# 変調フィルタリングを用いた映画中の音声区間検出*

○向 奈津美，北口 直，金寺 登（石川高専），荒井 隆行，藤樫 佑樹，古賀 綾子（上智大・理工），
△吉井 順子（フジヤマ），船田 哲男（金沢大）

## 1 はじめに

映画や動画コンテンツ等の字幕作成作業には音声区間検出が必須である。音声区間を自動検出するためには，映画中に含まれる様々な背景雑音や BGM などの対処が必要である。

音声区間の検出には，パワーやゼロ交差数を用いた方法[1] ,線形予測に基づく方法[2] など数多く提案されている。また，音声と音楽の識別に関しては，4Hz 付近の変調エネルギーを利用した方法などが報告されている[3]。

本報告では，変調スペクトル成分中で，1～16Hz の変調周波数バンドにほとんどの音声認識情報が存在するという実験結果に基づき[4,5]，変調フィルタリングを施したケプストラムの変動情報を用いて映画中の音声区間検出を行った結果を報告する。

## 2 音声区間検出特徴量

### 2.1 変調フィルタリングを用いた音声区間検出特徴量

Fig. 1 に示すように，まず音声から PLP ケプストラムの時間軌跡を求める。次に，PLP ケプストラムの一部（$C_L$～$C_H$）に対して変調フィルタリング(バンドパスフィルタリング)を行う。変調フィルタリングした各 PLP ケプストラム $C_L$～$C_H$ に対して一定区間（20 フレーム）における標準偏差（$\sigma_L$～$\sigma_H$）を求める。さらに,各標準偏差の和を特徴量とする。

### 2.2 高域スペクトルの傾き

予備調査の結果，雑音と音楽の高域スペクトルの傾きは平坦であることが多かった。一方，音声の高域スペクトルは右下がりである傾向が見受けられた。そこで，4～6kHz のスペクトルの傾きを求め，-1 倍したものを音声区間検出特徴量とする。

## 3 評価実験

### 3.1 実験条件

評価実験には，ミュージカル映画(英語版)の 1406s(約 23 分)を使用した。300ms 以上非音声区間が継続した場合に音声区間終了とみなして正解データを作成した。全音声区間数は 299 区間，正解音声時間は 615s であった。

評価結果を以下に示す誤り率，再現率，適合率で算出した。

$$(\text{音声・非音声})\text{誤り率} = \frac{\text{誤った音声・非音声フレーム数}}{\text{音声・非音声フレーム数}}$$

$$(\text{音声・非音声})\text{再現率} = \frac{\text{正解した音声・非音声フレーム数}}{\text{音声・非音声フレーム数}}$$

$$(\text{音声・非音声})\text{適合率} = \frac{\text{正解した音声・非音声　フレーム数}}{\text{出力結果での音声・非　音声フレーム数}}$$

### 3.2 ケプストラム係数の選択

本報告では 2.1 節で述べたように，PLP ケプストラムの一部（$C_L$～$C_H$）を選択して利用する。$C_L$，$C_H$ を変化させた場合の総誤り率を Table 1 に示す。なお，変調フィルタリング条件は，総誤り率が最も低くなるものを使用した。Table 1 の結果より $C_1$～$C_7$ を用いた場合の誤り率が最も低かった。従って，以下の実

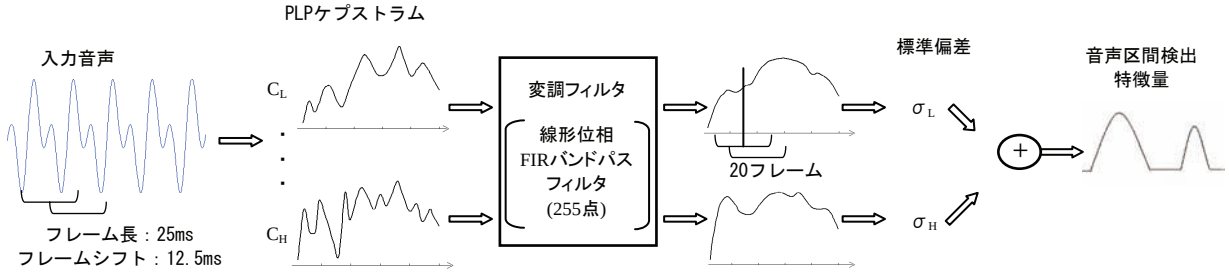


Fig. 1　変調フィルタリングを用いた音声区間検出特徴量

* Voice activity detection using modulation filtering for captioning movie by N. Mukai, S. Kitaguchi, N. Kanedera (Ishikawa National College of Technology), T. Arai, Y. Fujikashi, A. Koga (Sophia University), J. Yoshii (Fujiyama Inc.), T. Funada (Kanazawa University) .

Table 1　ケプストラム係数 $C_L$～$C_H$ を用いた場合の総誤り率[%]

| $C_L$ \ $C_H$ | $C_0$ | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $C_6$ | $C_7$ | $C_8$ |
|---|---|---|---|---|---|---|---|---|---|
| $C_0$ | 21.85 | 20.90 | 19.69 | 19.30 | 18.38 | 17.89 | 17.66 | 17.35 | 36.26 |
| $C_1$ | | 22.25 | 18.63 | 18.38 | 17.57 | 16.78 | 16.85 | 16.65 | 16.70 |
| $C_2$ | | | 20.86 | 19.56 | 18.12 | 17.18 | 17.33 | 17.08 | 17.20 |
| $C_3$ | | | | 22.96 | 19.74 | 17.74 | 17.72 | 17.51 | 17.64 |
| $C_4$ | | | | | 21.38 | 18.11 | 18.09 | 17.87 | 17.92 |
| $C_5$ | | | | | | 20.27 | 20.24 | 19.16 | 18.60 |
| $C_6$ | | | | | | | 26.45 | 22.14 | 21.01 |
| $C_7$ | | | | | | | | 23.32 | 22.40 |
| $C_8$ | | | | | | | | | 27.08 |

Table 2　変調フィルタリング条件を変化させた場合の総誤り率[%]

| $F_L$[Hz] \ $F_H$[Hz] | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 16 | 100 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 28.41 | 22.35 | 20.00 | 18.95 | 17.53 | 17.10 | 16.72 | 16.65 | 16.77 | 16.73 | 16.98 | 17.54 |
| 1 | | 21.97 | 19.73 | 18.78 | 17.55 | 16.97 | 16.98 | 16.95 | 17.02 | 17.09 | 17.30 | 17.76 |
| 2 | | | 21.07 | 19.57 | 18.25 | 17.97 | 17.89 | 17.97 | 17.99 | 18.07 | 18.40 | 18.92 |
| 3 | | | | 21.79 | 19.59 | 18.88 | 18.48 | 18.62 | 18.82 | 18.88 | 19.14 | 19.97 |
| 4 | | | | | 21.88 | 20.01 | 19.86 | 19.70 | 19.86 | 19.86 | 20.02 | 21.26 |
| 5 | | | | | | 23.16 | 22.01 | 21.18 | 21.16 | 21.14 | 21.40 | 22.78 |
| 6 | | | | | | | 24.83 | 23.05 | 22.53 | 22.34 | 22.24 | 24.05 |
| 7 | | | | | | | | 26.29 | 25.07 | 23.48 | 23.31 | 25.36 |
| 8 | | | | | | | | | 27.02 | 25.43 | 23.77 | 26.89 |
| 9 | | | | | | | | | | 26.60 | 24.78 | 28.54 |
| 10 | | | | | | | | | | | 25.98 | 30.21 |
| 16 | | | | | | | | | | | | 37.77 |

験では $C_1$～$C_7$ を選択して使用する。

### 3.3 変調フィルタリング条件

変調フィルタリングを変化させた場合の総誤り率を Table 2 に示す。ここで Table 2 中の $F_L$ は変調バンドパスフィルタの低域遮断周波数，$F_H$ は変調バンドパスフィルタの高域遮断周波数を表している。Table 2 より 0～8Hz の総誤り率が最も低かった。これは, 文献[4,5] の変調周波数バンドの 1～16Hz, 特に 2～8Hz にほとんどの音声認識情報が存在するという結果に符合している。

### 3.4 評価結果

Table 3 に音声区間抽出特徴量として，短時間パワー，高域スペクトルの傾き，変調フィルタリングによる特徴量，変調フィルタリングによる特徴量と高域スペクトルの傾きの重み和を用いた場合の評価結果を示す。Table 3 より，変調フィルタリングによる特徴量，高域スペクトルの傾き，短時間パワーの順で総誤り率が低かった。さらに，変調フィルタリングによる特徴量と高域スペクトルの傾きの重み和を用いた場合，総誤り率がわずかながら改善された。

Table 3 評価結果[%]

| | 総誤り率 | 音声誤り率 | 音声適合率 | 非音声誤り率 | 非音声適合率 |
|---|---|---|---|---|---|
| パワー | 38.42 | 27.68 | 54.60 | 46.77 | 71.20 |
| 傾き | 24.33 | 26.34 | 71.57 | 22.76 | 79.03 |
| 変調 | 16.66 | 25.52 | 85.58 | 9.76 | 81.97 |
| 変調+傾き | 16.05 | 17.41 | 81.07 | 15.00 | 86.26 |

## 4　おわりに

映画中の音声区間検出方法として変調フィルタリングを用いた方法と高域スペクトルの傾きを用いる方法を提案した。今後は，誤り率を改善するために，正しく音声区間を検出できなかった部分の分析を行う予定である。

### 参考文献


[1] L.R.Rabiner, M.R.Sambur, BSTJ, 54, (2), 287-315,1975.

[2] 藤樫佑樹，古賀綾子，荒井隆行，金寺 登，吉井順子，音講論(秋), 33-34, 2005.

[3] E. Scheirer and M. Slaney, ICASSP-97, 1331-1334, 1997.

[4] 金寺 登，荒井隆行，船田哲男，電子情報通信学会論文誌 D-Ⅱ, J84-D-Ⅱ(7), 1261-1269, 2001.

[5] N. Kanedera, T. Arai, H. Hermansky, M. Pavel, Speech Communication, 28, 43-55, 1999.