Real Adaboostによる音声区間検出*

◎松田博義，滝口哲也，有木康雄 (神戸大)

1 はじめに

雑音下において音声認識を行う際，音声非音声の判定により音声区間検出 (VAD: Voice Activity Detection) を行う必要がある．静かな状況ではゼロクロッシング法などにより区間検出を行うことが可能である．しかし雑音下，特に音声の大部分が雑音に埋もれてしまっているような状況においては，従来の手法では十分な結果を得ることができない．

本稿では，雑音に対するロバストな音声区間検出の手法として，音声非音声の判定に Real Adaboost を用いることを提案する．提案手法の有効性を示すため，音声非音声の判定を行う従来手法の一つとして広く用いられている，GMM(Gaussian Mixture Model) との比較を行う．

2 Real Adaboostによる音声非音声の識別

Adaboost[1] は，単純な識別機を複数組み合わせることによって精度の高い識別器を構成する Boosting 法の中でも顕著な性能を示す手法である．今回は Adaboost をさらに発展させた，Real Adaboost という手法を用いて音声非音声の識別を行った．以下にアルゴリズムの詳細を述べる．

学習データ数を n，繰り返し回数を M とする．これらの値はあらかじめ決定しておく必要がある．学習データを $x_i(i = 1, \cdots, n)$，各データの重みを $w_i(i = 1, \cdots, n)$ とする．x_i としては音声の各フレームから取り出された MFCC(Mel Frequency Cepstrum Coefficient) を用いる．各データ x_i にはあらかじめ $y \in \{-1, +1\}$ を与えておく．すなわちデータ x_i が音声であれば $y = +1$，データ x_i が非音声であれば $y = -1$ とする．

1. 各データの重みを $w_{1i} := \frac{1}{n}$ で初期化する．
2. $m = 1, \cdots, M$ で以下を実行する．
   (a) w_{mi} を確率分布として，x_i から重複を許して n 個，重みつきリサンプリングしたものを x'_i とする．
   (b) x'_i に対して弱識別器 $f_m(x)$ を構成する．弱識別器 $f_m(x)$ は信頼度を生成するものでなければならない．ここでは弱識別器として，CART による 2 分木を用いた [2].
   (c) こうして得られた弱識別器 $f_m(x)$ を用いて $c_m(x_i)$ を得る．

$$c_m(x_i) = \frac{1}{2} \log\left(\frac{f_m(x_i)}{1 - f_m(x_i)}\right) \tag{1}$$

   (d) 各データ x_i の重みを $w_{(m+1)i}$ に更新する．

$$w_{(m+1)i} = \frac{w_{mi} e^{-y_i c_m(x_i)}}{\sum_{r=1}^{n} w_{mr} e^{-y_r c_m(x_r)}} \tag{2}$$

3. 最終的な出力として強識別器 $F(x)$ を得る．

$$F(x) = \sum_{m=1}^{M} c_m(x) \tag{3}$$

通常，sign 関数により出力を $\{-1, +1\}$ にするが，ここでは各フレームごとの信頼値を用いるため，$\sum_{m=1}^{M} c_m(x)$ により算出された値をそのまま用いることにした．

閾値 θ を決定し，入力データ x に対し $F(x) \geq \theta$ であれば音声，$F(x) < \theta$ であれば非音声とする．

Adaboost の重要な性質として，誤ったデータに対するサンプリング重みを増し，次回以降の学習でそれらのデータを重点的に学習する，ということがあげられる．それにより，前段の弱識別器が誤識別を起こしてしまったデータに対しても，後段の弱識別器が正しく識別するため，最終的に正しい識別結果を得ることができる．

3 音声区間検出

音声区間が分断されることを避けるため，式 (4) により，隣接する n フレーム間でスムージングを行う。

$$F'(x_i) = \frac{1}{n} \sum_{j=i-\frac{n}{2}}^{i+\frac{n}{2}} F(x_j) \tag{4}$$

得られた $F'(x_i)$ が，閾値 θ 以上であれば音声，以下であれば非音声とし，暫定的な音声区間を得る．

こうして得られた音声非音声の区間から，連続時間が短いものを取り除くことにより，最終的な音声区間を得る．

---

*Voice Activity Detection with Real Adaboost. by Hiroyoshi Matsuda , Tetsuya Takiguchi, Yasuo Ariki (Kobe Univ.).

4　音声区間検出実験

4.1　実験条件

音声として学習に用いたデータは，研究用音声データベース (ASJ)Vol.1 より，男性 8 名計 1200 文，女性 8 名計 1200 文である．非音声として学習に用いたデータは，車内にて録音された 5 分間程度の走行音である．

実験用に用いたデータは，アイドリング時及び高速道路走行時にて録音された発話データである．どちらも男性 4 名，女性 4 名，各話者 100 発話で計 800 発話である．発話内容は日本各地の地名である．SN 比はアイドリング時でおよそ 10〜25 dB，平均約 17dB，高速道路走行時でおよそ 0〜5 dB，平均約 2 dB である．アイドリング時，高速道路走行時ともに背景雑音として排気音，走行音等が含まれるのみで，音楽，クラクション，ウィンカー音などは含まれていない．

実験は，検出された音声区間があらかじめ与えておいた正解区間より大きいものを正しく検出された区間とした．正解区間と関係の無い区間を検出したものは湧き出しとした．それらを用いて，次式により評価値を計算した．

$$Recall = \frac{\text{正しく検出された区間の総数}}{\text{検出すべき区間の総数}} \quad (5)$$

$$Precision = \frac{\text{正しく検出された区間の総数}}{\text{検出された区間の総数}} \quad (6)$$

すべてのデータは 12,000 Hz にリサンプリングし，低域に集中する車内雑音を取り除くため，カットオフ周波数 200 Hz をもつハイパスフィルタを適用している．

音声の特徴量としては，MFCC を用いる．フレーム幅 32 ms，シフト幅は 8 ms で，CMS を行っている．

なお，Real Adaboost で学習を行なう際，繰り返し回数は 100 回とした．

4.2　比較対象

比較対象として，音声非音声の判定に GMM を用いたもの [3] との比較を行う．[3] より，3 フレーム，5 フレーム，7 フレームをまとめてひとつの特徴量として扱ったものとの比較を行った．GMM は音声として，男性，女性の 2 モデル，非音声として，車内雑音と計 3 つのモデルを使った．各フレームから得られた男性と女性の尤度の内，値が大きいほうを音声の尤度，車内雑音の尤度を非音声の尤度として用る．式 (7) により，得られた尤度比 $L(x)$ が，閾値 θ 以上であれば音声，以下であれば非音声とする．

$$L(x) = \frac{P(x \mid GMM_{speech})}{P(x \mid GMM_{noise})} \quad (7)$$

Table 1　Idling

Idling	Real Adaboost	GMM(3frame)	GMM(5frame)	GMM(7frame)
Recall	98.63 %	98.50 %	98.63 %	98.88 %
Precision	98.87 %	99.24 %	99.25 %	99.50 %

Table 2　Highway

Highway	Real Adaboost	GMM(3frame)	GMM(5frame)	GMM(7frame)
Recall	94.88 %	84.75 %	87.88 %	90.25 %
Precision	95.23 %	95.09 %	95.13 %	95.97 %

ここで $P(x \mid GMM_{speech})$ は音声の尤度，$P(x \mid GMM_{noise})$ は非音声の尤度である．GMM の混合数はすべて 32 混合とした．

4.3　実験結果

アイドリング時，及び高速道路走行時の発話データに対する実験結果を表 1 及び表 2 に示す．

表 1 より，提案手法はアイドリング時のような SN 比の良い状況下において GMM と同程度の識別率を持っていることが確認できる．表 2 より，高速道路走行時のような SN 比の悪い状況において，GMM がアイドリング時に比べ大きく識別率を落としたのに対し，提案手法はアイドリング時と比べ大きな差が無いことから，提案手法の雑音に対する頑健性が確認できる．

5　まとめ

音声区間検出を行う際，音声非音声の判定を Real Adaboost によって行うことを提案した．実環境で収録されたデータに対して区間検出を行い，音声非音声を判定する代表的な手法の一つである GMM と比較することにより，本手法の有効性を確認した．

今後の課題としては，他の音声特徴の使用，音楽，ウィンカー音など走行音以外の雑音が入ったデータでの実験，オフロード走行時などの更に SN 比が悪い状況下での実験などが上げられる．

参考文献

[1] Freund Y.　and Schapire R. E. :,“A decision-theoretic generalization of on-line learning and an application to boosting”, Journal of Comp. and System Sci., 55, pp119-139, (1997).

[2] http://research.graphicon.ru/general-projects/about-us.html.

[3] Norbert Binder, Konstantin Markov, Rainer Gruhn, Satoshi Nakamura:“SPEECH NON-SPEECH SEPARATION WITH GMMS”, 日本音響学会講演論文集 2001 年 10 月, pp141-142.