

雑音重み推定と音声 GMM を用いた既知雑音除去*

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．様々な手法により、非定常な雑音にもロバストな認識が実現できるようになってきた．しかしながら、突発的で短時間しか続かないような雑音であっても認識率の低下が見られるが、このような雑音は追跡することが困難であるため除去することは難しい．我々は、突発的な雑音を検出し、どのような音かを識別することができれば、突発的な雑音の除去につながると考え、雑音の検出と識別についての提案を行った [1]．本稿ではその結果を元に、検出された突発的な雑音の除去法について述べる．検出された雑音の種類が分かるため、雑音スペクトルの形状は事前に保存しておくことができる．しかしその雑音の強さは分からないため、強さを推定する必要がある．本研究では、雑音の強さを表す重みを導入し、その重みの値を推定しながら GMM を用いて雑音を除去する手法について述べる．

2 雑音の検出と識別

クリーン音声に雑音が重畳した場合に、その区間を検出し、どのような雑音であるか判定する [1]．検出、識別とも、事前に用意したクリーン音声と雑音重畳音声によって学習した AdaBoost で行われる．検出はフレームごとに雑音が重畳しているかどうかの判定を行う．雑音が重畳していると判定されたフレームは、識別において、マルチクラスの判別ができるように拡張した AdaBoost を用いて、あらかじめ決められたいくつかのクラスへと分類される．この識別結果を使用することによって、音声に重畳した雑音が既知であるという前提で、除去を考える．

3 提案手法

雑音除去手法として、GMM を用いた手法が提案されている [2]．本研究ではこの手法に雑音の強さの推定を組み込み使用した．以下でその手法について述べる．

3.1 雑音重畳音声

観測信号の k 次元目のメルフィルタバンク特徴量 $X(k)$ はクリーン音声 $S(k)$ 、雑音 $N(k)$ を使って

$$X(k) \approx S(k) + N(k) \quad (1)$$

と書くことができる．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 $N(k)$ の形状を知ることができる．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

$$X(k) \approx S(k) + \alpha \cdot N(k) \quad (2)$$

とする．また、 $X(k), S(k), N(k)$ の対数をそれぞれ $X_{\log}(k), S_{\log}(k), N_{\log}(k)$ として、(2) 式の対数をとって対数メルフィルタバンクに変換すると以下のようになる．

$$\begin{aligned} X_{\log}(k) &= \log \left(e^{S_{\log}(k)} + \alpha \cdot e^{N_{\log}(k)} \right) \\ &= S_{\log}(k) + \log \left(1 + \alpha \cdot e^{N_{\log}(k) - S_{\log}(k)} \right) \\ &= S_{\log}(k) + G(\alpha, N_{\log}(k), S_{\log}(k)) \end{aligned} \quad (3)$$

ここで、右辺の第 2 項を推定することができれば、クリーン音声の対数メルフィルタバンクが推定できる．

3.2 GMM を用いた雑音除去

クリーン音声の対数メルフィルタバンクの M 混合 GMM が平均ベクトル μ_i^S 、対角共分散行列 Σ_i^S を使って以下の通り与えられている．

$$p(S_{\log}) = \sum_{i=1}^M w(i) N(S_{\log}; \mu_i^S, \Sigma_i^S) \quad (4)$$

このとき観測信号 $X_{\log}$ のモデルを同様の混合数の GMM を使用して以下のように仮定する．

$$p(X_{\log}) = \sum_{i=1}^M w(i) N(X_{\log}; \mu_i^X(\alpha), \Sigma_i^X) \quad (5)$$

$$\mu_i^X(\alpha) = \mu_i^S + \mu_i^G(\alpha) \quad (6)$$

$$\mu_i^G(\alpha) = \log(1 + \alpha \cdot \exp(\mu^N - \mu_i^S)) \quad (7)$$

$$\Sigma_i^X = \Sigma_i^S \quad (8)$$

ここで混合重み $w(i)$ は (4) 式と同様の値である． μ^N は雑音の対数メルフィルタバンクの平均値であり、除去を行うフレームが識別されたときにどのクラスに分類されたかによって決定している． $G(\alpha, N_{\log}(k), S_{\log}(k))$ は $p(X_{\log})$ を使って μ_i^G の混

*Noise reduction using estimation of noise weight and clean speech GMM. by MIYAKE Nobuyuki, TAKIGUCHI Tetsuya, ARIKI Yasuo (Kobe University)

合ごとの重み付き平均とする．すなわち以下の式になる．

$$G(\alpha, N_{\log}, S_{\log}) = \frac{\sum_{i=1}^M p_i(X_{\log}) \cdot \mu_i^G(\alpha)}{\sum_{i=1}^M p_i(X_{\log})} \quad (9)$$

$$p_i(X_{\log}) = w(i)N(X_{\log}; \mu_i^X(\alpha), \Sigma_i^X) \quad (10)$$

そして，この G を $X_{\log}$ から引くことでクリーン音声の対数メルフィルバンクを求める．

$$S_{\log}(k) = X_{\log}(k) - G(\alpha, N_{\log}(k), S_{\log}(k)) \quad (11)$$

3.3 雑音重みの推定

(9) 式において， α は未知の値である．従って $G(\alpha, N_{\log}(k), S_{\log}(k))$ を算出するためには α を何らかの方法で決定する必要がある．本稿では $p_i(X_{\log})$ を最大化するように α の値を混合ごとに決定する．したがって α は M 個現れそれぞれを α_i とする．このとき $p_i(X_{\log})$ を最大化する α_i は

$$\frac{\partial}{\partial \alpha_i} \log p_i(X_{\log}) = 0 \quad (12)$$

を解くことで求まる．(10) 式を代入して

$$\frac{\partial}{\partial \alpha_i} \log N(X_{\log}; \mu_i^X(\alpha_i), \Sigma_i^X) = 0 \quad (13)$$

となる．しかし，これは解析的にとくことは難しい．そこで，ここではニュートン法を用いることで決定する．これは

$$f^1(\alpha_{i,t}) = \frac{\partial}{\partial \alpha_{i,t}} \log N(X_{\log}; \mu_i^X(\alpha_{i,t}), \Sigma_i^X) \quad (14)$$

$$f^2(\alpha_{i,t}) = \frac{\partial^2}{\partial \alpha_{i,t}^2} \log N(X_{\log}; \mu_i^X(\alpha_{i,t}), \Sigma_i^X) \quad (15)$$

とおき，以下の演算を収束するまで繰り返す手法である．

$$\alpha_{i,t+1} = \alpha_{i,t} - \frac{f^1(\alpha_{i,t})}{f^2(\alpha_{i,t})} \quad (16)$$

この時の収束値は最適な α_i の近似値になる．こうして求めた α_i を使用して，(9) 式の重み付き平均を算出し，(11) 式にしたがって雑音を除去する．

4 実験

ATR の特定話者単語データベースを用いた．話者数 4，各話者 2720 発話を学習で使用し，500 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] からスピーク音，電話音，紙を破る音，粒子状のものをばら撒く音，ベルの音，ホルン

Table 1 AdaBoost による雑音の検出・識別結果

	再現率	適合率	F値	識別率
0 dB	0.948	0.987	0.967	0.963
5 dB	0.869	0.992	0.926	0.891

Table 2 単語認識率

	SNR	除去なし	16混合	32混合	64混合
重み推定なし	0 dB	57.2	60.8	64.7	63.0
	5 dB	67.1	65.8	65.8	64.1
提案手法	0 dB	57.2	88.6	88.1	88.7
	5 dB	67.1	90.9	91.0	91.5

音の 6 種類の雑音を使用した．このデータベースには 1 種類の雑音につき 50 パターンの雑音が収録されており，30 パターンを検出・識別の学習用と雑音の平均値の算出に用い，残りの 20 パターンはテストに用いた．テストデータには，ランダムで選んだ雑音が 100 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．SNR は 5 dB と 0 dB に設定した．検出・識別の結果を表 1 に示す．表 1 の評価は [1] と異なり，フレーム数を用いた場合の評価になっている．これらの結果を用いて，除去した場合の認識率を評価した．特徴量は MFCC12 次元，音響モデルには 5 状態 4 混合の音素 HMM を用いた．結果を表 2 に示す．この時，「重み推定なし」とは，重み係数 $\alpha = 1$ として GMM による除去を適用したときの結果である．結果を見ると，重みを使用することで認識率は大幅に改善していることが分かる．GMM の混合数による差はあまり見られなかった．

5 おわりに

これまでに提案した，雑音の検出と識別の結果を用いて，雑音を除去する手法について述べた．実験により特定話者の小語彙における認識では大幅な認識率の改善を得ることができた．今後は不特定話者大語彙における実験を行い認識率を評価する予定である．また未知の雑音の除去法についても考案していく予定である．

参考文献

- [1] 三宅，滝口，有木，音講論 (春)，141-142，2007．
- [2] Segura *et al.*, Proc. EuroSpeech, Vol.1, pp. 221-224, 2001．
- [3] Nakamura *et al.*, "Sound Scene Database in Real Acoustic Environments," Oriental COCOSA Workshop, 1998.