音声情報を用いた会議における雰囲気把握支援方法の検討

加藤 雄一[1]　大江 展弘[2]　陳 金姫[2]　平野 靖[3]　梶田 将司[3]　間瀬 健二[3]

名古屋大学工学部電気電子・情報工学科[1]

名古屋大学大学院情報科学研究科[2]　名古屋大学情報連携基盤センター[3]

1　はじめに

近年，記録媒体の小型化，大容量化が進み，会議やミーティングなどを録音・録画し，議事録として用いる研究が盛んに行われている[1-3]．これらの研究には，ノンバーバルな情報を利用し，要約や検索の手がかりを与えるもの[1,2]，カメラとマイクロホンによる会議風景の収録と半自動的な議論の構造化の仕組みを実現したものがある[3]．しかし，会議全体を録画・録音したものすべてを閲覧するには時間がかかりすぎるため実用的ではない．一方，従来の議事録では発言内容が要約され，文章で記述されるので，比較的短時間で閲覧できるが，会議の雰囲気に関する情報は欠落してしまう．そのため，議事録を閲覧した上で，どのような雰囲気で会議が行われたかを知りたいという要求がある．この要求に対し，雰囲気の把握を補助する方法が必要となる．

本稿では，上記の要求に対し，音声の強調部分と発話区間に着目し，会議における雰囲気のひとつである「盛り上がり」を検出する方法を検討する．

2　盛り上がりの抽出方法

ここでは，本研究の雰囲気把握支援方法に必要な「盛り上がり」の抽出方法について述べる．

2.1　盛り上がりの定量化

会議における「盛り上がり」とは何かを考える．会議特有の「盛り上がり」のひとつとして議論が白熱している状態が考えられる．このようなときは，人間の声は自然と大きくなる．さらに，それにより発話者同士の音の重なり（クロストーキング）が多くなる．つまり，どちらの場合も音声のパワー（強さ）が大きくなることがわかる．

以上のことから，本稿では，通常の発話音声のパワーと比べて大きくなっている部分を会議における「盛り上がり」のひとつであるとし，「盛り上がり」の定量化を行う．

2.2　発話区間の検出

発話区間は一瞬の時刻ではなく，ある程度の時間幅を持つものである．本稿では，雰囲気を掴むために必要な発話区間を検出し，瞬間的なノイズに対応するために，それぞれのピンマイクから得た音声のパワーs_i（i=1, 2, . . . , s_{max}）を計算し(s_{max} は音声データ数)，0.1 秒単位で，ヒューリスティックに定めた閾値 t_1（無発話区間の 5 倍程度の値，本稿では 2000 とした）以上の値をとる回数 r_j を計算する．その回数と 0.1 秒前の回数との差分 Δr_j を式(1)により算出し，Δr_j が 0 より大きければ発話しているとみなす．(t_{max} は会議音声の秒数の 10 倍)

$$\Delta r_j = r_j - r_{j-1} (j = 1,2,...,t_{\max}) \qquad (1)$$

2.3　盛り上がり度の計算

議論が白熱すると，声は大きくなり，クロストーキングを多く発生させる．よって，会議における「盛り上がり」は，瞬間ではなく時間の幅を持つと考えられる．

そこで，発話区間中において 0.1 秒単位の平均パワーの値 m_j が閾値 t_2（t_1 の 1.5 倍，本稿では 3000 とした）以上のとき，1 秒後までの m_j が t_2 以上の値をとる回数を計算する．その値を「盛り上がり度(h_j)」とする．こうすることで，議論が白熱している度合いが表され，「盛り上がり」を定量化できる．直感的には，m_j が t_2 以上になったときに，その時点付近の m_j が t_2 以上になっている個数を求めたものが「盛り上がり度(h_j)」である．

式(2)において $u(x)$は 1-0 型ヘビサイド関数である．1-0 型ヘビサイド関数は x が 0 以上であれば 1 を返し，0 より小さいならば 0 を返す関数である．

Study of atmosphere grasp supporting method that uses voice information at conference

[1] Yuichi Kato
Dept. of Information Engineering, School of Engineering, Nagoya University
[2] Nobuhiro Ohe, Chen Jinji
Graduate School of Information Science, Nagoya University
[3] Yasushi Hirano, Shoji Kajita, Kenji Mase
Information Technology Center, Nagoya University

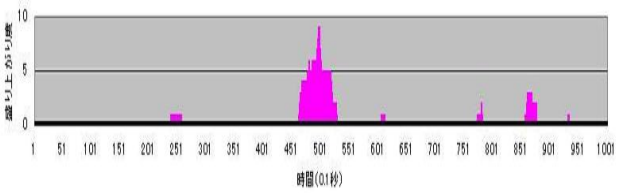

図１　盛り上がり度の推移

$$m_j = \frac{\sum_{i=(j-1)*4410}^{j*4410} |s_i|}{4410} \quad (j = 1,2,...,t_{\max})$$

$$t_2 = 1.5 * t_1 \qquad (2)$$

$$h_j = \sum_{k=(j-1)*10}^{(j-1)*10+10} u(t_2 - m_k) \quad (j = 1,2,...,t_{\max})$$

3　評価実験

前章で求めた盛り上がり度の計算の有効性を検証するために，基礎実験を行った．ここでは，実験方法と結果について述べる．

3.1　準備

会議における音声の取得は，会議参加者それぞれに，胸の辺りにピンマイクをつけ，同時に声を取得する．ピンマイクは SONY の ECM-T145 を 4 つ使用し，AD/DA 変換ボードは interface 社の PCI-3120 を使用した．サンプリング周波数は 44.1kHz，分解能は 12bit で録音する．

被験者 4 人にピンマイクをつけ，100 秒間の擬似会議を行った．議題はあらかじめ決めており，その議題について自由に討論をし，音声の取得を行った．

後日，会議参加者・不参加者合わせて 4 人にその会議の音声を聞いてもらい，個人個人の主観に基づいて，この会議の中で一番盛り上がっていると思われる部分の開始時刻と終了時刻を示してもらった．

3.2　結果

2.3 の手順に従って計算した，盛り上がり度の推移を図１に示す．図１は，縦軸が盛り上がり度，横軸が時間を表している．

図１から，最も盛り上がり度が高い区間は，47～52 秒であった．「この会議で一番盛り上がっていると思われる時間はいつか」という質問に対する調査結果は，「47～49 秒」「40～51 秒」「48～50 秒」「46～49 秒」であった．このうち 3 人の意見は，検出された盛り上がりの区間に含まれる．また，「40～51 秒」と答えた中にも盛り上がりとして検出された区間が含まれる．個人個人によって，盛り上がりの開始時刻と終了時刻は異なるが，本稿で提案した方法によって，会議における「盛り上がり」を求めることができたと考えられる．

4　おわりに

本稿では，会議における雰囲気のひとつである「盛り上がり」について検討した．会議において人間が主観的に感じる盛り上がり区間と，本稿で定義した「盛り上がり」の区間とが，ほぼ一致したことによって，本稿で論じた方法によって「盛り上がり」を求めることができることがわかった．しかし，「盛り上がり」には少なくとも 2 種類があり，本稿で論じた，発話者の音声のパワーが大きいものと，クロストーキングによるものとがある．

今後の課題として，多くの会議を収録して，「盛り上がり」の種類，雰囲気の種類を見つけ出し，分離する必要が挙げられる．雰囲気の種類が増えることによって，会議の雰囲気をより一層把握しやすくなると考えられる．また，検出された雰囲気を検索や要約の手がかりとすること[4]で，雰囲気の有用性を示していく必要がある．

謝辞

本研究は文部科学省cc-Societyプロジェクトの支援により行われた．

参考文献

[1] 大江 展弘，平野 靖，梶田 将司，間瀬 健二，“全方位画像を用いた会議記録・閲覧システム”，第 3 回情報科学技術フォーラム(FIT2004) 一般講演論文集，pp. 557-558 (2004)．

[2] 角康之，間瀬健二，”インタラクション・コーパス構築の試みとしてのミーティング・キャプチャ”，ヒューマンインタフェースシンポジウム 2002，pp. 241-244(2002)．

[3] K.Nagao, T.Shimizu, and K.Kaji，“Discussion Mining: Knowledge Discovery from Semantically Annotated Discussion Records”, In Proceedings of the International Workshop, “From Semantic Web to Semantic World”(2003)

[4] 日高 浩太，町口 恵美，竹内 順二，水野 理，中島 信弥，” 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”，インタラクション 2003 論文集，pp．17-24(2003)．