I-029

メタデータ生成のための歓声と応援の分類手法

A Classification Method of Crowd Noise for metadata generation

佐野 雅規† 住吉 英樹† 八木 伸行†
Masanori Sano Hideki Sumiyoshi Nobuyuki Yagi

1. まえがき

将来のサーバー型放送では，番組に付与されたメタデータを手がかりに，ユーザーは興味ある部分だけを視聴することができるようになる．このサービスの有力な候補としてスポーツ番組を挙げることができ，試合のダイジェストなど，その需要は高いと言える．このサービスを実現するためには，放送局でメタデータを付与する必要がある．このメタデータをできるだけ効率的に制作するため，我々は番組制作の過程で得られる様々な情報を利用した制作手法を研究してきた．

そのひとつとして，サッカー番組を対象にハイライト部分の抽出に，会場音（中継の制作業務で必ず設置されるスタジアム全体の音を収音するマイク音）を利用し，イベント内容の抽出にアナウンサーの実況音声の音声認識結果を使用したメタデータ自動生成を試みてきた[1]．会場音を用いた理由は，観客の沸いた部分がハイライトシーンの最有力候補であるからである．放送音声を用いた歓声部分の抽出例として[2]があるが，我々は放送局の立場から会場音を用いて実験した．ところが，多くの試合に適応してみたところ，[1]で用いた歓声抽出手法では，サポーター（応援団）による結束した応援（以後，組織的応援と記述）も誤って抽出することがわかった．そこで，これを除くために，スペクトル包絡線を使用する手法を開発した．

2. 提案手法

従来の手法[1]では，会場音の短時間パワーを計算し，その時間的変化のみを解析した．図１はこの短時間パワーの時間的変化を示したもので，後半に存在する山の部分が，試合中に観客が沸いた典型的な変化に相当する．我々はこのような部分を，時刻と共に多様な変化をする会場音に合わせて，リアルタイムに抽出する動的閾値処理を用いている．

今回，この動的閾値処理の結果から，前述した組織的応援シーンを除くために，新たにスペクトル包絡線の解析を加えた．提案する手法を図２に示す．スペクトル包絡線の解析は，動的閾値により抽出された区間の短時間パワーを入力とする．はじめに入力された区間の先頭の一部分を選択し，その区間の FFT 計算を行う．次に計算結果を基にスペクトル包絡線を求め，正規化を行う．正規化された包絡線に対し，歓声が沸いた場合と，組織的応援の場合の特徴の有無を判定し，歓声が沸いた区間の情報だけを出力する．

スペクトル包絡線を求める際に，区間の先頭部分だけを用いる理由は，続いておこる様々な音源の影響を除去するためである．例えばスポーツ観戦においては，注目する何かが発生すると，歓声が沸き，続いて拍手や指笛による雑音の発生，また会場のアナウンスが入る場合も多い．このため動的閾値で抽出された区間全体を用いると，解析対象に後者の音が含まれてしまい，区別できなくなる可能性が高くなるからである．

また，正規化する理由はスペクトル包絡線の大小ではなく，形状を判定するためである．本手法では歓声が沸いた場合と組織的応援の場合をスペクトル包絡線の形状で分類する．従って求めた包絡線の値の大小には関係なく，その相対的な形（スペクトルパワーの構成）が重要となる．このため正規化を行うことでデータを一律に扱うことが可能となる．この正規化された包絡線を用いて，歓声と組織的応援を分類する．サッカー２試合分について，動的閾値により得られる全区間を対象にスペクトル包絡線を分析したところ，図３に示すような事象毎に共通するパターンが得られた．(a)は本手法において抽出目的となる歓声が沸いた部分の包絡線パターンであり，(b)(c)(d)は全て組織的応援のパターンである．特に(c)は会場が静まっていて，サポーターの歌声のみが大きく響いている区間に相当し，(d)はサポーターがそろって「ウォーィ」と単発的に叫んでいる区間（一種のかけ声のようなもの）に相当する．

これらのパターンを用いて，歓声が沸いたシーン(a)とそれ以外の３つのシーン(b)(c)(d)を分類する．ここで図３の中に明るいグレーで示した２つの範囲を，上注目部，下注目部と呼ぶことにする．まず，歌声が含まれる場合の包絡線(b)(c)は，(a)(d)に比べると，複数の細いピークが存在することが特徴となる．従って上注目部での細いピークの数を数えることで，歌声を含んでいるかどうかを判別する．

次に，歌声を含まない組織的応援である単発的な叫び声

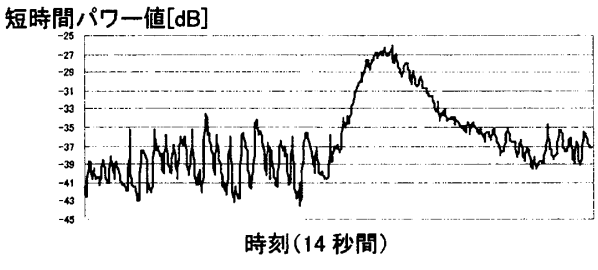


図1 歓声が沸いた際の短時間パワー

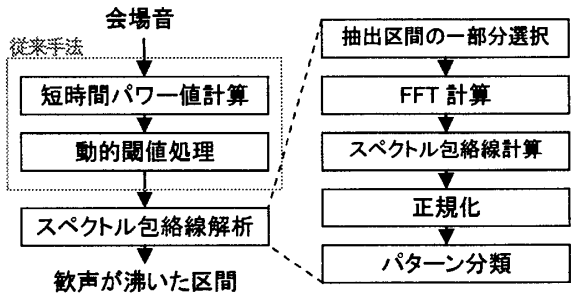
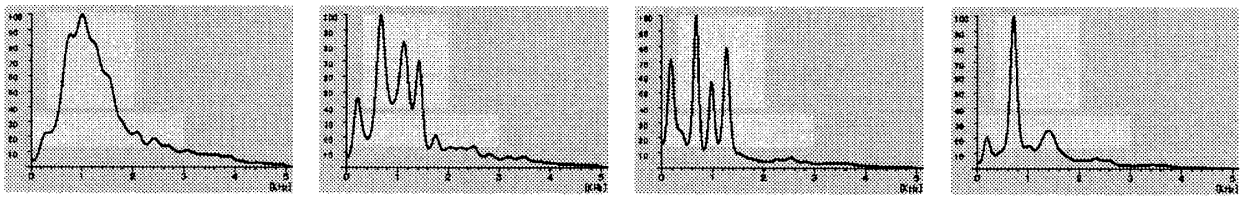


図2 提案する手法

† NHK 放送技術研究所（知能情報処理）



(a) 歓声が沸いた　(b) 組織的応援(歌声あり)　(c) 組織的応援（ほぼ歌声のみ）　(d) 組織的応援(単発的叫び)

図３　正規化したスペクトル包絡線のパターン（縦軸：正規化スペクトルパワー，横軸は周波数 0～5

の場合(d)は，下注目部での周波数成分に注目すると歓声が沸いたシーン(a)と区別できる．(a)の下注目部においては連続して広い周波数の成分を含んでいるが（ホワイトノイズ的），(d)ではある部分の周波数しか含まれていない．

従って，上述した特徴の有無を調べることにより，表１に示すようなシーンの分類が可能となる．

表１　包絡線パターンの特徴とシーン分類

	広周波数成分	狭周波数成分
複数ピークなし	歓声沸いた	単発的応援
複数ピークあり	歌声あり	歌声のみ

３．実験と考察

実験は平成 15 年のＪリーグ６試合を対象に行った．実験に先立ち，各パラメータを次のように決定した．従来手法の部分は文献[1]の値をそのまま用い，今回追加した包絡線の解析部分については次のように決めた．包絡線を求める際の低ケフレンシー成分は，経験則から 40 ポイントとした．また，包絡線パターンの特徴分析については，２試合のデータ(136 カ所)について調査し，次のように定めた．上注目部は，パワー値 40 以上，周波数帯域は 366～2153[Hz]で定義し，この中の細いピーク（高さ 30 以上，幅 430[Hz]以内）を数えた．ピークの高さは，ピークの頂点と谷（左右の高い方）までの差と定義し，幅はピークの頂点から 30 低い値での周波数幅とした．また，下注目部での周波数成分の特徴は，パワー値 15,25,35 全てにおいて，それぞれ 1507, 968, 645[Hz]以上連続した周波数成分を含む場合に，表１の広い周波数成分を含むとした．

表２に，パラメータを導き出す際に用いた２試合を除いた，４試合分の実験結果を示す．表中の( )内の数字は，動的閾値により抽出されたシーン数を示す．また，適合率の欄にある[ ]は，従来手法による適合率を示す．

実験結果から，どの試合についても高い再現率(95.5%)と適合率(97.8%)で，歓声の沸いたシーンを検出できていることがわかる．再現率と適合率を比べると，適合率の方が若干高くなっている．メタデータ作成の観点からすると，この検出結果に多少のゴミデータ（組織的応援など）が入ることは問題なく，むしろ検出もれをなくす方が重要と言える．検出にもれたシーンを調べると，コーナーキックやゴール前でのフリーキックを得た瞬間，相手チームのミスへの弱い歓声やブーイング，歓声は沸いているが強い指笛の音が入っているシーンであった．これらのスペクトル包絡線はみな似ており，図３の(a)と(d)の中間的なものであった．従ってこれらは，パラメータを適切に変化させることで検出できるようになると考える．

適合率を用いて従来手法との比較をすると，全体で平均22%改善されている．メタデータの制作作業では抽出された全てのシーンについて，番組を見て確認する必要がある．少しでも不適切な部分を減らすことは，作業効率の向上につながり，本手法による改善は大きいと考える．

また，表２から試合によって動的閾値処理で検出される区間の数に 23～100 とバラツキがあり，１つ１つの試合においても歓声と組織的応援の割合に大きな差があることがわかる．これは会場そのものの違いや観客の数，そして試合を行っているチームのサポーターによる振る舞いが大きく違うことに起因する．しかしながら，このように会場音を形成する要因が大きく異なる場合であっても，本手法は歓声が沸いた区間を精度良く抽出できると言える．

表２　歓声が沸いたシーンの検出結果

試合 (検査区間数)	検出数	再現率	適合率 [従来]
試合１　(100)	72	94	95 [69]
試合２　(48)	33	91	96 [66]
試合３　(66)	45	97	100 [68]
試合４　(23)	23	100	100 [100]
４試合平均 (237)	173	95.5	97.8 [75.8]

４．まとめ

本稿ではスポーツ番組にメタデータを付与する際に，スタジアムの会場音を用いて，精度良く歓声の沸いた区間を検出する手法について報告した．従来手法では取り除くことができなかった組織的な応援部分を，周波数解析を加えることにより自動的に除外することができた．具体的にはスペクトル包絡線を求め，そのパターンの特徴から歓声と組織的応援を分類した．実験結果は再現率 95.5%，適合率 97.8%で歓声が沸いた部分を検出できており，本手法の有効性が確認できた．メタデータ制作の観点からは，できるだけ再現率を高める必要があり，今後は音声以外の情報と組み合わせることで精度をあげる手法について研究していく．

参考文献

[1] M.Sano, H.Sumiyoshi, M.Shibata, and N.Yagi,"Generating Metadata from Acoustic and Speech Data in Live Broadcasting," Proc. 30th ICASSP, MSP-P2.4, Vol.II, pp.1145-1148, Philadelphia, U.S.A., Mar. 2005

[2] 渡部隆志，二反田直己，長谷山美紀，北島秀夫，“ファジィクラスタリングを用いたサッカー映像におけるオーディオインデキシングに関する考察，”信学技報，ITS2004-57, E2004-191, Feb. 2005