2L-04

料理映像の索引付けのための音響解析手法の検討*

須場 康貴[†,‡], 浜田 玲子[††], 井手 一郎[†††], 坂井 修一[‡], 田中 英彦[‡]

[†,††,‡]{yas, reiko, sakai, tanaka}@mtl.t.u-tokyo.ac.jp , [†††]ide@nii.ac.jp

[†] 日立工業専門学院, [††] 東京大学大学院工学系研究科

[†††] 国立情報学研究所, [‡] 東京大学大学院情報理工学系研究科

1 はじめに

近年のマルチメディアデータの増大に伴い、その解析がますます重要となりつつある。そのため複数メディアを統合的に処理する手法が注目されている。

我々は、このような統合メディア処理手法の研究の一環として、料理映像を題材とした研究を行っている。浜田 [1] や三浦 [2] は、料理映像中の動画のみを用いて映像とテキスト教材を自動で対応づけたが、その正解率は60〜85%であり正確性に欠ける。またこの精度は、対象となる映像によって非常に流動的であったと報告されている。

そこで本稿では、従来では検討されていなかった映像中の音データを解析し、それらの自動検出を行うための音響解析手法について検討する。将来的にはその解析結果を用いて、料理映像とテキスト教材の対応付けの補助を行い、統合メディア処理システムの精度向上かつ安定化を目指す。

2 調理音の検出

料理映像には、一般的に対応するテキスト教材が存在することが多い。料理テキスト中の調理動作を表す動詞部分と料理映像中の調理動作を対応づけるため、調理動作時に生じる調理音に着目し、これを検出する手法を検討する。

2.1 調理音の分類

料理番組中の調理音を、その特徴から以下の 3 種類に分類した。

- 連続性音…焼く音、揚げる音など時間的に連続した広帯域性雑音
- インパルス性音…切る音、器具がぶつかる音などインパルス性の強い雑音
- ランダム性音…混ぜる音、その他複雑な音などランダム的な雑音

我々はこれら各々の調理音について、その検出手法を検討している。本稿では、このうち連続性音の検出手法について述べる。

2.2 連続性音の検出

2.2.1 検出手法

連続性音は時間的に連続した広帯域性雑音である。その特徴として、(1) 音圧レベルが可聴周波数のほぼ全域に分布、(2) 高音域の音圧レベルが比較的大きい、(3) 区間幅は数秒〜数十秒以上、などが挙げられる。これらの特徴をふまえ、以下の検出手法を提案する。

1. 映像から取り出した音響ストリームを微小区間 s に時間細分化し、各々について高速フーリエ変換をかける。時刻 t での周波数 f の強度を $i(t, f)$ とする。
2. 音声帯域を排除するためにカットオフ周波数 f_c のハイパスフィルタ（HPF）をかける。
3. HPF を通過した強度 $i(t, f)|_{f>f_c}$ を積算し、これを I_n とする。
4. 定区間 τ 内での I_n の最低値が閾値 T 以上の時、これを広帯域雑音部分と見なし、抽出する。但し、I_{n-1} と I_n、および I_n と I_{n+1} が共に閾値をまたいでいる場合、I_n については考慮しないものとする。

---

*"Audio analysis method for cooking video indexing"

Yasutaka Suba[†,‡], Reiko Hamada[††], Ichiro Ide[†††]
Shuichi Sakai[‡], Hidehiko Tanaka[‡]
[†]Hitachi Technical College
2-17-2 Nishinarusawa-cho, Hitachi-shi, Ibaraki 316-0032, Japan
[‡]Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[††]Graduate School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
[†††]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

2.2.2 検出実験

料理映像 9 本（計 63 分）について、上記手法を用いて連続性音検出実験を行った。

実験に用いたデータの諸元を表 1 に示す。

表 1: 実験データ諸元

対象番組	キュービー 3 分クッキング
映像時間	7 分 ×9 本 =63 分
オーディオ形式	WAV 形式
サンプリングレート	44,100 sample/sec
チャンネル	モノラル

また、前節提案手法で示した各要素値を下記のように選定した。

$$s = \frac{2,048[\text{sample}]}{44,100[\text{sample/sec}]} \approx 46.4 \times 10^{-3}[\text{sec}]$$

$$f_c = 5,000[\text{Hz}], m = 20, \tau = s \times m \approx 0.93[\text{sec}]$$

$$T = \frac{\sum_N \min\left(I_{mn}, I_{mn+1}, \ldots, I_{mn+m-1}\right)}{N}$$

$$N = \frac{\text{Total Time}}{m}$$

表 2 および図 1 に検出結果を示す。なお正解区間は、料理映像を見て、画像または音から連続性音が発生している区間とみえる部分を人手で決定した。

表 2: 連続性音検出結果

正解時間	正検出	誤検出	検出洩れ	再現率	適合率
1021 [sec]	937 [sec]	42 [sec]	84 [sec]	92%	96%

$$\left(\text{再現率} = \frac{\text{正検出}}{\text{正検出} + \text{検出洩れ}}, \text{適合率} = \frac{\text{正検出}}{\text{正検出} + \text{誤検出}}\right)$$

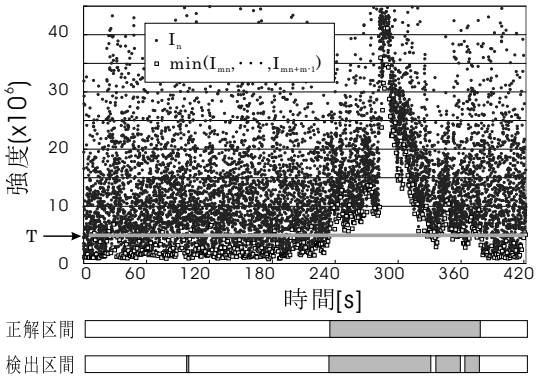

図 1: 検出結果のグラフ（一例）

実験の結果、再現率 92%、適合率 96%と非常に良好な結果が得られた。

誤検出の主な原因は、人の音声の倍音や時間幅の広いインパルス性音を検出したことである。また検出洩れの主な原因は、連続性音のレベル低下によるものである。しかし、再現率、適合率共に高精度であり、本手法は有効であると考えられる。

3 今後の課題

今後は、連続性音検出手法の精度向上を目指すと同時に、他の 2 種類の調理音（インパルス性音、ランダム性音）の検出手法の検討を進める。

連続性音の検出について、本稿では閾値 T を用いて評価した。しかし将来索引付けやデータベース作成などを行うにあたり、閾値 T を連続性音の検出基準として用いないことにする。これは、複数の調理音が同時に発せられた場合の誤検出や検出洩れを防ぐためである。今後は、連続性、インパルス性、ランダム性のそれぞれの調理音の度合いを総合的に評価し、音データと映像の対応付けを行う。

また、本稿では調理音を 3 種類に大別したが、更に細かく分類し、各々の調理動作や材料などに準じた調理音の検出手法も検討する。更に、音声の倍音などの誤検出防止のため、音声とサウンドの分離などについても検討していく方針である。

4 おわりに

本稿では、料理映像における音響解析手法の一つとして、時間的連続性のある広帯域な調理音の検出手法を提案した。また、提案手法を用いて連続性調理音の検出実験を行い、良好な結果を得ることができた。

今後は、連続性音検出手法の精度向上、他の調理音の検出手法、更に細かい分類での調理音の検出手法、また音声とサウンドの分離手法などについて検討する。更に将来は、これらの解析結果を利用した索引付けやデータベース作成など、様々な応用が考えられる。

参考文献

[1] R. Hamada, I. Ide, S. Sakai, H. Tanaka: ”Associating Cooking Video with Related Textbook”, Proc. ACM Multimedia 2000 Workshops, pp.237-241, Nov. 2000.

[2] 三浦宏一, 浜田玲子, 井手一郎, 坂井修一, 田中英彦: “ 料理映像の構造解析による手順との対応づけ ”, 第 62 回情報処理学会全国大会, No6R-9, Vol.3, pp.31-32, Mar. 2001.