論文 / 著書情報
Article / Book Information

論題(和文)	目的音GMM尤度基準スペクトル補正法の諸評価
Title(English)	
著者(和文)	篠崎 隆宏, 古井 貞熙
Authors(English)	Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	日本音響学会2009年秋季講演論文集, , No. 1-1-10, pp. 31-32
Citation(English)	, , No. 1-1-10, pp. 31-32
発行日 / Pub. date	2009, 9

目的音GMM尤度基準スペクトル補正法の諸評価*

○篠崎隆宏, 古井貞熙 (東工大)

1 はじめに

電話やインターネットを介した音声対話システムなどにおいて，様々な雑音環境に対して頑健に動作するオンライン認識システムを実現することは非常に重要である．音声認識システムに入力される音声に重畳される雑音は加算性の雑音と乗算性の雑音に分類することができ，スペクトル領域ではそれぞれ音声に対する加算項および乗算項として表現される．これら加算項および乗算項が推定できれば，元の音声を雑音重畳音声から周波数ごとにアフィン変換を行うことで求めることができる．そこで，これまでに音声GMMの尤度を基準として0.5秒程度の音声区間毎に加算項および乗算項を推定しスペクトル領域で雑音補償を行う手法を提案し，音声認識実験において有効性を示した．提案法は認識器とは独立したフロントエンドとして動作し，また比較的短い時間スケールでの雑音変動への対応が可能である．本稿では本手法について，使用するGMMの混合数と認識性能の関係やケプストラム平均正規化法と組み合わせた場合の効果などについて評価を行う．

2 GMMを用いたスペクトル補正変換

本節では目的音GMM尤度基準スペクトル補正法(TGSC) [1]について，簡単に説明する．雑音重畳音声の振幅スペクトルを n_ω，乗算性雑音および加算性雑音の推定値をそれぞれ a_ω，b_ω とすると，クリーン音声スペクトル x_ω は式(1)により推定できる．

$$x_\omega = \frac{n_\omega}{a_\omega} - \frac{b_\omega}{a_\omega}. \tag{1}$$

より一般的に，パラメタ a_ω および b_ω に依存して補正変換を行う関数を $f(n_\omega, a_\omega, b_\omega)$ と表すことにする．変換パラメタ a_ω および b_ω を要素とするベクトルをそれぞれ A および B とすると，それらは式(2), (3)に示すように音声GMMの尤度を最大とする値として推定できる．

$$L(A,B) = \sum_t L_{GMM}(Y_t), \tag{2}$$

$$\{A_{opt}, B_{opt}\} = \underset{A,B}{\operatorname{argmax}} \{L(A,B)\}. \tag{3}$$

ここで，Y_t は補正後のスペクトルから導出した音声認識特徴量である．実際の補正変換 $f(n_\omega, a_\omega, b_\omega)$ としては式(1)では補正後の振幅スペクトルが負になってしまうことがあることに対応し，また最急上昇法によるパラメタ最適化を可能とするため，式(4)に示す連続関数を用いる．

$$\begin{aligned} f &= \max\left\{a_\omega^2 \cdot n_\omega - b_\omega^2,\ 0.1 n_\omega\right\} \\ &\approx log\left(exp\left(a_\omega^2 \cdot n_\omega - b_\omega^2\right) + exp\left(0.1 n_\omega\right)\right). \end{aligned} \tag{4}$$

式(4)において a_ω^2 および b_ω^2 を2乗で定義しているのは，音声項に対する係数としての値が常に正になることを保証するためである．

3 実験条件

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．音声の分析条件は窓幅25msフレームシフト10msであり，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．音響モデルおよび雑音補正に用いるGMMはクリーンコンディションで作成したものを用いた．GMMはAg-EM法 [2]によるパラメタ推定およびAgCV法 [3]による混合数最適化法を用いて学習した．

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，各発話のはじめの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(SS法) [4]と同じ結果となる．以下では定数を用いて初期化した場合をTGSCc，SS法と同様に初期化した場合をTGSCssと表記する．変換パラメタは50フレーム（0.5秒）を一つの区画として重複の無い区画ごとに求め，補正変換はパラメタ推定と同じ区画に対して適用した．したがって計算時間を除いた最小遅延時間は0.5秒である．音響モデルの学習および認識処理はコーパス付属のスクリプトに基づいており，このためコーパスガイドラインにおける実験カテゴリは0である．

ケプストラム平均正規化(CMS) [5]については，適用しない条件および適用する条件それぞれについて実験を行った．CMSを適用する場合においてTGSC法との組み合わせは，TGSC法により補正の後CMSを発話単位で適用することにより行った．その際，TGSC法ではCMSをかけずに学習したGMMを使用した．

*Evaluation of the target speech GMM-based spectral compensation method. by SHINOZAKI, Takahiro and FURUI, Sadaoki (Tokyo Institute of Technology)

Table 1 *Number of iterations of the gradient ascent to optimize TGSCss and word accuracy. Zero-th iteration is the result of spectral subtraction.*

	SNR						
# iter	clean	20	15	10	5	0	-5
0	98.7	92.2	83.7	66.5	41.6	21.8	11.6
1	98.8	94.1	86.6	71.0	46.5	23.8	12.5
2	99.0	94.9	88.2	72.9	48.9	24.9	13.1
5	99.0	95.8	89.8	75.1	50.0	25.5	12.8
10	98.3	95.8	89.6	74.0	48.0	23.5	12.1

Table 2 *Number of iterations of gradient ascent and real time factor (RTF).*

# iter	0	1	2	5	10
RTF	0.02	0.63	1.2	3.2	6.1

4 実験結果

表 1 に最急上昇法の繰り返し数と単語認識率の関係を示す．表において繰り返し数 0 は SS 法の結果であり，また補正変換に用いた GMM の混合数は 422 である．表より認識率の向上について繰り返し一回目の効果が大きく 5 回目でほぼ最大の効果が得られることが分かる．繰り返しを 10 回とすると SNR によっては 5 回目より若干低下がみられるが，これは過学習のためと考えられる．実験に用いたプログラムは octave スクリプトを用いたものであり余り高速化に重点をおいた実装とはなっていないが，参考として表 2 に最急上昇法の繰り返し数と計算時間の関係を示す．TGSC はパラメタ最適化を行うため SS 法より計算量が多くなるが，繰り返し数が 1 のときは実時間以下で動作している．以下では最急上昇法の繰り返し数は 5 回とする．

図 1 に雑音補正に用いる GMM の混合数と単語認識率の関係を SNR=10dB の場合について示す．図より TGSCss は混合数 2 以上で，GMM を用いた最適化を行わない SS 法と比較して高い認識性能が得られていることが分かる．また，認識率の向上は混合数の増加とともに大きくなり，およそ 400 混合あたりでほぼ収束している．

表 3 に CMS を用いた場合の結果を示す．GMM の混合数は 422 である．変換パラメタ b_ω^2 を定数で初期化した TGSCc と CMS を組み合わせた場合，CMS のみを用いた場合と比較して SNR=-5 から 15dB の条件で有意に認識率が向上した．また TGSCss と CMS を組み合わせた場合，SS と CMS を組み合わせた場合と比較して全ての条件で認識率が有意に向上し，CMS と組み合わせた場合においても TGSC が認識率の向上に有効であることが示された．

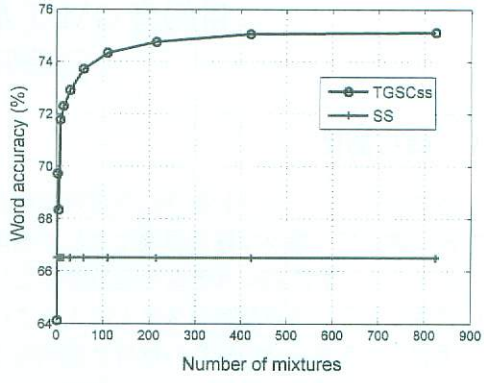

Fig. 1 *Number of mixtures of a GMM for TGSCss and word accuracy when SNR=10dB.*

Table 3 *SNR(dB) and Word accuracy when CMS was applied. "C" is CMS, "T" is TGSCc with CMS, "Css" is CMS with spectral subtraction, "Tss" is CMS with TGSCss.*

	SNR						
	clean	20	15	10	5	0	-5
C	99.5	95.7	85.5	58.3	31.0	21.4	13.4
T	99.0	94.5	**86.5**	**67.4**	**44.4**	**26.0**	**15.5**
Css	99.1	94.6	89.6	77.4	56.0	31.1	15.5
Tss	**99.3**	**97.0**	**93.3**	**83.5**	**62.4**	**35.2**	**16.9**

5 おわりに

これまでに提案した目的音 GMM 尤度基準スペクトル補正法 (TGSC) の諸条件における評価を行った．TGSC が最急上昇法の繰り返し数を 1 とすることで認識率を向上させながら実時間で動作することを実証し，また GMM の混合数としては 2 混合から効果が得られ 400 混合でほぼ最大の効果が得られることを示した．さらに TGSC 法を CMS と組み合わせた実験を行い，CMS 法との組み合わせにおいても TGSC が有効であることを示した．

謝辞 本研究は科研費 (21700188) の助成を受けたものである．

参考文献

[1] 篠崎，古井，音講論 (秋)，1-2，2008.
[2] Shinozaki et al., Proc. ICASSP, pp. 4405–4408, 2008.
[3] Shinozaki et al., Proc. Interspeech, pp. 2382–2385, 2008.
[4] Boll, IEEE Trans. ASSP, Vol. 27, No. 2, pp. 113-120, 1979.
[5] Atal, JASA, vol. 55, no. 6, pp. 1304–1312, 1974.