GMMに基づく音声特徴量の時間変動を考慮した突発性雑音の除去*

◎三宅信之, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

雑音が混入することで音声認識率は低下する．そのため，雑音除去に関する様々な研究がなされてきた．我々はこれまでに，突発的な雑音に対する除去法を提案した [1]．これはまずフレームごとに突発的雑音があるかどうかの判定を行い，重畳している雑音の識別を行い，雑音と音声の静的特徴を利用して除去を行うものであった．この方法では最初の雑音の有無の判定で，雑音が未検出であった場合，除去が行われず，認識に悪影響を及ぼす可能性がある．そこで本研究では前フレームからの時間変動と静的特徴の両方を考慮した雑音除去法を使用し，今まで除去が行われなかったフレームに対してもある程度除去を行う．また，以前の静的特徴量のみを利用した除去法に比べ，変動成分を考慮することで，雑音が検出されたフレームに対しても，雑音除去の効果がより高くなると期待できる．

2 雑音の検出と識別

ここで用いる検出・識別手法は文献 [1] と同様であり，まずフレーム単位で突発的雑音があるかどうかをBoostingで作成した識別器で判定し，雑音があると判断された場合，それに最も近いクラスを，同様にBoostingで作成した識別器で識別する．そして識別結果のクラスの平均値を以下の雑音の特徴量として使用する．なお，この段階で雑音がないと判定された場合，以下で述べる除去時の雑音の特徴量はすべて0として扱う．

3 提案手法

本研究で扱うタスクにおいて，雑音除去に音声の時間変動を利用することは，既提案手法 [1] を利用した場合、雑音が検出できなかったフレームに対しても，雑音除去が可能になるため、以前の手法よりも高い雑音除去効果が得られると期待できる．そこでLiらによって提案された音声の動的特徴を利用した雑音除去法 [2] を用い，それに雑音の強さの推定を組み込むことで突発的雑音の除去を行う．

3.1 雑音重畳音声

フレーム t における，観測信号の k 次元目のメルフィルタバンク特徴量 $X_t(k)$ はクリーン音声 $S_t(k)$，雑音 $N_t(k)$ を使って

$$X_t(k) \approx S_t(k) + N_t(k) \tag{1}$$

と近似できる．2節による識別結果によって重畳している雑音の種類が何であるかが与えられるので，雑音 $N_t(k)$ の形状は既知だと仮定する．しかしSNRの推定は行われていないため，雑音の強さは分からない．そこで雑音の強さを表す重み α を雑音に掛け合わせて

$$X_t(k) \approx S_t(k) + \alpha \cdot N_t(k) \tag{2}$$

とする．このことから対数メルフィルタバンク $x_t(k) = log(X_t(k))$ は音声と雑音の対数メルフィルタバンク $s_t(k), n_t(k)$ を使って以下のように書き表される．

$$\begin{aligned} x_t(k) &= \log\left(e^{s_t(k)} + \alpha \cdot e^{n_t(k)}\right) \\ &= s_t(k) + \log\left(1 + \alpha \cdot e^{n_t(k) - s_t(k)}\right) \\ &= s_t(k) + g\left(\mathbf{s}_t, \mathbf{n}_t, \alpha\right) \end{aligned} \tag{3}$$

3.2 音声特徴量の推定

音声特徴量の推定値として以下の値を考える [2]．

$$\hat{\mathbf{s}}_t = \int \mathbf{s}_t p(\mathbf{s}_t|\mathbf{s}_{t-1}, \mathbf{x}_t, \mathbf{n}_t) d\mathbf{s}_t \tag{4}$$

これを以下のクリーン音声の対数メルフィルタバンクとその Δ の M 混合GMMを使って求める．

$$p(\mathbf{s}_t, \Delta\mathbf{s}_t) = \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{s}_t; \mu_m^{\mathbf{s}}, \Sigma_m^{\mathbf{s}}) \mathcal{N}\left(\Delta\mathbf{s}_t; \mu_m^{\Delta\mathbf{s}}, \Sigma_m^{\Delta\mathbf{s}}\right) \tag{5}$$

$\mu_m^{\mathbf{s}}$, $\mu_m^{\Delta\mathbf{s}}$ は平均ベクトル，$\Sigma_m^{\mathbf{s}}$, $\Sigma_m^{\Delta\mathbf{s}}$ は対角共分散行列である．また本稿では $\Delta\mathbf{s}_t = \mathbf{s}_t - \mathbf{s}_{t-1}$ として扱う．$\mathbf{s}_t$ と $\mathbf{n}_t$，$\mathbf{x}_t$ と $\mathbf{s}_{t-1}$ はそれぞれ無相関であると仮定して式 (4) を変形すると

$$\hat{\mathbf{s}}_t = \frac{\sum_m c_m \int \mathbf{s}_t p(\mathbf{s}_t|m, \mathbf{s}_{t-1}) p(\mathbf{x}_t|\mathbf{s}_t, \mathbf{n}_t) d\mathbf{s}_t}{p(\mathbf{x}_t|\mathbf{n}_t)} \tag{6}$$

この時，$p(\mathbf{s}_t|m, \mathbf{s}_{t-1})$ は以下のような正規分布として表すことができる．

$$p(\mathbf{s}_t|m, \mathbf{s}_{t-1}) = \mathcal{N}\left(\mathbf{s}_t; \nu_m, \Phi_m\right) \tag{7}$$

$$\nu_m = \Sigma_m^{\Delta\mathbf{s}}(\Sigma_m^{\mathbf{s}} + \Sigma_m^{\Delta\mathbf{s}})^{-1}\mu_m^{\mathbf{s}} + \Sigma_m^{\mathbf{s}}(\Sigma_m^{\mathbf{s}} + \Sigma_m^{\Delta\mathbf{s}})^{-1}(\mathbf{s}_{t-1} + \mu_m^{\Delta\mathbf{s}})$$

$$\Phi_m = \Sigma_m^{\mathbf{s}}\Sigma_m^{\Delta\mathbf{s}}(\Sigma_m^{\mathbf{s}} + \Sigma_m^{\Delta\mathbf{s}})^{-1}$$

$p(\mathbf{x}_t|\mathbf{s}_t, \mathbf{n}_t)$ も同様に正規分布と仮定し，

$$p(\mathbf{x}_t|\mathbf{s}_t, \mathbf{n}_t) = \mathcal{N}(\mathbf{x}_t; \mathbf{s}_t + g(\mathbf{s}_t, \mathbf{n}_t, \alpha), \Sigma_x) \tag{8}$$

*Noise reduction considering temporal variation of speech based on GMM. by MIYAKE Nobuyuki, TAKIGUCHI Tetsuya, ARIKI Yasuo (Kobe University)

ここで，$g(\mathbf{s}_t, \mathbf{n}_t, \alpha) \approx g(\mu_m^{\mathbf{s}}, \mathbf{n}_t, \alpha)$, $\Sigma_x \approx \Sigma_m^{\mathbf{s}}$ と近似する．これらを利用し，式 (4) における積分を以下のように近似する．

$$\int \mathbf{s}_t p(\mathbf{s}_t|m, \mathbf{s}_{t-1}) p(\mathbf{x}_t|\mathbf{s}_t, \mathbf{n}_t) d\mathbf{s}_t$$
$$\approx \left[w_1 \cdot \nu_m + w_2(\mathbf{x}_t - g(\mu_m^{\mathbf{s}}, \mathbf{n}_t, \alpha))\right] \cdot N_m(\mathbf{x}_t) \quad (9)$$

本稿では $w_1 = 0.5$, $w_2 = 0.5$ と設定し，さらに $N_m(\mathbf{x}_t) \approx p(\mathbf{x}_t|\mathbf{n}_t, m)$ とする．これは以下のように定義する．

$$p(\mathbf{x}_t|\mathbf{n}_t, m) = \mathcal{N}(\mathbf{x}_t; \mu_m^{\mathbf{x}}, \Sigma_m^{\mathbf{x}}) \quad (10)$$
$$\mu_m^{\mathbf{x}} \approx \mu_m^{\mathbf{s}} + g(\mu_m^{\mathbf{s}}, \mathbf{n}_t, \alpha)$$
$$\Sigma_m^{\mathbf{x}} \approx \Sigma_m^{\mathbf{s}}$$

このことから $p(\mathbf{x}_t|\mathbf{n}_t)$ は以下のように表される．

$$p(\mathbf{x}_t|\mathbf{n}_t) = \sum_m c_m p(\mathbf{x}_t|\mathbf{n}_t, m) \quad (11)$$

$\mathbf{s}_{t-1}$ には前フレームの推定値を，n_t は識別結果から得られた雑音の平均値を利用することで，$\mathbf{s}_t$ の推定値は次の式から求められる．

$$\hat{\mathbf{s}}_t = \sum_m \gamma_m(\mathbf{x}_t)\left[w_1 \cdot \nu_m + w_2(\mathbf{x}_t - g(\mu_m^{\mathbf{s}}, \mathbf{n}_t, \alpha))\right] \quad (12)$$

$$\gamma_m(\mathbf{x}_t) = \frac{c_m p(\mathbf{x}_t|\mathbf{n}_t, m)}{\sum_m c_m p(\mathbf{x}_t|\mathbf{n}_t, m)}$$

3.3 雑音重みの推定

(12) 式において，雑音の強さとして与えた α は未知の値である．従って (12) 式を計算するためには α を何らかの方法で決定する必要がある．本稿では $p(\mathbf{x}_t|\mathbf{n}_t)$ を最大化するように EM アルゴリズムを用いて α の値を決定する．以下のよう関数を設計し，

$$Q(\alpha^{(k)}, \overline{\alpha}) = \sum_m p(\mathbf{x}_t, m|\mathbf{n}_t, \alpha^{(k)}) \log p(\mathbf{x}_t, m|\mathbf{n}_t, \overline{\alpha}) \quad (13)$$

$$\alpha^{(k+1)} = \underset{\overline{\alpha}}{\operatorname{argmax}}\, Q(\alpha^{(k)}, \overline{\alpha}) \quad (14)$$

を収束するまで繰り返すことで目的とする α の値を求めることができる．ここで，$p(\mathbf{x}_t|m, \mathbf{n}_t, \alpha)$ は (10) 式に c_m をかけたものである．こうして求めた α を使用して，(12) 式にしたがって音声特徴量を推定する．

4 実験

ATR の特定話者単語データベースを用いて実験を行った．男女各 2 名を利用し，各話者 2000 発話，計 8000 発話を学習データとして，各話者 500 発話，計 2000 発話をテストデータとして利用した．雑音は RWCP の非音声ドライソース [3] に含まれる 105 種類の雑音を利用した．このデータベースには 1 種類の雑音につき 100 パターンの雑音が収録されており，50 パターンを検出・識別の学習と雑音の平均値の算出に用い，残りの 50 パターンはテスト用に用いた．テストデータには，ランダムに選んだ雑音が 30 ms 〜 250 ms 程度継続するように切り出し，SNR を調整して重畳させた．この時，ひとつの発話に複数個の雑音を重畳させた．SNR は 5 dB, 0 dB, -5 dB に設定し，それぞれの場合について実験を行った．フレーム単位における検出・識別の結果を表 1 に示す．雑音の種類が非常に多いため識別率は低い値となった．この結果を用いて，除去した場合の認識率を評価した．特徴量は MFCC ＋Δ＋ΔΔ 36 次元，音響モデルには 5 状態 4 混合の音素 HMM を用いた．除去時に用いた GMM の混合数は 16, 32, 64, 128 の 4 つであり，最も良かったものを結果として載せている．なお，混合数による差は最大でも 2 % 程度であった．結果を表 2 に示す．ここで，baseline とは除去を行わなかった場合の認識率であり，previous は以前に提案した静的特徴を利用した場合の手法，そして proposed は本稿で提案した手法である．結果を見ると，どの SNR でも以前の手法に比べ認識率が改善していることが分かる．また，雑音が無い場合の音声認識率は 98.8 % であった．

Table 1　雑音の再現率，適合率，識別率 (%)

	5 dB	0 dB	−5 dB
Recall	82.0	89.7	95.2
Precision	82.7	83.1	83.3
Classification	28.3	40.4	47.0

Table 2　単語認識率 (%)

	SNR		
	5 dB	0 dB	−5dB
baseline	61.9	50.5	40.3
previous	81.4	77.8	73.2
proposed	88.2	85.1	80.4

5 おわりに

音声の時間変動を考慮して雑音除去する手法について提案した．動的特徴を利用することで，以前よりも高い効果を得ることができた．

参考文献

[1] 三宅ら，SP2007-100, pp.25-30, 2007-12.
[2] Li *et al.*, Proc. ICASSP2002, pp. 829-832, 2002.
[3] Nakamura *et al.*, Oriental COCOSDA Workshop, 1998.