E-027

汎用・識別的特徴量を用いた音声区間検出
Voice Activity Detection using GA-based Informative Feature

奥田 博也† 田村 哲嗣‡ 速水 悟‡
Hiroya Okuda Satoshi Tamura Satoru Hayamizu

1. はじめに

近年音声認識技術は様々な場面で活躍している．例えば，カーナビゲーションシステムや携帯電話などで用いられている．しかし雑音環境下では，認識率が低下してしまう問題点があり，雑音に対し何らかの前処理を施す必要がある．その 1 つとして音声区間検出(Voice Activity Detection，VAD)が広く用いられている．VAD とは，入力された音声信号を音声区間と非音声区間に分類する手法である．検出された音声区間に対してのみ認識を行うことで，認識性能を改善できる．過去の研究では，MFCC[1]や 3 次キュムラント[2]を特徴量として VAD を行ない，雑音に対する頑健性を検証している．本稿では，VAD に汎用・識別的特徴量(GA-based Informative Feature，GIF)[3]を用いて，検出性能を検証し，考察する．

2. 汎用・識別的な特徴量を用いた音声区間検出

2.1 クラスター分析

後述する特徴量抽出では変換行列をクラスを基に学習で求める必要がある．このため，入力ベクトルについてクラスター分析を行う．本来，VAD では識別するクラスは音声と非音声の 2 クラスだが，GIF への変換ではクラス数を拡張させる必要があるため，音声 3，非音声 2 の計 5 クラスとした．

2.2 汎用・識別的な特徴量

汎用・識別的な特徴量(GIF)は田村らによって提案された特徴量である[3]．入力ベクトルに対し二段階で変換を施すことで，最終的な識別用の特徴量を計算する．はじめに，それぞれのクラスごとに入力ベクトルが該当クラスに属するか否かを判別する二値写像を形成する．そしてこれらを統合することで第一の変換を作成する．次に，入力ベクトルに第一の変換を行って得られた中間ベクトルにおいて，クラスごとに平均を求め，分散が最大になるよう，第二の変換を作成する．同時に，特徴量次元間の無相関化と特徴量次元の削減を行う．これらの変換を得るために，遺伝的アルゴリズム(Genetic Algorithm，GA)を用いる．

N次元の入力特徴量空間を$\boldsymbol{X}$，入力ベクトルを$\boldsymbol{x} \in \boldsymbol{X}$，出力される$M$次の特徴量空間を$\boldsymbol{Z}$，$\boldsymbol{x}$から得られる出力ベクトルを$\boldsymbol{z} \in \boldsymbol{Z}$とする．本稿では，$\boldsymbol{x}$から$\boldsymbol{z}$を得る際に以下のように二段階の変換(線形変換)を行う．図 1 に概略を示す．

第一の変換では$\boldsymbol{x}$からC次元の中間ベクトル$\boldsymbol{y}$を(1)式により計算する(Cは識別すべきクラスの数)．

$$\boldsymbol{y} = \boldsymbol{A} \times (\boldsymbol{x}^T \quad 1)^T \tag{1}$$

Aは$C \times (N+1)$次元の変換行列である．次に第二の変換では，(2)式のように中間ベクトル$\boldsymbol{y}$から次式により識別特徴量$\boldsymbol{z}$を計算する．

$$\boldsymbol{z} = \boldsymbol{B} \times \boldsymbol{y} \tag{2}$$

Bは$M \times C$次元の行列である($1 \leq \mathrm{M} \leq \mathrm{C}$)．本稿では，GIF への入力ベクトルに，フィルタバンク 24 次元を用いた．GIF を用いた VAD の概略を図 2 に示す．

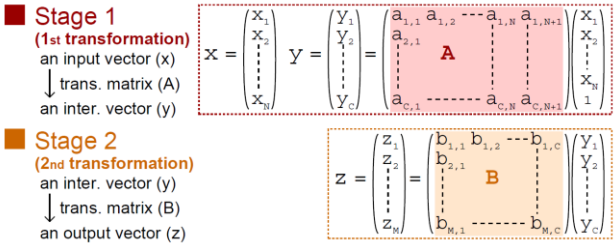

図 1　GIF の概略

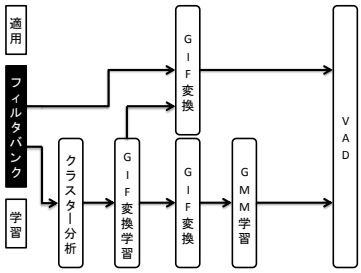

図 2　GIF を用いた VAD の概要

2.3 GMM を用いた VAD アルゴリズム

本稿では，GMM(Gaussian Mixture Model)を用いた VAD を実装した．はじめに，学習データから得られた GIF を用いてモデルの学習を行った．VAD を行う際にはフレームごとに対数尤度$L(i)$を計算し，閾値以上なら音声区間，閾値未満なら非音声区間とした．音声区間の尤度$L_v(i)$，非音声区間の尤度$L_o(i)$は(3)，(4)式のように GMM を用いて計算する．iはフレーム番号，$\boldsymbol{o_i}$はiフレーム目で観測された特徴ベクトル，πは混合重み，Kは混合数，$x = v$またはoである．

$$L_x(i) = \sum_{k=1}^{K} \pi_{k,x} N(\boldsymbol{o_i} | \boldsymbol{\mu}_{k,x} \boldsymbol{\Sigma}_{k,x}) \tag{3}$$

$$L(i) = \log L_v(i) - \log L_o(i) \tag{4}$$

ただし，音声区間は 100ms 以上，非音声は 500ms 以上の連続した区間があるものとする．

CENSREC-1-C[4]のベースラインと同様に，各データについて初期閾値THR_{int}を判別分析法[5]で決定する．その後，初期閾値を用いて(5)，(6)式のように閾値THRを変動させる．

$$THR = THR_{int} + k \cdot \alpha \tag{5}$$

$$\alpha = \frac{L_h - L_l}{40} \tag{6}$$

† 岐阜大学大学院工学研究科
‡ 岐阜大学工学部

ここで，L_hは閾値以上の尤度の平均，L_lは閾値未満の尤度の平均である．kを変動させることで閾値を自動的に決定できる．本稿では$k=-40$，-39，…，39，40とし，81 通りの閾値について VAD を行った．

3. 評価実験

3.1 実験条件

CENSREC-1-C[4]を用いて VAD を行った．CENSREC-1-C は VAD の評価基盤であり，収録されている音声は男性話者，女性話者それぞれ 52 名の合計 104 名による連続数字読み上げである．クリーン音声と人工的に作成したシミュレーション雑音，実環境データが含まれている．本稿では Subway，Babble，Car，Exhibition，Restaurant，Street，Airport，Station のシミュレーション雑音を用い，SNR はそれぞれ 20dB，10dB，5dB，0dB，-5dB とした．

GIF と GMM の学習には，CENSREC-1-C の男性 26 名，女性 26 名の計 52 名分のクリーン音声を用いた．識別は学習と異なる話者 52 名に対して行った．GIF の出力ベクトルは 4 次元で，GMM の混合数は音声，非音声それぞれ 32 である．VAD の評価は 10 ミリ秒ごとにフレーム単位で行い，評価尺度は (7)，(8) 式の FRR(False Rejection Rate) と FAR(False Acceptance Rate) を用いた．雑音別の評価は EER(Equal Error Rate)を用いた．EER は FRR と FAR が等しくなるときの結果である．

$$FRR = \frac{N_{FR}}{N_s} \times 100 \quad (7)$$

$$FAR = \frac{N_{FA}}{N_{ns}} \times 100 \quad (8)$$

N_sは音声フレームの総数，N_{FR}は非音声と検出された音声フレームの数，N_{ns}は非音声フレームの総数，N_{FA}は音声と検出された非音声フレームの数である．

3.2 実験結果

クリーン環境の GIF の特徴量を図 3 に示す．図 3 をみると得られた特徴量は音声区間で 1 次元目の値が高くなっており，VAD に有効であると考えられる．GIF の尤度による VAD は対数パワーのみを用いたベースラインと同程度の識別性能であった．また GIF1 次元目を用いた VAD の追加実験として 2.3 節のように閾値を設けて VAD を行ったところ，ほぼ同等の結果が得られた．1 次元目を用いて閾値処理したときの SNR 別の結果を図 4 に示す．また，表 1 は雑音別の識別結果である．数値は$100-\mathrm{EER}$を表し，大きいほど性能が良い．Babble，Exhibition，Restaurant，Airport 雑音下では，識別率の改善が見られた．GIF は比較的非定常なこれらの雑音下において VAD 性能の向上が見られた．GIF は音声の周波数的な特徴を表現できていると考えられるため，単純なパワーよりも良くなったことが理由として挙げられる．

4. まとめ

本稿では GIF を用いた VAD を提案した．GIF の尤度による VAD と GIF の 1 次元目による VAD の性能を評価した．SNR 別ではベースラインと同程度であるが，雑音の種類によっては性能の改善が見られた．今後は実環境での検出性能の検証，画像特徴量と統合したマルチモーダル音声区間検出への応用が挙げられる．

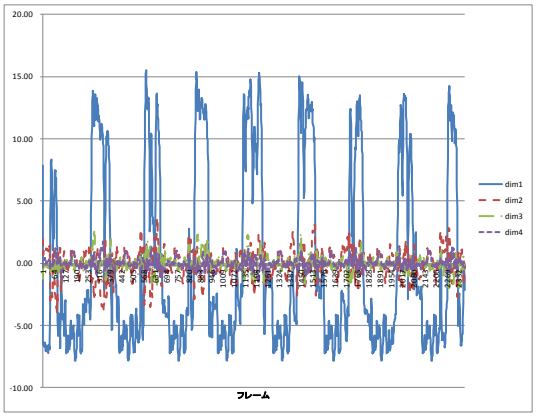

図 3　GIF 変換結果

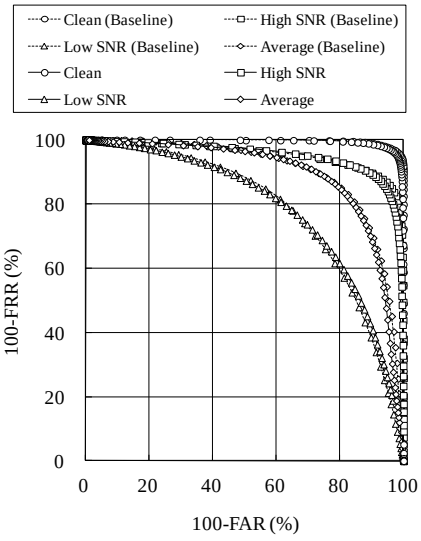

図 4　SNR 別 VAD 結果

表 1　雑音別の識別結果(100-EER)[%]

	Subway	Babble	Car	Exhibition
Baseline	82.10	81.58	85.50	85.2
GIF	81.23	82.35	84.35	87.23
	Restaurant	Street	Airport	Station
Baseline	81.30	83.05	82.35	81.83
GIF	82.93	78.83	84.10	79.28

参考文献

[1] 和田 直哉, 早坂 昇, 宮永 善一, 畑岡 信夫, "メルケプストラムを用いたロバスト音声区間検出", 電子情報通信学会技術研究報告.DSP, ディジタル信号処理, Vol.103, No.146 , pp.25-30, 2003.
[2] 松田 博義, 滝口 哲也, 有木 康夫, "3 次キュムラント音声特徴を用いた音声区間検出",電子情報通信学会技術研究報告. SP, 音声 Vol.106, No.263, pp.37-42, 2006
[3] 田村 哲嗣, 田上 陽嗣, 速水 悟, "GIF-SP:汎用・識別的な特徴量を用いた音声認識性能の改善", 電子情報通信学会技術研究報告, SP2011-92, Vol.111, No.364, pp.119-124, 2008.
[4] 北岡 教秀, 山田 武志, 柘植 覚, 宮島 千代美, 西浦 敬信, 中山 雅人, 傳田 遊亀, 藤本 雅清, 山本 一公, 滝口 哲也, 黒岩 眞吾, 武田 一哉, 中村 哲, "CENSREC-1-C:雑音下音声区間検出評価基盤の構築", 情報処理学会研究報告, 2006-SLP-63, Vol.2006, No.107, pp.1-6, 2006
[5] 大津 展之, "判別および最小 2 乗基準に基づく自動しきい値選定法", 電子通信学会論文誌, Vol.J63-D, No.4, pp.349-356, 1980