E-028

ウェアラブル端末向けシングルチャネル音声区間検知の一検討

A Study on Single-channel Voice Activity Detection for Wearable Device

須藤　隆†
Takashi Sudo

1．はじめに

考えて話すことが脳の活性化に繋がり，認知症予防になると言われている[1]．そこで，手首装着のリストバンド型ウェアラブル端末にマイクを 1 個搭載し、そのウェアラブル端末の装着者だけが発話した時間を音声区間検知と話者照合を併用することで計測して，認知症予防に活用することを目指している．

本研究では，手首装着のリストバンド型ウェアラブル端末でハンズフリーシングルマイク環境での音声区間検知の検討を行ったので報告する．手首装着であるために服の袖による衣擦れ音がマイクに入り込んだり，ウェアラブル端末であるために家庭内での TV 音や施設内での館内放送などの生活雑音がマイクに入り込んだりし，音声区間検知の性能劣化の原因となる．近年，音声信号と音楽信号を識別する研究が進んでおり，放送コンテンツのメディア処理や音声・音楽符号化技術に活用されている[2,3,4].

そこで，これらの音楽が含まれる生活雑音に対してロバストにするために，従来は放送コンテンツのメディア処理向けに音声と音楽を区別するために用いている特徴量 5 種類(零交差数の分散(VZC)，Spectral Flux(SF)，Spectral Flux の分散(VSF)，Cepstrum Flux(CF)，Block Cepstrum Flux(BCF))[2]を，音声区間と非音声区間を区別する音声区間検知に適用し，ウェアラブル端末向けとして処理量が小さい線形識別関数による音声区間検知を用いて，有効性を評価した．

2．音声区間検知の従来特徴量

従来，音声区間検知で音声区間と非音声区間を区別するための特徴量は，様々な特徴量が用いられており[3,5]，零交差数やパワーの古典的な特徴量，周波数スペクトルの概形・基本周波数・調波成分・線スペクトル周波数・メル周波数ケプストラム係数(MFCC)など音声の性質を利用した特徴量，SNR など雑音の情報を利用した特徴量，4Hz 変調エネルギーなど時間情報を利用した特徴量に大別される[3].

音声信号と音楽信号は音響的な特徴が似通っているため，従来の音声区間検知で用いられている特徴量では，音楽が含まれる環境雑音向けの音声区間検知は困難な場合がある．

3．ウェアラブル端末向け音声区間検知

本研究では，従来研究対象としては検討されて来なかった，手首装着のリストバンド型ウェアラブル端末に搭載したマイクでのシングルチャネル音声区間検知法を提案する．

手首装着のウェアラブル端末に搭載したマイクには，TV 音や施設内での館内放送など音楽が含まれる雑音が入るため，音声区間検知の性能改善として，音声と音楽を区別するために用いている特徴量を用いることを提案する．

†（株）東芝　研究開発センター, Toshiba Corporation Corporate Research & Development Center

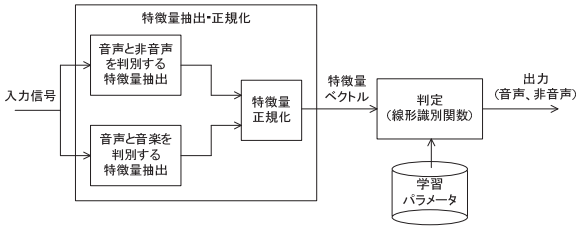

図 1：提案法の構成図

また，判別にはウェアラブル端末で実現可能な小さい処理量の線形識別関数[4]を用いる．図 1 に構成図を示す.

3.1　特徴量

音声と音楽を区別するために用いていた特徴量 5 種類(VZC，SF，VSF，CF，BCF)[2]を音声区間検知に用いる．また，音声区間と非音声区間を区別するために用いられていた従来特徴量も併せて利用する．

VZC は，複数のフレームから構成されるブロックと呼ぶ単位内での零交差数の分散である．SF は，隣り合ったフレーム間における振幅スペクトルのユークリッドノルムであり，VSF はブロック内での SF の分散である．CF は，現フレームの LPC ケプストラムベクトルと，現フレームから時間的に先行する複数のフレームにおける LPC ケプストラムベクトルとのユークリッドノルムの平均値であり，BCF はブロック内での CF の平均値である．なお，すべての特徴量は正規化して用いる．

3.2　線形識別関数による学習と判別

学習データを用いて事前に回帰分析をして，重回帰分析式を線形識別関数とする線形判別[4]を利用する．

学習データに用いるk番目の入力パラメータセット$\boldsymbol{x}^k$を，抽出したn個の特徴量のベクトルとして式(1)で表す．

$$\boldsymbol{x}^k = (x_1{}^k, x_2{}^k, \cdots, x_n{}^k) \quad \cdots (1)$$

音声区間を 1，非音声区間を-1 として，使用する学習データの正解区間を人手でラベル付けしておき，入力パラメータセット$\boldsymbol{x}^k$が属する正解区間y^kを式(2)で表す．

$$y^k = \begin{cases} -1, & \text{非音声区間} \\ 1, & \text{音声区間} \end{cases} \quad \cdots (2)$$

学習データであるN個の入力パラメータセット$\boldsymbol{x}^k$に対して，式(3)の評価値$f(\boldsymbol{x})$と正解区間y^kとの二乗誤差をそれぞれ求める．それらの二乗誤差の和Eを式(4)と求め，Eが最小となる重回帰分析式から係数ベクトル$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_n)$を求める．

$$f(\boldsymbol{x}) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \quad \cdots (3)$$

$$E = \sum_{k=1}^{N} \{y^k - f(\boldsymbol{x}^k)\}^2 \quad \cdots (4)$$

学習によって決定した係数ベクトル$\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_n)$を用いて式(3)によって評価値$f(\boldsymbol{x})$をフレーム単位で計算し，$f(\boldsymbol{x}) > 0$ならば音声区間，$f(\boldsymbol{x}) < 0$ならば非音声区間と判定する．なお，ブロック単位で算出する特徴量を考慮して，ブロック分の処理遅延を持たせている．

音声区間判定に必要な処理は式(3)のみであり，k-NN や GMM 等の学習による他の判定処理と比較して，処理量はウェアラブル端末で実現可能な程度に小さい．

4. 評価

4.1 評価データ

音声区間検知の基本性能を評価するために，CENSREC-1-C 評価環境[6]から，雑音重畳していない約 30 秒の音声データ 104 名分を評価および学習音声データとし，ノイズ 8 種類(Subway, Babble, Car, Exhibition, Restaurant, Street, Airport, Station)を評価ノイズデータとして利用した．

また，手首装着のウェアラブル端末でのハンズフリーシングルマイク環境を模擬するために，InvenSense 社製 MEMS マイク ADMP441 を用いたレコーダを片腕の手首に装着して，打撃音・拍手音・衣擦れ音・TV 音・風切音・館内放送をサンプリング周波数 8kHz で 3 人分集音し，評価ノイズデータとして利用した．打撃音は複数の部屋で壁や机を叩いた音であり，拍手音は両手で拍手をする音であり，衣擦れ音は長袖服を着て室内・屋外を歩行した際の音であり，TV 音は室内で録画したニュース番組・バライティ番組・ドラマを視聴したときの音であり，風切音は風の強い複数の日に静止・歩行した際の音であり，館内放送は商業施設で静止・歩行した際の音である．

口元から手首装着マイクまでの距離とそれが動くことも考慮して，評価音声データと評価ノイズデータを SNR が 20dB，10dB，5dB になるように重畳させて評価および学習データを作成した．音声データの 2 名分を学習に，102 名分を評価に用いた．

4.2 評価方法

音声区間検知で音声区間と非音声区間を区別するために従来用いられている特徴量 5 種類(フレームパワー(RMS)[3], スペクトルエントロピー(SE)[5]，短時間 SNR(SNR)[3]，メル尺度での短時間 SNR(Mel-SNR)[5]，スペクトル間余弦値(SC)[5])だけを利用して，線形識別関数による音声区間検知をする方式(Baseline)と，これらに音声と音楽を区別するために用いていた特徴量 5 種類を併用する提案法との比較評価を実施する．なお，1 フレームを 20ms とし，短時間 SNR は，所定区間での入力信号のフレームパワーの最小値をフロアノイズパワーとし，入力信号のフレームパワーとフロアノイズパワーの比で求める．また，メル尺度での短時間 SNR では音声が入力されていない区間の平均スペクトルから推定する．

評価方法は，CENSREC-1-C 評価環境[6]を利用し，すべての評価データに対してフレーム単位での音声区間検出を求め，平均等価エラー率(EER)で評価する．

4.3 評価結果

ノイズ種類ごとの平均 EER として評価結果を図 2 に示す．CENSREC-1-C 評価環境[6]の 8 種類ノイズは，平均として 1 つにまとめた．提案法は Baseline と比較して，平均 EER が 1.5～29.4%低下し，すべての条件で低下した．平均 EER は SNR20dB で平均 8.8%，SNR10dB で平均 6.5%，SNR5dB で平均 6.9%低下した．平均 EER は，TV 音では平均 2.7%，館内方法では平均 7.3%低下した．

また，SNR10dB のときの各特徴量と正解区間y^kとの相関係数の絶対値を図 3 に示す．BCF は拍手音・TV 音・館内放送と，SF は TV 音と相関係数の絶対値が 0.6 以上と高く，EER 低下に寄与したと考える．

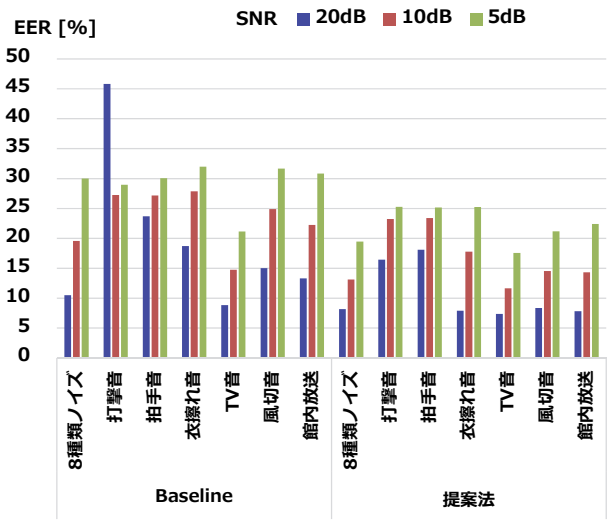

図 2：平均 EER による評価結果

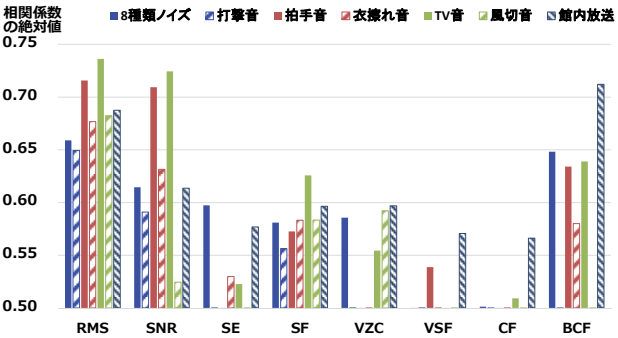

図 3：相関係数の絶対値による評価結果

5. まとめ

手首装着のウェアラブル端末に搭載したマイクでのシングルチャネル音声区間検知を検討した．音声と音楽を区別するために用いていた特徴量 5 種類をウェアラブル端末で想定される生活雑音に対して検討し，BCF は拍手音・TV 音・館内放送と，SF は TV 音と相関係数の絶対値が高くなった．これら特徴量を併用した線形識別関数による音声区間検知の結果，平均 EER が平均 7.4%改善できた．

参考文献

[1] 大武ら,``認知症予防を目的とする共想法による会話活性度の解析と評価,'' 日本機械学会 福祉工学シンポ, MF142, Oct. 2007.
[2]谷口ら, ``音声・音楽識別を目的とした特徴量の検討,'' 信学技報 SP2002-135, Dec. 2002.
[3] 石塚ら, ``音声区間検出技術の最近の研究動向,'' 日本音響学会誌, Vol.65, No.10, 2009.
[4] 米久保ら,``背景音にロバストな音声・音楽信号の識別方式の検討,'' FIT2008, E-021, 2008.
[5] 山本ら, ``雑音にロバストな音声と非音声の判別技術,'' 東芝レビュー, Vol.64, No.12, 2009.
[6] IPSJ SIG-SLP, ``雑音下音声区間検出評価環境(CENSREC-1-C),'' 2006.