K-36

音声強調に着目したマルチメディアコンテンツ要約技術

A New Multimedia Content Summarization Technique based on Automatic Speech Emphasis Extraction

◎日高 浩太† 水野 理†† 中嶌 信弥†
Kota Hidaka Osamu Mizuno Shinya Nakajima

1. はじめに

マルチメディアコンテンツの増加に伴い、短時間でコンテンツを視聴可能な要約技術が求められている。メタデータ付与[1]による要約もあるが、大量のコンテンツの要約には自動化が必要となる。マルチメディアコンテンツに含まれる音声情報を利用した要約は、音声コンテンツに加え、映像コンテンツにも適応できる利点がある。既往の技術には、音声認識を用いた手法[2]や、キーワード検出による手法[3][4]があるが、実環境の音声での認識精度、検出精度の課題がある。本稿では、音声の強調部分（音声強調）に着目して要約する方法を提案する。イントネーションの“形”に着目して音声強調を抽出する手法[5]があるが、実環境での抽出が課題である。本提案方法では、音声の高さ、強さ、速さに対応する特徴量を統計的に分析し、音声が強調となる確率（強調確率）と平静となる確率（平静確率）を求め、音声強調度として抽出する。多種多様な音声を対象に、音声強調度を基にして、ユーザの指定した任意の長さにマルチメディアコンテンツを要約する技術について報告する。

2. マルチメディアコンテンツ要約方法

2-1. 要約方法概要

本稿では、音声強調を基に要約する。音声強調のみを要約に用いると聴取し難いため、音声強調を含むある程度の長さを再生区間として用いる。図 1 にマルチメディアコンテンツ要約の処理を示す。①音声強調を抽出するために、分析する区間として音声小段落を定義する。聴取し易い区間として音声段落を定義する。音声の無声区間から音声小段落を、パワから音声段落を抽出する。②音声のピッチ（高さ）、パワ（強さ）、スペクトル変化量（速さ）を抽出し、音声小段落の音声が強調となる確率（強調確率）、平静となる確率（平静確率）を求める。③強調度を定義し、強調度を基に任意の長さの要約を生成し、要約音声を再生する。

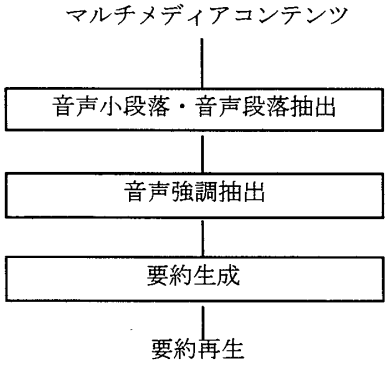

図 1 マルチメディアコンテンツ要約処理

† NTTサイバーソリューション研究所
†† NTTサイバースペース研究所

2-2. 音声段落抽出方法

音声小段落は、音声が Ts 以上の無声区間を両端とする音声区間と定義する。音声段落は、音声小段落の平均パワと、音声小段落の後半部分の平均パワとの比が、閾値以上である音声小段落を両端とする音声区間とする。

音声小段落の平均パワが P、音声小段落が n 個の有声区間で構成され、有声区間の平均パワが $p_j(j=1,2,...,q)$ であるとき、定数α、βを用いて

$$\frac{\sum_{j=q-\alpha}^{q} p_j}{P} < \beta \quad (1)$$

を満たす音声小段落を両端とする音声区間を音声段落と定義する。

2-3. 音声強調抽出方法

音声強調を抽出するために、音声特徴量から音声の強調確率と平静確率を求める。

学習用データを作成するために、音声資料に対し音声の強調部分と平静部分にラベルを付与した（強調ラベル、平静ラベル）。作業者が音声を聴取し、音声が強調、もしくは平静と感じた区間をラベリングする。強調ラベル区間、平静ラベル区間の音声特徴量（ピッチ、パワ、スペクトル変化量）を LBG 法で、ベクトル量子化し、一つのコードブックを作成する。強調確率、平静確率の抽出区間 X がフレーム数 n から構成され、量子化された音声特徴量のコードが $C_l(l=1,2,...n)$ の場合の強調確率 P_{Xemp}、平静確率 P_{Xnrm} について述べる。フレーム f のコードが C_f のとき、フレーム f の強調確率 $P_E(f)$、平静確率 $P_N(f)$ は

$$P_E(f)=\lambda_{exp1}p_{emp}(C_f|C_{f-1}C_{f-2})+\lambda_{exp2}p_{emp}(C_f|C_{f-1})+\lambda_{exp3}p_{emp}(C_f) \quad (2)$$

$$P_N(f)=\lambda_{nrm1}p_{nrm}(C_f|C_{f-1}C_{f-2})+\lambda_{nrm2}p_{nrm}(C_f|C_{f-1})+\lambda_{nrm3}p_{nrm}(C_f) \quad (3)$$

の形で線形補完して求める。重み係数λ_{empi}、$\lambda_{nrmi}(i=1,2,3)$は学習用データから削除補間法で推定する。任意のフレームで式(2)、式(3)を求め、強調確率 P_{Xemp}、平静確率 P_{Xnrm} は

$$P_{Xemp}=\prod_l P_E(l) \quad (4)$$

$$P_{Xnrm}=\prod_l P_N(l) \quad (5)$$

となる。

2-4. 音声強調抽出実験と結果

本手法を用い、強調抽出実験を行った。音声資料として、会話、会議、講演、テレビ番組、映画の音声（合計約 2780 分）を用いて学習用データと評価データを作成した。学習用データは強調ラベル 1956 区間、平静ラベル 1955 区間であり、評価データは強調ラベル 1711 区間、平静ラベル 1712 区間である。ラベル区間は平均 1.45 秒で、学習用データは約 58 分であった。音声資料の音声特徴量を抽出し、学習用データを用いてコードブックサイズ 32 のコードブックを作成した。

音声特徴量について報告する。自己相関法により基本周波数 $f0$、パワ p を求める。分析フレーム長 50ms、フレーム周期 50ms とした。自己相関係数が 0.7 以上の $f0$、p を用い、フレーム区間の平均を求め、それぞれ差成分を求める。発話速度に対応する特徴として、現フレームを中心に幅 1s 内の動的尺度[6]のピーク本数 dp を計測する。これと、現フレームの開始、終了時刻の前後 450ms を中心

表1　音声強調抽出実験結果

	強調ラベル		平静ラベル	
	再現率	適合率	再現率	適合率
Close	73%	75%	76%	74%
Open	74%	77%	77%	74%

とする幅 1s 内の dp との差成分を求める。

学習用データの各ラベルについて式(4)、(5)の強調確率と平静確率を比較し、被験者の設定したラベルとの再現率、適合率で評価した(close 実験)。評価データの各ラベルについて同様の実験を行い評価した(open 実験)。表 1 に結果をまとめる。両実験の再現率、適合率で 70%以上であった。

2-5.　任意の長さの要約方法

要約生成においては、任意の音声小段落 S（フレーム数 L）の強調確率 P_{Semp}、平静確率 P_{Snrm} を式(4)、(5)から求める。図 2 に会話音声における音声小段落毎の強調確率と平成確率の比のプロットを例として示す。更に、強調度を

$$\frac{\log P_{Semp} - \log P_{Snrm}}{L} \tag{6}$$

で定義する。音声小段落の強調度を降順に、音声小段落を一つでも含む音声段落を抽出し、抽出した音声段落の総延長時間 T_Gを計測する。音声の全長の 1/X 時間（要約率）、あるいは T_S時間に要約する場合、$T_G \fallingdotseq T_S$となるまで音声段落を抽出する。抽出した音声段落を、時系列順に再生し、任意の長さに要約する。

3.　マルチメディアコンテンツ要約システム

本稿の要約技術をシステム化した。図 3 にマルチメディアコンテンツシステムを示す。ユーザは、映像ファイルを選択し、任意の視聴時間を入力し、要約ボタンをクリックする。式(6)の強調度を降順にし、要約区間を決定し、再生区間を表示し、マルチメディアコンテンツが要約再生される。本システムを用いてマルチメディアコンテンツ要約実験を行った。

4.　マルチメディアコンテンツ要約実験と評価

マルチメディアコンテンツ要約実験と評価を行った。実験条件として、音声特徴量は前述した強調抽出実験と同様であり、音声小段落、音声段落抽出は前述した Ts=400ms、α=3、β=3.2 で抽出した。

ドキュメント、映画、スポーツのコンテンツ要約実験について述べる。表 2 に実験の結果を示す。映画の音声は英語であった。評価として、本技術による要約(要約率 1/X、T_S)と、再生時間が T_Sとなるように、10 秒間再生を一定間隔で繰り返したもの（一定間隔再生）を、被験者 10 名が比較評価した。実験では、平均 9.2 人の被験者が本技術を選択した。

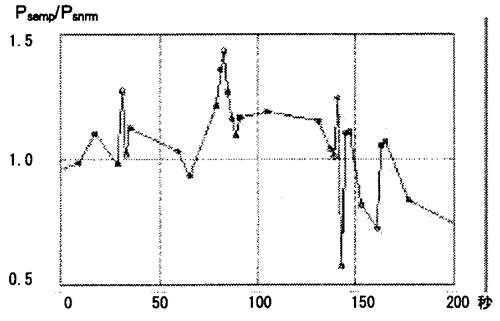

図2　音声小段落の強調確率と平静確率の比のプロット

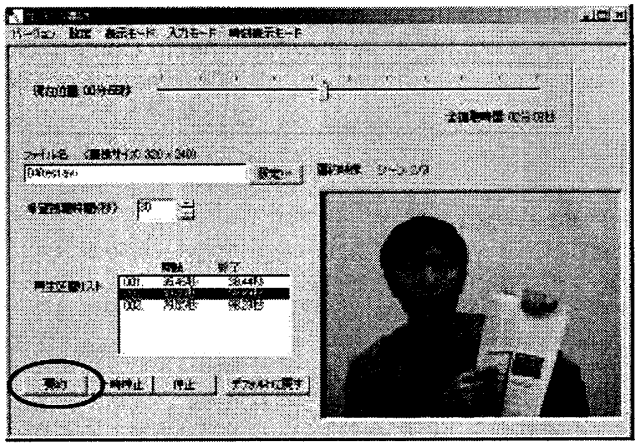

図3　マルチメディアコンテンツ要約システム

表2　コンテンツ要約実験結果

	コンテンツ	全長	要約時間	要約率	評価
1	ドキュメント	45分	90秒	1/30	9
2	映画（英語）	90分	90秒	1/60	10
3	映画（英語）	90分	90秒	1/60	7
4	映画（英語）	120分	120秒	1/60	10
5	スポーツ	100分	60秒	1/100	10

本技術を音楽コンテンツで実験した。歌謡曲 10 曲を対象に、1 曲 1 音声段落を抽出する。被験者 10 名は抽出した音声段落を“さび”と感じたかを評価した。平均 8.5 人/曲が“さび”であるとした。

5.　まとめ

音声の強調部分を基にマルチメディアコンテンツを要約する技術について報告した。音声の強調確率、平静確率を求めた。音声強調抽出の精度は、70%以上であった。多種多様なマルチメディアコンテンツにおいても、音声強調を抽出できるとの結果が得られた。強調度を定義し、多言語を含むマルチメディアコンテンツを任意の長さに要約し、一定間隔再生と比較評価した。80%以上の被験者が本技術を選択したことから、本提案方法が有用であるとの結果を得た。更に、音楽コンテンツの“さび”抽出を検討し、楽音コンテンツへの可能性を示した。

今後は、本技術による要約視聴した際の、マルチメディアコンテンツに対する“興味”を評価する予定である。

謝辞

日頃活発な議論をして頂いたサイバーソリューション研究所マルチメディア端末プロジェクト小川克彦プロジェクトマネージャ、町口恵美氏、竹内順二氏、武井香氏に深く感謝致します。

参考文献

[1] 堀, オーディオビジュアル複合情報処理シンポジウム 2001 論文集, pp. 3-10, 2001
[2] 堀, 古井, 信学論 VOL.J85-D-2, pp.200-209, 2002
[3] 川崎, 川俣, 山本, 板橋, 音講論, pp. 239-240, 2000-03
[4] 木山, 伊藤, 岡, 信学技法, SP95-35, 1995-06
[5] Chen, Withgott, Proc. ICASSP-92, pp. I-229-232(1992)
[6] 嵯峨山, 音講論, pp. 589-590, 1979-03