音声の感性情報に着目したマルチメディアコンテンツ要約技術

日高 浩太　町口 恵美　竹内 順二　水野 理　中嶌 信弥

NTT サイバーソリューション研究所

要旨

本稿では、マルチメディアコンテンツの音声の強調部分に着目し、強調音声を集めることでコンテンツを要約する。強調音声を声の“高さ”、“強さ”、“速さ”に対応する特徴量とその変化量を統計処理して抽出する。更に、強調音声の“強調度”を求めることで、ユーザの指定する任意の長さに要約する技術を提案する。一般のマルチメディアコンテンツの音声においても、強調音声を約 80%の高い精度で抽出した。音声の感性情報として強調音声を基にコンテンツを要約することで、オリジナルの雰囲気や感性情報を内包したプレビュの生成を実現した。本提案方法の要約と一定間隔再生によるリファレンスデータとの対比較評価実験の結果、80%以上の被験者が本技術を選択しており、その有効性を確認した。

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

Kota Hidaka　Megumi Machiguchi　Junji Takeuchi　Osamu Mizuno　Shinya Nakajima

NTT Cyber Solution Laboratories

Abstract

This paper describes a new multimedia summarization technique based on automatic extraction of emphasized speech. The proposed method estimates emphasized portions of speech at high accuracy, by using prosodic parameters such as pitch, power and speaking rate. As the method does not employ any speech recognition technique, it enables highly robust estimation under noisy environments. To extract emphasized portions of speech, the method introduces a metric, “degree of emphasis” that indicates how much emphasized each speech segment is. Given an article, the method computes the degree of emphasis for each speech segment in it. When a user requests the summary content of the article, the method collects the emphasized segments referring to the user specified ‘summarization rate.’ Preference experiments were performed in which subjects were conducted to select better summary content out of ones created by our method and ones made by a fixed interval approach. The preference rate of our method was 80%, and this result suggests that the proposed method can regenerate proper summary contents.

1. はじめに

多チャンネル時代の到来により、マルチメディアコンテンツが増加しているが、1 人 1 日当たりのコンテンツ視聴時間が比例して増加しているわけではない(1 人 1 日 4 時間)[1]。膨大なコンテンツの中から、ユーザが如何に嗜好のコンテンツに出会えるかが重要となってきている。ドラマ、映画等のコンテンツは、プレビュを頼りに選択することが多いが、ユーザは受動的に視聴するに過ぎず、インタラクティブ性はなかった。本稿では、有限のコンテンツ視聴時間を有効利用する為に、ユーザが希望する視聴時間を与え、システムが指定された時間のプレビュを生成する要約技術を提案す

る。プレビュ視聴後、オリジナルに対するユーザの興味を誘発する為、オリジナルの雰囲気や感性情報を内包した要約を行なう。

マルチメディアコンテンツの要約技術には、映像情報を利用したアプローチと音声情報を利用したアプローチの二つが考えられる。映像処理技術には、カット点検出、テロップ認識、カメラワーク検出等の映像インデクシング技術[2]がある。サムネール自動生成等のイベント一覧表示技術としての利点はあるが、任意の長さのプレビュ生成や、オリジナルの雰囲気、感性情報を内包した要約への適応は困難である。一方、音声情報を利用したアプローチとしては、音声認識結果をテキスト要約する手法[3]や、音韻的な類似度を計算してキーワードを検出し、キーワードを基に要約する手法[4][5]がある。音声認識ベースのアプローチは、タスク依存や、実環境における認識精度など実用上の問題点は少なくない。また、テキスト要約では発話時の強弱、ニュアンスといった感性情報が認識の過程で除外されてしまっており、感性情報を包括したプレビュ生成には適していない。キーワード検出による手法も同様に、実環境における検出精度の課題があり、またキーワード提示の意味合いが濃く、本稿の目的とする要約は行えない。

本稿では、音声の感性情報として音声の強調部分（強調音声）に着目し、強調音声を集めることでコンテンツを要約する。強調音声の強調度を定義し、強調度を用いてユーザが指定した任意の長さ(オリジナルの 1/10,1/100 等)に要約する。強調音声抽出の既往の技術に、音声の高さ、強さの特徴を用いて、イントネーション（ピッチパタン）の上昇、下降などのパタンに着目した手法[6]がある。主にインタビュー音声をターゲットとしており、BGM など頻繁に存在するドラマ、映画等の多種多様な音声の強調抽出は難しい。本稿では、より頑健性が高いコンテンツ要約方法を目的とする為、高さ、強さに加え発声速度をパラメータとし、ある特定のパタンに着目せずノンパラメトリックな統計的手法で強調音声を抽出する。音声に含まれる感情の研究[7][8][9][10][11]では感情認識の手法[11]があるが、実環境の音声での認識精度に課題があり、本稿の目的とする要約への適応は困難である。本提案方法によれば、実環境の音声においても精度よく強調音声を抽出し、強調音声を基にコンテンツを要約することで、オリジナルの雰囲気や感性情報を内包することが可能となる。

2. コンテンツ要約方法

2-1. 概要

マルチメディアコンテンツを対象に、強調音声を抽出し、強調音声を基にコンテンツを要約してプレビュを生成する。強調音声のみでは短すぎて意味が通じない為、文に相当する単位として音声段落を定義し、強調音声を含む音声段落をプレビュに用いる。図 1 にコンテンツ要約処理を示す。①強調音声抽出の為の音声特徴量として、音声の基本周波数（高さ）、パワ（強さ）、スペクトル変化量（速さ）を抽出し、これらを統計処理して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。②プレビュに用いる区間として、音声段落を抽出する。③強調音声の強調度を定義し、強調度を用いて任意の長さにコンテンツを要約してプレビュを生成する。

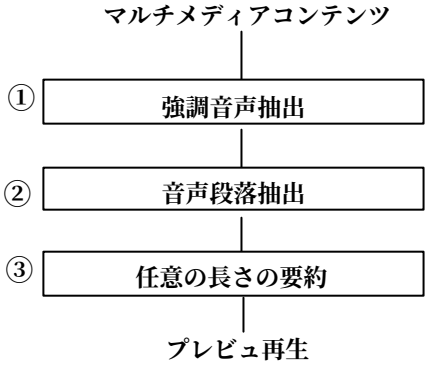


図1.コンテンツ要約処理

2-2. 強調音声の分類

強調音声を抽出する為に、強調音声のバリエーションを調べた。作業者 2 名の合意の下、5 名程度の会議音声約 10 時間を聴取して、発話における強調を 8 つに分類した。表 1 に各強調の特長を示し、括弧内には強調が出現する状況の例を示す。強調される言葉には感嘆、嫌悪、苛立ち、怒り等の感情が内包されていることがある[12]。郡は東京語におけるこの種の強調を、“発音による強調法”、“強意を示す副詞や接辞の使用”、“修辞表現”、“表現意図の強調”の 4 つに分類した[12]。更に“発音による強調法”を、普段の言い方よりも力んだ発音、母音の延伸と促音・撥音の挿入、テンポの遅延と語末拍上昇、拍ごとのポーズ挿入に小分類しているが、本稿で分類した表 1 の 1、2、3、5、6、7 の強調は前述の“発音による強調法”の類型に当て

はまることがわかる。一方、4 の強調は言い直しや、言い淀み[13]に対応している。8 の強調については、直前までの声の強さ、高さとの差が強調と関係がある事を示している。

表1. 発話における強調の分類

1	強く、伸ばす （重要な事を述べる時）
2	強く、高くなる （重要な事を述べる時）
3	話し始めを伸ばす （話題変更を主張し、意見に注目させる時）
4	高低差、強弱差が激しくなる （苦笑い、焦り、ごまかす時）
5	急激に高くなる （周囲に同意を求める、問いかける時）
6	徐々に強くなる （念を押す時）
7	速く、強く、高くなる （発話権を主張、保持する時）
8	直前までに比べ、急激に弱く、低くなる （本音や秘密を述べる時）

2-3. 強調音声の持続時間

強調音声の持続時間を調べた。会議音声約 10 時間を聴取し、作業者 5 名の合意の下、表 1 の 8 つの基準に該当する強調音声となる区間を 2228 個設定した。本稿では、この区間を強調ラベルとする。図 2 は強調音声の持続時間のヒストグラムである。強調ラベルの持続時間の平均は 0.699s で、標準偏差が 0.414 であり、最大は 3.546s、最小は 0.075s であった。強調ラベルの持続時間が 1s 以下であるのは 81%の割合であり、1.5s 以下であるのは 96%の割合であった。これにより、本稿では、強調音声抽出の為の分析単位を 1s、分析単位のシフト幅を 0.5s とした。

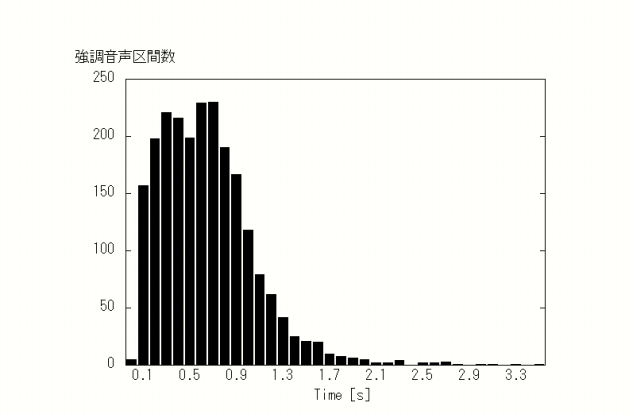
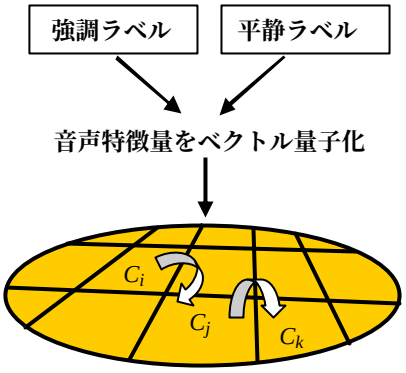


図2.強調音声の持続時間のヒストグラム

2-4. 強調音声抽出方法

強調音声を抽出する為に、音声特徴量から発話の強調確率と平静確率を求める。

予め、学習データを作成する。図 3 に学習データ作成の手順を示す。



コードの時間遷移 $C_i \rightarrow C_j \rightarrow C_k$

コード C_k	*unigram*	*bigram*	*trigram*
強調確率	$Pemp(C_k)$	$Pemp(C_k\|C_j)$	$Pemp(C_k\|C_iC_j)$
平静確率	$Pnrm(C_k)$	$Prnm(C_k\|C_j)$	$Pnrm(C_k\|C_iC_j)$

図3.学習データ作成手順

前述の強調ラベルに加え、作業者 5 名の合意の下、平静ラベルは表 1 の基準に該当せず、発話が平静であるものとした。

強調ラベル区間、平静ラベル区間の音声から音声特徴量を抽出する。表 2 に音声特徴量を示す。“高さ”、“強さ”に対応する基本周波数、パワ、これらの差成分を抽出し、“速さ”に対応する特徴として、動的尺度[14]の単位時間当たりのピーク本数を計測し、差成分を求める。

表2. 音声特徴量

基本周波数のフレーム平均	$\overline{f0}$
前後差成分	$\pm\Delta\overline{f0}$
パワのフレーム平均	$\overline{P}$
前後差成分	$\pm\Delta\overline{P}$
動的尺度のピーク本数(単位時間)	dp
差成分	$\pm\Delta dp$

LBG 法[15]でベクトル量子化し、コードブックを作成する。あるコード C_k について、強調の出現確率 $p_{emp}(C_k)$、平静の出現確率 $p_{nrm}(C_k)$を求める。コードの時間遷移の順が C_i、C_j、C_k の場合の、コード C_k が出現する条件付確率として、強調について $p_{emp}(C_k|C_j)$、$p_{emp}(C_k|C_iC_j)$を、平静について $p_{nrm}(C_k|C_j)$、$p_{nrm}(C_k|C_iC_j)$を求める。強調確率、平静確率の算出の為、予め任意のコードについて、unigram、bigram、trigram に相当する前述の確率を求めておく。

強調音声抽出の分析区間の強調確率、平静確率を算出する。分析区間 X がフレーム数 L から構成され、量子化された音声特徴量のコードが $C_l(l=1,2,..L)$であるとする。あるフレーム f の強調確率 $P_E(f)$、平静確率 $P_N(f)$を

$$P_E(f) = l_{e1}P_{emp}(C_f \mid C_{f-1}C_{f-2}) + l_{e2}P_{emp}(C_f \mid C_{f-1}) + l_{e3}P_{emp}(C_f) \quad (1)$$

$$P_N(f) = l_{n1}P_{nrm}(C_f \mid C_{f-1}C_{f-2}) + l_{n2}P_{nrm}(C_f \mid C_{f-1}) + l_{n3}P_{nrm}(C_f) \quad (2)$$

の形で線形補完して求める。重み係数 l_{ei}、$l_{ni}(i=1,2,3)$は学習データから削除補間法で推定する。全フレームで式(1)、(2)を求め、強調確率 P_{Xemp}、平静確率 P_{Xnrm} は

$$P_{Xemp} = \prod_l P_E(l) \quad (3)$$

$$P_{Xnrm} = \prod_l P_N(l) \quad (4)$$

となる。式(3)、(4)より強調確率が平静確率より大であれば分析区間 X は強調音声とした。

2-5. 音声段落抽出方法

プレビュに用いる音声段落を抽出する。音声段落は音声の基本周波数とパワの時間変化、ポーズ情報の全てに関連するが、本稿では複数話者が自然に会話する際、発話の“間”が特に重要であると考えた。“間”に相当するものとして、音声が無声となる時間情報を用いる。図 4 は会話音声における音声段落を示している。Ⅰ)自己相関係数、Ⅱ)パワ、Ⅲ)有声無声判定、Ⅳ)基本周波数であり、矩形波で示される有声無声判定（有声:1、無声:0）から d 以上の無声区間で囲まれる区間を音声段落とした。矢印が示す範囲は d=1s とした場合の音声段落である。パワと無声区間から音声段落を抽出した既往の研究[16]と同程度の性能を確認した。

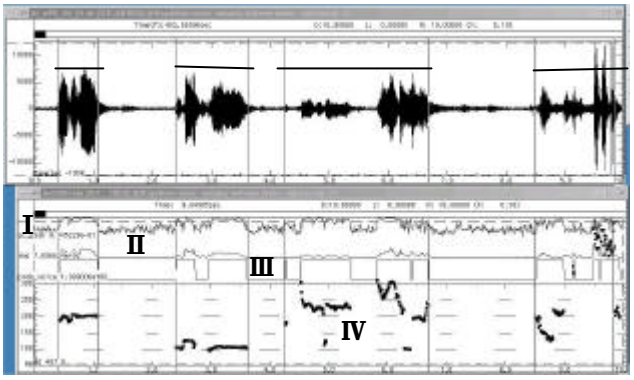


図4.無声区間から抽出した音声段落

2-6. 任意の長さの要約方法

前述の強調音声抽出では、分析区間 X が強調音声か否かを判定したが、ユーザが指定した任意の長さに要約するには、“どの程度の強調であるか”という強調度が必要となる。本稿では、式(3)、(4)の強調確率、平静確率から強調音声の分析区間の強調度 K_X を、

$$K_X = \frac{\log P_{Xemp} - \log P_{Xnrm}}{L} \quad (5)$$

で定義する。図 5 に、分析区間毎の強調度 K_X の例を示す。図 6 に任意の長さの要約方法の概略を示す。強調度 K_X が一定値以上の分析区間を一つでも含む音声段落を要約に使用する。要約に使用する音声段落の総和 T_G（秒）が要約時間となる。長さ T_S（秒）に要約する場合、T_G　T_S となる強調度の閾値 K_T を求める。K_X>K_T となる分析区間を含む音声段落を、時系列順に再生し、任意の長さに要約する。

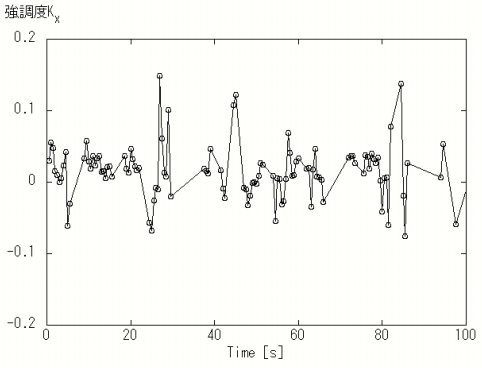


図5.　分析区間毎の強調度 K_X の例

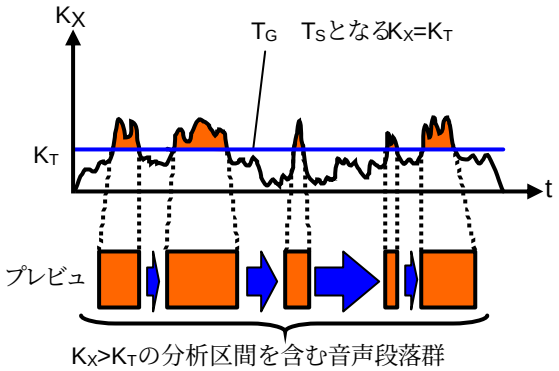


図6． 任意の長さの要約方法

2-7． コンテンツ要約システム

ユーザが指定した任意の長さでマルチメディアコンテンツのプレビュを生成できる要約システムを開発した。図 7 にシステムのインターフェースを示す。インタラクションとしてユーザが、①映像ファイルを選択し、②任意の視聴時間を入力し、③要約ボタンをクリックすると、システムは指定の長さにプレビュを生成する。

図7． コンテンツ要約システム

3． 実験と評価

3-1． 強調音声抽出実験と評価

強調音声抽出実験を行った。音声資料として、発話者 5 名程度の会議音声約 10 時間を用いて学習データと評価データを作成した。学習データは強調ラベル 707 区間、平静ラベル 807 区間であり、評価データは強調ラベル 173 区間、平静ラベル 193 区間である。学習データの音声特徴量を抽出し、コードブックサイズ 256 のコードブックを作成した。

学習データの各ラベルについて式(3)、(4)の強調確率と平静確率を比較し、作業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った(open 実験)。結果を表 3 に示す。両実験の再現率、適合率で約 85%を得た。

一般のマルチメディアコンテンツの音声においても強調音声を抽出できるか実験した。音声資料として、前述の会議音声に加え、談話、講演、ニュース、ドラマ、映画の音声（合計約 2780 分）を用いて学習データと評価データを作成した。学習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。学習データは計約 58 分であった。コードブックサイズ 64 とし、前述と同様の実験を行った。表 4 に結果を示す。open 実験、close 実験共に再現率、適合率は約 80%であった。作業者が強調ラベル、平静ラベルを設定する際の揺らぎが 1 割程度存在することを考慮すれば、再現率、適合率 80%の値は強調音声抽出の目標値となると考える。本方法はマルチメディアコンテンツの音声においても精度よく強調音声を抽出できると言える。

表3. 会議音声における強調音声抽出実験結果

	強調ラベル		平静ラベル	
	再現率	適合率	再現率	適合率
Close	89%	87%	88%	91%
Open	84%	91%	92%	87%

表4. マルチメディアコンテンツの音声における強調音声抽出実験結果

	強調ラベル		平静ラベル	
	再現率	適合率	再現率	適合率
Close	82%	80%	80%	82%
Open	78%	80%	79%	79%

強調音声抽出精度の学習データ依存性を調べた。図 8 は、open 実験における強調ラベル、平静ラベルの再現率と学習データ数の関係を示す。コードブックサイズは 64 とした。横軸の学習データ数は強調ラベル数と平静ラベル数の和である。学習デ

ータ数が 1500 近傍を境に、強調ラベル、平静ラベルの再現率の差が極小となり、学習データ作成は各ラベル数 750 程度が良いことがわかる。

本方法は強調音声を“高さ”、“強さ”に加え、“速さ”に対応する特徴量を用いて抽出することを特徴としている。強調音声の“速さ”の影響を評価する為、強調ラベルについて、表 2 に示す音声特徴量の *dp*、±Δ*dp* を使用した場合と、使用しない場合の強調音声抽出の再現率を調べた。図 9 に open 実験における再現率と、コードブックサイズの関係を示す。何れのコードブックサイズにおいても“速さ”の特徴量を用いた方が、再現率が良いことから、強調音声抽出には“高さ”、“強さ”、“速さ”の全ての特徴を用いることが良いことがわかる。

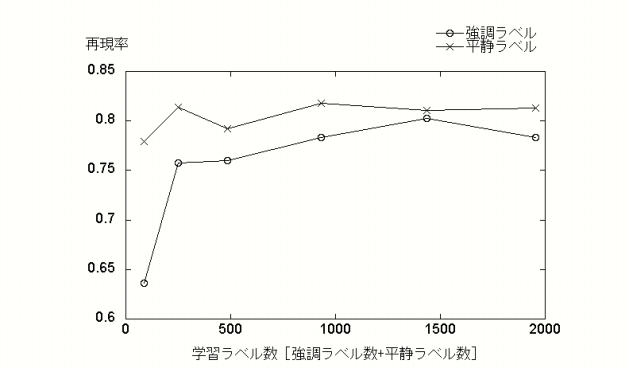


図8． 強調音声抽出精度の学習データ依存性

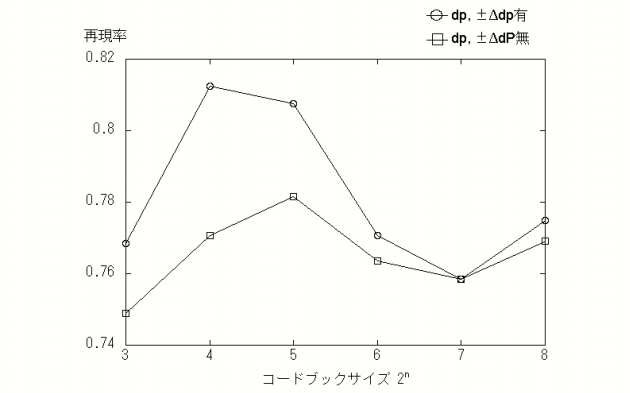


図9． 強調音声抽出精度の“速さ”の特徴の影響

3-2． 任意の長さの要約実験と評価

任意の長さに要約する実験を行った。コンテンツは 7 つで、各コンテンツは 3 名の話者が、あるテーマについて自由会話している映像である。任意の長さの要約の評価として、本方法のプレビュとリファレンスデータを対比較した。リファレンスデータは 10 秒間再生を一定間隔で繰り返した、無作為な抽出によるプレビュである(一定間隔再生)。プレビュの長さをオリジナルの 1/10、1/15、1/30 に変化させて対比較した。被験者 10 名が「直感的によいと感じた」方を選択した。表 5 に評価結果として、本方法を選択した被験者の割合を示す。各テーマの括弧内はオリジナルの長さである。概ね 80%の被験者が本方法を選択したことから、無作為なプレビュよりも、強調音声を基に生成したプレビュが良いとの結果が得られた。また、プレビュの長さが短くなるにつれて、本方法の評価が高い。プレビュが短ければ、本方法のプレビュと一定間隔再生との重なりの部分が少なくなることから、より本方法の有用性が評価されたと考える。

表5. 本方法のプレビュと一定間隔再生との対比較評価結果

会話のテーマ	1/30	1/15	1/10	平均
IT 社会(600s)	100%	80%	70%	83%
環境問題(600s)	80%	100%	80%	86%
日本の景気(900s)	80%	80%	70%	76%
自分の故郷(900s)	90%	90%	70%	83%
BSE 問題(500s)	60%	90%	90%	80%
今年のプロ野球(1200s)	100%	90%	70%	86%
最近の 10 代(1800s)	90%	50%	70%	70%
平均	85%	82%	74%	81%

3-3． 本要約技術の絶対評価と検討

本方法の要約技術の絶対評価を行った。評価用映像として表 6 に示す 6 つのコンテンツを用いた。各コンテンツのオリジナルは約 10 分である。プレビュは 1 分の長さとした。1 分としたのは、ユーザの負担にならないプレビュの長さであると考えたからである。被験者はプレビュを視聴した後、オリジナルを全て視聴する。プレビュは本方法のものと、前述と同様の一定間隔再生の二つを用意した。被験者 11 名を 2 つのグループ（α班 6 名、β班 5 名）に分けた。α班は奇数番号のコンテンツについて本方法のプレビュを視聴し、偶数番号について一定間隔再生を視聴する。β班はα班の逆のパタンのプレビュを視聴する。表 7 に評価項目を示す。プレビュを視聴後に表 7 の A～C の項目を評価する。オリジナルを視聴後に、D～F の項目を評価する。回答は「はい」、「どちらかというとはい」、「どちらかというといいえ」、「いいえ」の何

れか一つを強制選択した。

表6. 要約技術の絶対評価用映像コンテンツ

		コンテンツ名
1		4 人の卒業生の同窓会
2		2 組の夫婦の親交
3		人形劇「新しい村長は？」
4		人形劇「4 匹のお見合い」
5		談話「旅行先を決めよう」
6		談話「私の中の英雄」

表7. 評価項目

A	感性情報（盛上がり）のシーンが多数あったと思うか
B	全体のコンテンツを見てみたいと思うか
C	コンテンツの特徴や雰囲気が伝わるか
D	感性情報（盛上がり）の視点で、両者に差はなかったか
E	プレビュは、オリジナルの雰囲気を内包していたか
F	プレビュは、オリジナルの要約といえるか

図 10 は本方法と一定間隔再生の絶対評価の比較である。6 つのコンテンツの平均を各項目について棒グラフで示した。全ての評価項目で本方式の評価が高く、特に評価項目 A、D、E が顕著であり、感性情報を包括するには無作為な生成よりも強調音声を基にプレビュを生成する本方法が良いとの結果を得た。

本方法のみの評価をより詳細に見る。図 11 に評価項目に対する 11 人の回答の平均を棒グラフで、コンテンツ毎に示した。評価結果から、概ね良いとされている。特に評価項目 B、C、E から、本方法の要約技術で生成したプレビュがオリジナルの雰囲気や感性情報を内包し、全体のコンテンツの視聴を促すことがわかる。評価項目 D、F では、コンテンツによっては否定的な意見もある。本稿では音声の盛り上がった場面の集まりが要約としたが、それ以外に考慮しなければならない点、例えば映像情報の利用などが示唆された。

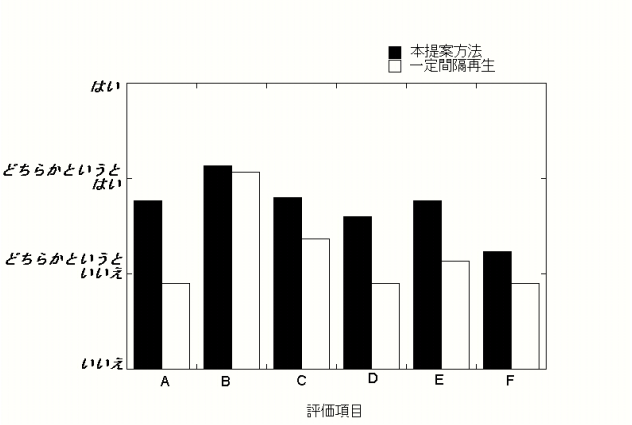


図10． 本方法と一定間隔再生の絶対評価の比較

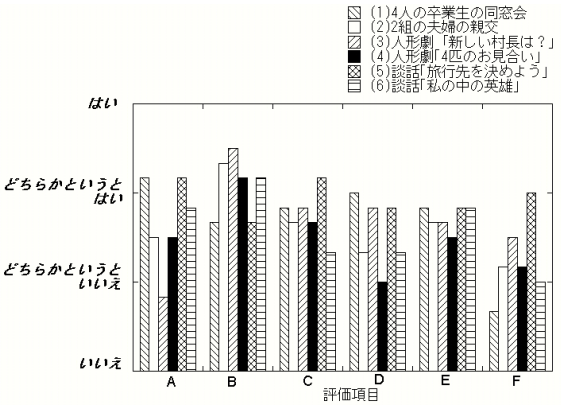


図11． コンテンツ要約技術の絶対評価

4． まとめ

マルチメディアコンテンツの音声の強調部分を基に要約する技術について報告した。音声の“高さ”、“強さ”、“速さ”をパラメータとし、特定のパタンに着目せずノンパラメトリックな統計的手法で強調音声を抽出した。強調音声抽出実験を行い、分析区間の強調確率と平静確率を算出し、会議音声において再現率、適合率は約 85%の高い精度で強調音声を抽出できるとの結果が得られた。一般のマルチメディアコンテンツの音声においても、強調音声を約 80%の高い精度で抽出した。分析区間の強調度を求め、音声段落を抽出し、強調度が一定値以上の分析区間を一つでも含む音声段落をプレビュに使用した。強調度を用いることで、ユーザが指定する任意の長さに要約することが可能になった。本方法のプレビュと一定間隔再生を対比較し、80%以上の被験者が本技術を選択したこ

とから、強調音声を基に要約することの有用性を確認した。更に、本方法を絶対評価し、本方法がオリジナルの雰囲気や感性情報を内包し、オリジナルに対するユーザの興味を誘発する効果があるとの結果が得られた。

謝辞

日頃活発な議論をして頂いたNTT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、武井香氏、松浦晴美氏、NTTサイバースペース研究所メディア処理プロジェクト阿部匡伸グループリーダに深く感謝致します。

文献

[1] 情報メディア白書2002年版, 電通, pp.135, 2002.

[2] Tonomura Y., Akutsu A., Taniguchi Y., Suzuki G.: Structured Video Computing, IEEE Multimedia, Vol. 1, No. 3, pp.34-43, Fall 1994.

[3] 堀智織, 古井貞熙: 単語抽出による音声要約文生成法とその評価, 電子情報通信学会論文誌 D-Ⅱ, Vol. J85-D-Ⅱ, No.2, pp.200-209, 2002.

[4] 木山次郎, 伊藤慶明, 岡隆一: Incremental Reference Interval-free連続DPを用いた任意話題音声の要約, 電子情報通信学会技術研究報告, Vol. 95, No. 123(SP95 25-36), pp.81-88, 1995.

[5] 川崎剛史, 川俣眞人, 山本幹雄: 韻律情報を考慮した音声要約の一方法, 日本音響学会講演論文集, pp.239-240, 2000-03.

[6] Chen F R., Withgott M.: The use of emphasis to automatically summarize a spoken discourse, Proc IEEE Int. Conf. Acoust. Speech Signal Process., Vol 1, pp.I.229-232, 1992.

[7] 櫻庭京子, 今泉敏, 筧一彦: 幼児・児童の感情表現における音響的分析 –「ぴかちゅう」にこめられた感性情報-, 電子情報通信学会技術研究報告, SP99-86, pp.1-7, 1999.

[8] 光本浩士, 柳田益造, 大多和寛, 田村進一: 皮肉音声の音響的特徴に基づいた判別, 電子情報通信学会技術研究報告, SP99-39, pp.17-24, 1996.

[9] 上床弘幸, 小林豊, 新美康永: 音声の感情表現の分析とモデル化, 電子情報通信学会技術研究報告, SP92-131, pp65-72, 1993.

[10] 武田昌一, 大山玄, 朽谷綾香, 西澤良博: 日本語における「怒り」を表現する韻律的特徴の解析, 日本音響学会誌, Vol. 58, pp.561-568, 2002.

[11] Tosa N., Nakatsu R.: Life-Like Communication Agent –Emotion Sensing Character "MIC" and Feeling Session Character "MUSE", IEEE International Conference on Multimedia Computing and Systems, pp.12-19, 1996.

[12] 郡史郎: 講座日本語と日本語教育2・日本語の音声・音韻（上）, 杉藤美代子編, pp.316-342., 明治書院

[13] 田窪行則: 音声言語の言語学的モデルをめざして –音声対話管理標識を中心に-, 情報処理学会誌, Vol. 36, No. 11, pp1020-1026, 1995.

[14] 嵯峨山茂樹, 板倉文忠: 音声の動的尺度に含まれる個人性情報, 日本音響学会講演論文集, pp.589-590, 1979-03.

[15] Linde Y., Buzo A., Gray R. M.: An algorithm for vector quantizer design, IEEE Trans. Commun., Vol. Com-28, pp.84-95, 1980.

[16] 日高浩太, 水野理, 中嶌信弥: 発話の強調自動抽出による音声要約技術, 日本音響学会講演論文集, pp.99-100, 2002-9.