雑音環境に依存しない汎用的な Joint GMM を学習し，次に特定の雑音環境のデータのみを用いて，その環境に適するような Joint GMM へとパラメータを再学習する．その後，再学習したパラメータを連結し，雑音の種類数と同じだけの Super Vector を作成する．得られた Super Vector に主成分分析を施し，高次元の Joint GMM のパラメータを低次元の重みベクトルで表現できるようにする．その結果，未知の雑音環境下であったとしても，低次元の重みベクトルを推定するだけで Joint GMM を適応することが可能になる．この Eigen Joint GMM を用いた特徴量変換手法を，以下 E-J-GMM 法と記述し，その手順について詳しく説明していく．

3.1 事前の学習

まず，学習データ中の全種類の雑音環境のパラレルデータを用いて，環境非依存の Joint GMM，$\boldsymbol{\lambda}^0$ のパラメータを学習する．その後，雑音のタイプや SNR の異なる雑音環境毎の学習データを用いて，$\boldsymbol{\lambda}^0$ を環境 e に適した $\boldsymbol{\lambda}^e$ へ再学習する．この際，平均ベクトル $\boldsymbol{\mu}^y$ のパラメータのみを再学習する．次に，環境依存 $\boldsymbol{\lambda}^e$ の平均ベクトル $\boldsymbol{\mu}_s^{y^e}$ を連結し，以下のように Super Vector $\boldsymbol{SV}^e$ を作成する．

$$\boldsymbol{SV}^e=\left\{\boldsymbol{\mu}_1^{y^e},\cdots,\boldsymbol{\mu}_s^{y^e},\cdots,\boldsymbol{\mu}_S^{y^e}\right\}. \tag{10}$$

得られた複数の Super Vetors に対して主成分分析を施すことで，Super Vector の主成分行列 $\boldsymbol{PC}^m$ とバイアスベクトル $\boldsymbol{BV}$ を得る．

$$\boldsymbol{BV} = \{\boldsymbol{b}_1,\cdots,\boldsymbol{b}_s,\cdots,\boldsymbol{b}_S\}, \tag{11}$$

$$\begin{aligned}\boldsymbol{PC}^1 &= \{\boldsymbol{c}_1^1,\cdots,\boldsymbol{c}_s^1,\cdots,\boldsymbol{c}_S^1\},\\ &\vdots\\ \boldsymbol{PC}^M &= \{\boldsymbol{c}_1^M,\cdots,\boldsymbol{c}_s^M,\cdots,\boldsymbol{c}_S^M\}.\end{aligned} \tag{12}$$

ただし，m は主成分のインデックスであり，M は用いる主成分の数である．

3.2 変換

これら $\boldsymbol{BV},\boldsymbol{PC}$ を用いて，ある環境 e に適する Joint GMM の平均ベクトルは，重みベクトル $\boldsymbol{w}$ によって下記のように記述できる．

$$\begin{aligned}\boldsymbol{\mu}_s^e &= \boldsymbol{B}_s\boldsymbol{w}+\boldsymbol{b}_s, \text{ただし}\\ \boldsymbol{B}_s &= \{\boldsymbol{c}_s^{1\mathsf{T}},\cdots,\boldsymbol{c}_s^{M\mathsf{T}}\}\end{aligned} \tag{13}$$

つまり，この Joint GMM を用いた変換は下記のようになる．

$$\hat{\boldsymbol{x}}_t=\sum_{s}p(s|\boldsymbol{y}_t,\boldsymbol{\lambda}^e)\left\{\boldsymbol{\mu}_s^x+\boldsymbol{\Sigma}_s^{xy}{\boldsymbol{\Sigma}_s^{yy}}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{B}_s\boldsymbol{w}-\boldsymbol{b}_s\right)\right\}. \tag{14}$$

3.3 入力環境への適応

次に，教師なしで入力雑音環境に適する重みベクトル $\boldsymbol{w}$ を決定する手順について説明する．推定では，以下の最尤推定を行う．

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\int p(\boldsymbol{x},\boldsymbol{y}^{(\mathrm{IN})}|\boldsymbol{\lambda}^e)d\boldsymbol{x}, \tag{15}$$

$$= \underset{\boldsymbol{w}}{\operatorname{argmax}}\,p(\boldsymbol{y}^{(\mathrm{IN})}|\boldsymbol{\lambda}^e). \tag{16}$$

ただし，$\boldsymbol{y}^{(\mathrm{IN})}$ は入力未知環境の雑音付加音声特徴量の系列を表し，．このとき以下の関数を繰り返し最大化することによって，教師なしで重みベクトルを推定することができる．

$$Q(\boldsymbol{w},\boldsymbol{w}^{\mathrm{new}})=\sum_{t=1}^{T}\sum_{s=1}^{S}p(s|\boldsymbol{y}_t^{(\mathrm{IN})},\boldsymbol{\lambda}^e)\log p(s|\boldsymbol{y}_t^{(\mathrm{IN})},\boldsymbol{\lambda}^{e\,\mathrm{new}}) \tag{17}$$

更新式は以下のようになる．

$$\boldsymbol{w}^{\mathrm{new}}=\left\{\sum_{s=1}^{S}\overline{\gamma_s^{(\mathrm{IN})}}\boldsymbol{B}_s^{\mathsf{T}}\boldsymbol{\Sigma}_s^{yy^{-1}}\boldsymbol{B}_s\right\}^{-1}\sum_{s=1}^{S}\boldsymbol{B}_s^{\mathsf{T}}\boldsymbol{\Sigma}_s^{yy^{-1}}\overline{\boldsymbol{y}_s^{(\mathrm{IN})}} \quad (18)$$

ただし，

$$\overline{\gamma_s^{(\mathrm{IN})}} = \sum_{t=1}^{T} p(m_s|\boldsymbol{y}_t^{(\mathrm{IN})},\boldsymbol{\lambda}^e) \quad (19)$$

$$\overline{\boldsymbol{y}_s^{(\mathrm{IN})}} = \sum_{t=1}^{T} p(m_s|\boldsymbol{y}_t^{(\mathrm{IN})},\boldsymbol{\lambda}^e)\left(\boldsymbol{y}_t^{(\mathrm{IN})}-\boldsymbol{b}_s\right) \quad (20)$$

なお，最初の事後確率を求める際には，環境非依存の $\boldsymbol{\lambda}^0$ のパラメータを用いる．E–J–GMM 法の概要を Fig. 1 に示す．

4 Eigen–SPLICE

この節では，著者らが既に提案している Eigen–SPLICE についても簡単に説明しておく [5]．Eigen–SPLICE は，従来の SPLICE を主成分分析によって改良したものである．これは，J–GMM 法を主成分分析によって E–J–GMM に改良したものと対応している．しかし，E–J–GMM 法と違い，Eigen–SPLICE は変換関数を適応するためには，入力音声の始端前と終端後が雑音のみの区間であるという仮定を設け，擬似パラレルデータを作成する必要がある．その一方で，E–J–GMM 法とは異なり，Eigen–SPLICE は適切な重みを繰り返し計算ではなく，擬似パラレルデータを用いて解析的に求めることができる．

$$\hat{\boldsymbol{x}} = \sum_{s} p(s|\boldsymbol{y},\boldsymbol{\lambda}^y)\left(\boldsymbol{A}_s\boldsymbol{y}+\boldsymbol{r}_s\right), \quad (21)$$

従来の SPLICE は上記の式のような方法で特徴量変換する．ここで，$\boldsymbol{\lambda}^y$ は雑音付加音声の特徴量の確率密度を GMM でモデル化したもので，$\boldsymbol{A}_s, \boldsymbol{r}_s$ はそれぞれ事前に学習した変換行列と補正ベクトルである．これを J–GMM 法を E–J–GMM に改良したのと同様の手順で主成分行列とバイアスベクトルを得て，下記のように重みベクトル $\boldsymbol{w}$ で表現できるように改良する．

$$\hat{\boldsymbol{x}}_t = \sum_{s} p(s|\boldsymbol{y}_t,\boldsymbol{\lambda}^y)\left(\boldsymbol{A}_s\boldsymbol{y}_t+\boldsymbol{B}_s\boldsymbol{w}+\boldsymbol{b}_s\right). \quad (22)$$

詳しくは [5] を参照されたい．

5 実験

それぞれの手法の性能を比較するために，雑音環境における音声認識用のデータベース AURORA–2[8] を用いて以下のような実験を行った．

5.1 実験条件

まず，AURORA–2 の train セットの 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 Joint GMM $\boldsymbol{\lambda}^0$ を学習する．次に，それぞれの雑音環境のデータを用いて再学習を行い，雑音環境毎の $\boldsymbol{\lambda}^e$ を得る．その後，$\boldsymbol{\mu}_s^y$ の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．そして，AURORA–2 の test セットに対して認識実験を行った．test A セットは train セットと同じ種類の雑音環境で，test B 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，test B セットは学習環境とは異なる，未知の雑音環境である．

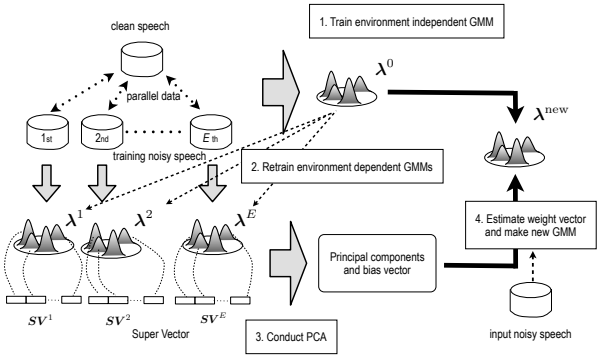

Fig. 1 Overview of Eigen Joint GMM method

Table 1 Word recognition accuracies [%] in the four cases of (a) no enhancement, (b) Joint GMM method, (c) Eigen Joint GMM method, (d) SPLICE and (e) Eigen-SPLICE

	(a)	(b)	(c)	(d)	(e)
test A	51.45	77.34	76.97	78.02	81.86
test B	44.86	77.60	78.68	77.87	81.28

今回，E–J–GMM 法で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 3 とし，入力一発声毎に 5 回更新を繰り返すことで重みを推定した．簡単の為，MFCC13 次元に対して変換を行い，変換後の特徴量から Δ と ΔΔ を計算し，MFCC+Δ+ΔΔ の 39 次元として認識した．認識用の音響モデルは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．また，GMM の混合数は 256 混合とした．

また，比較対象の Eigen–SPLICE では，雑音付加音声特徴量の GMM の混合数は E–J–GMM と同様の 256，予備実験で高い性能を示した主成分数は 6，擬似パラレルデータは 8 発声分で，その他は同様の条件で特徴量変換した場合の結果も示した．認識結果は SNR が 0[dB] から 20[dB] の Accuracy の平均を用いて評価した．

5.2 結果と考察

認識結果を Table 1 に示す．また，test B の詳細結果についても Table 2 に示す．

Table 2　Word recognition accuracies [%] (a) without enhancement, baseline, (b) with the Joint GMM method, (c) with the Eigen Joint GMM mthod.

(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	98.07	96.70	98.33	98.09	97.80
15dB	96.53	93.26	95.94	95.90	95.41
10dB	89.75	83.10	90.49	89.39	88.18
5dB	75.04	59.22	71.37	72.69	69.58
0dB	42.46	28.93	37.76	39.06	37.05
Avg.	80.37	72.24	78.78	79.03	77.60

(c)	Rest.	Street	Airport	Station	Avg.
20dB	98.56	97.40	97.55	98.18	97.92
15dB	97.67	95.92	95.38	96.39	96.34
10dB	91.19	89.33	90.01	89.17	89.92
5dB	77.22	66.87	72.83	67.42	71.09
0dB	46.70	32.32	41.46	31.97	38.11
Avg.	82.27	76.37	79.45	76.63	78.68

結果を見ると，単純な J–GMM 法であったとしても test B セットで 77.60%の精度を示しており，やや劣るものの従来の SPLICE と同程度の性能を発揮している．つまり，声質変換の分野で用いられていた Joint GMM の手法が雑音環境下における音声特徴量強調にも導入できるということが言える．また，E–J–GMM 法は test A において J–GMM 法と比較して 1.6%の誤り増加が見られたものの，未知の雑音環境である test B においては相対的に 4.8%の誤り削減が見られた．しかし，Eigen–SPLICE が test B セットにおいて SPLICE と比較して 15.4%の誤りの削減をしていることを考慮すると，改善の効果は限定的である．

Eigen–SPLICE と違い E–J–GMM 法では雑音付加音声とクリーン音声とが GMM のインデックスを共有しているため，事後確率 $p(s|\boldsymbol{y}_t, \boldsymbol{\lambda}^e)$ を求める際に必要な GMM が，雑音の特性に基づいて学習されていないことが原因であると考えている．

6　まとめ

本論文では，声質変換で用いられている EVC を雑音環境下における音声特徴量強調の分野にも導入し，その性能を SPLICE などの特徴量強調手法と比較した．この E–J–GMM の手法は，低次元の重みベクトルを入力の雑音環境に応じて適応することで，未知環境においても高精度の特徴量強調を実現する．E–J–GMM 法は著者らが既に提案している Eigen–SPLICE と比較して，発声の始端前と終端後が雑音のみの区間であるという仮定を必要としないという利点があるものの，適応による性能向上は限定的である．今回の結果により，仮定が妥当である場合においては，Eigen–SPLICE の手法が E–J–GMM 法よりも有効であるということが示された．

今後の展望としては，本報告で検証した特徴量変換の枠組みがどのような特徴量に対しても導入可能である利点を活かして，より耐雑音性の高い Noise Mean Normarization などを特徴量に施すことを検討している．また，特徴量強調の不確かさを認識段階で考慮する Uncertainty Decoding[9] に，この主成分分析を用いた特徴量強調を導入することも検討している．これらの試みによって，より雑音に頑健な音声認識フロントエンドが実現すると考えている．

参考文献

[1] 猿渡洋, 信学技報, SP2010-103, pp.1–6, 2011.

[2] Boll S., *IEEE Trans. Acoust. Speech, Signal prosecc.,* Vol.27, No.2, pp.113–120, 1979.

[3] Droppo J., *et al.*, *EUROSPEECH-2001*, pp.217–220, 2001.

[4] Stouten V., *KU Leuven, Ph.D. Thesis,* 2006.

[5] 千々岩圭吾, *et al.*, 音講論（秋）, pp.113–116, 2011.

[6] 戸田智基, *et al.*, 信学技報, SP2006-39, pp.25–30, 2006.

[7] Kain, A. *et al.*, *Proc. of ICASSP '98*, Vol.1, pp.285–288, 1998.

[8] Pearce D., *et al.*, *Proc. of ICSLP '00*, Vol.4, pp.29–32, 2000.

[9] Droppo, J. *et al.*, *Proc. of ICASSP '02*, Vol.1, pp.57–60, 2002.