論文 / 著書情報
Article / Book Information

論題(和文)	目的音GMMを用いたスペクトル補正フィルタの提案
Title(English)	Target Speech GMM-based Spectral Compensation Filter for Noisy Speech Recognition
著者(和文)	篠﨑 隆宏, 古井 貞熙
Authors(English)	Takahiro SHINOZAKI, Sadaoki FURUI
出典(和文)	日本音響学会2008年秋季講演論文集, , No. 1-1-1, p. 1-2
Citation(English)	, , No. 1-1-1, p. 1-2
発行日 / Pub. date	2008, 9

目的音GMMを用いたスペクトル補正フィルタの提案*

○篠崎 隆宏, 古井貞熙 (東工大)

1 はじめに

実環境の音声認識では、目的音声以外の音の重畳に対応する加算性雑音および音声伝達チャネルの特性に対応する乗算性雑音の認識性能への影響を減らすことが重要となる。これら雑音の効果はスペクトル領域ではそれぞれ単純な加算および乗算として表現されることから、最も直接的な雑音補正法として、スペクトル領域で入力信号に逆変換を適用しクリーン音声を推定することが考えられる。

これまでに耐雑音性を目的としてスペクトル領域で入力信号に変換を適用する手法として、スペクトルサブトラクション法 [1] や尤度を基準とした特徴量ベースのストキャスティックマッチング法 [2] などが提案されている。スペクトルサブトラション法では無音声区間などからノイズベクトルを求めクリーン音声の推定に用いることで加法性雑音に対応するのに対し、ストキャスティックマッチング法では認識結果 [2] または人手によるラベル [3] を用いたHMM尤度を基準として補正項を求め、加法性および乗算性雑音に対処する。ストキャスティックマッチングによる手法は目的音声のモデルを用いるという点でSeguraらによるモデルベースの補正法と似ているが、前者が音声モデルのみを用いて線形領域におけるスペクトル変換を推定するのに対し、後者では音声および雑音モデルのパラメタから近似的な期待値として対数領域での補正項を求め、また加算性雑音のみに対応することが相異点として挙げられる。

本研究では目的音声GMMによる尤度を基準としてスペクトル領域で補正を行い、数百ミリ秒程度の時間長に対して準定常的な加算性および乗算性雑音に対処する手法を提案する。提案法は尤度を基準として線形スペクトル領域での変換を行う点および雑音モデルが不要である点はストキャスティックマッチング法と同様であるが、尤度評価にGMMを用いるため音声ラベルを必要とせず、フロントエンドにおけるフィルタとして認識システムと独立して動作する特徴がある。また、提案法では雑音の定常性を仮定するスペクトルサブトラクション法では補正しきれない雑音の変動に対しても効果が期待できる。またさらにこの他に本研究の特徴として、スペクトル領域での変換の際に必要となるスペクトラルサブトラクションのフロアリングに相当する操作もパラメタ最適化において考慮されていることが挙げられる。

2 目的音GMMを用いたスペクトル補正フィルタ

2.1 スペクトル補正変換

クリーン音声のスペクトルを x_ω、音声伝達チャネルの特性を a_ω、目的音以外の加算性雑音を b_ω とすると、雑音重畳音声スペクトル n_ω は式 (1) のように表すことができる。

$$n_\omega = a_\omega \cdot x_\omega + b_\omega. \tag{1}$$

よって、もし a_ω および b_ω が与えられたとすると、式 (2) により雑音重畳音声からクリーン音声を求めることができる。

$$x_\omega = \frac{n_\omega}{a_\omega} - \frac{b_\omega}{a_\omega}. \tag{2}$$

以下ではより一般的にスペクトル補正変換をパラメタ a_ω および b_ω に依存して雑音重畳音声を変換する関数として $f(n_\omega, a_\omega, b_\omega)$ と表すことにする。また、a_ω および b_ω を要素とするパラメタベクトルを A および B とする。

2.2 変換パラメタの推定

提案法では一定時間区間における変換パラメタ A および B を、式 (3) に示す補正後のスペクトルから求めた音声認識特徴量 Y_t に対する対数GMM尤度和を最大化する値として、式 (4) のように求める。

$$L(A,B) = \sum_t L_{GMM}(Y_t), \tag{3}$$

$$\{A_{opt}, B_{opt}\} = \underset{A,B}{\operatorname{argmax}} \{L(A,B)\}. \tag{4}$$

式 (4) の局所最適解 $\{A_{opt}, B_{opt}\}$ は式 (5) に示す L の $p_\omega = a_\omega$ 又は b_ω に関する偏微分を用いて最急降下法により求めることができる。

$$\frac{\partial L}{\partial p_\omega} = \sum_t \sum_k \frac{\partial L^t_{GMM}}{\partial y^t_k} \frac{\partial y^t_k}{\partial f^t_\omega} \frac{\partial f^t_\omega}{\partial p_\omega}. \tag{5}$$

ここで、t は時刻, $\frac{\partial L^t_{GMM}}{\partial y^t_k}$ は対数GMM尤度の音声認識特徴量の第 k 要素 y^t_k に関する偏微分、$\frac{\partial y^t_k}{\partial f^t_\omega}$ は音声認識特徴量のスペクトル変換後の入力信号に関する偏微分、$\frac{\partial f^t_\omega}{\partial p_\omega}$ は変換関数のパラメタ p_ω に関する偏微分である。

*Target Speech GMM-based Spectral Compensation Filter for Noisy Speech Recognition, by Takahiro Shinozaki, Sadaoki Furui (Tokyo Institute of Technology)

特徴量が MFCC の場合、$\frac{\partial y_k}{\partial f_\omega}$ は式 (6) のように表すことができる。

$$\frac{\partial y_k}{\partial f_\omega} = \frac{\partial}{\partial f_\omega} \sum_j C_{k,j} \log \left(\sum_{\omega'} w_{j,\omega'} \cdot f_{\omega'} \right). \quad (6)$$

ここで、y_k は MFCC の第 k 要素、f_ω は補正後の振幅スペクトルベクトルの第 ω 要素、$C_{k,j}$ はコサイン変換行列の (k,j) 要素、$w_{j,\omega'}$ はフィルタバンク行列の (j,ω') 要素である。またデルタ特徴量は各特徴量次元ごとに時刻方向に数フレームの重み和をとったものであるから、対応する特徴量の偏微分の重み和として求めることができる。

3 実験条件

変換関数 $f(n_\omega, a_\omega, b_\omega)$ としては様々なバリエーションが考えられるが、本研究では式 (7) に示す変換を連続関数で近似した、式 (8) を用いた。

$$f = \max\left\{a_\omega^2 \cdot n_\omega - b_\omega^2,\ 0.1 n_\omega\right\} \quad (7)$$

$$\approx log\left(exp\left(a_\omega^2 \cdot n_\omega - b_\omega^2\right) + exp\left(0.1 n_\omega\right)\right). \quad (8)$$

式 (8) において、入力雑音重畳音声 n_ω は振幅スペクトルとした。また、a_ω, b_ω は変換の実数値パラメタ、$0.1n_\omega$ はフロアリング値である。

実験は（社）情報処理学会 音声言語情報処理研究会 雑音下音声認識評価ワーキンググループ 雑音下音声認識評価環境（AURORA-2J）のデータを用いて行った。特徴量は MFCC12 次元とそのデルタ、および C0 項のデルタの計 25 次元である。音響モデルはクリーンコンディションで作成したものを用いた。

変換関数において、a_ω^2 の初期値は 1.0 とした。また b_ω^2 の初期値については定数 (100) と、スペクトルサブトラクションと同様に各音声セグメントのはじめの 10 フレームより推定した雑音ベクトルの、2 通りを用いた。後者の場合、パラメタ更新を行わなければ従来のスペクトルサブトラクション法と同じ結果となることから、スペクトラルサブトラクション法と提案法の組み合わせと見ることができる。変換パラメタは 50 フレーム（500ms）を一つの区画として、区画ごとに求め、変換の適用を行った。最急降下法の繰り返し数は 5 回である。

4 実験結果

表 1 に AURORA-2J の A，B および C セットにおける正解精度の平均を示す。表で、BASE は耐雑音処理を行わないベースライン、TGSC-C が提案法において b_ω^2 の初期値に定数を用いた場合、SS がスペクトルサブトラクションを行った場合、TGSC-SS が提案法とスペクトラルサブトラクション法を組み合わせた場合である。

Table 1 Recognition accuracy

SNR(dB)	BASE	TGSC-C	SS	TGSC-SS
clean	99.6	99.0	98.8	99.2
20	93.2	92.1	91.7	94.4
15	81.7	82.7	84.8	**90.2**
10	60.0	67.0	70.1	**79.3**
5	35.6	**47.0**	46.5	58.3
0	19.0	26.1	23.2	32.0
-5	10.7	11.9	12.0	15.0

表より、BASE と TGSC-C を比べると、SNR が 15dB 以下の場合において認識率が向上していることが分かる。また、SS および TGSC-SS を比べると全ての SNR においてスペクトラルサブトラクション法の認識率よりもさらに認識率が向上しており、提案法はスペクトラルサブトラクション法と組み合わせた場合の相乗効果が良好であることが分かる。TGSC-C により、BASE と比較して相対的な誤り率が最大で 18%削減された。また TGSC-SS を用いることで、SS と比較して最大で 35%、BASE と比較して 48%、誤り率が削減された。

5 まとめ

目的音の GMM 尤度を基準に雑音重畳音声からクリーン音声への変換関数を推定し、得られた変換関数を用いてスペクトル領域で入力信号の補正を行うフィルタ手法を提案した。提案手法において、変換関数のパラメタを定数で初期化した場合、SNR が 15dB 以下の場合において認識率が向上した。また、スペクトルサブトラクション法と提案法を組み合わせた場合良好な相乗効果が得られ、SS 法による認識率の向上に加えて、更に大きな認識率の改善が得られた。

謝辞 本研究は経産省「情報家電センサー・ヒューマンインターフェースデバイス活用技術開発・音声認識基盤技術」プロジェクトの支援により行った。

参考文献

[1] S. F. Boll, IEEE Trans. ASSP, Vol. 27, No. 2, pp. 113-120, Apr 1979.
[2] A. Sankar, C. H. Lee, IEEE Trans. SAP, Vol. 4, No. 3, pp190-202, May 1996.
[3] D. Kim, D. Yook, IEE Electronics Letters, Vol. 40, No. 20, pp. 1313-1314, Sept 2004.