Vol.2010-AVM-69 No.10
2010/7/16

音声の有無による違いを考慮した Home video 簡易編集

高木幸一[†,††] 川田亮一[†] 篠崎隆宏[††] 古井貞 熙[††]

本稿では，音声の有無による違いを考慮した home video 簡易編集方式について報告する．音声情報を含む home video を編集する際に，映像・音声の両方を参照しながら編集する場合と，映像のみの情報を参照して編集する場合で結果が異なる．この違いを主観評価実験から明らかにする．さらに，その結果を利用して，音のみからしか得られない情報，すなわち，音としての本質的な情報のみを視覚化し，一般ユーザが視覚だけを使用して音声を含めたものと同じレベルの編集を行うことができる方式を提案する．主観評価の結果，映像・音声の両方を参照して編集する場合と大差ない結果が得られることを示す．

Home Video Trimming Method based on a Difference Depending on Presence or Absence of Audio Signals

Koichi Takagi[†,††] Ryoichi Kawada[†] Takahiro Shinozaki[††] and Sadaoki Furui[††]

This paper proposes a method for supporting trimming from home video on a mobile terminal without listening to the sound. It has two main contributions. First, we have analyzed the difference of trimming results between with and without listening to the sound, and derived what the necessary audio information not to be obtained from video is. Second, in consideration of the results, only the essential audio data is visualized on a small display of mobile terminal. An experimental result shows that the case of using the above-mentioned visualization data is comparable to that of providing both audio and video.

1. はじめに

近年，携帯端末の撮像デバイスの高性能化が進んでおり，10M 画素単位のものも珍しくなくなりつつある．そのような背景のもと，携帯端末で写真だけでなく Home Video を撮影する機会が増えてきている．特に，携帯端末はいつでも所持しているケースが多いことから，思い立った時に撮影を開始，終了を行うことができる利点がある．また，携帯端末内の蓄積スペースが大きくなってきているため，多少長時間の Home Video を撮影しても問題がない．ところが，それを他の人と共有しようとすると話は違ってくる．例えば，長時間 Home Video の場合，ネットワーク負荷の観点から，それを他の人に安易に送ることが困難になる，また，あまり長時間コンテンツだと，それを閲覧する意思がない限り見てもらえないなどの問題が生じる．そこで，簡単に重要な部分だけを抜き出したい（短時間コンテンツを作りたい．以下「短尺化」という．）という要求が生じる．さらに，この短尺化を，時間場所にとらわれることなく，携帯端末上で簡単にできるようになれば便利である．

ところで, Home Video には通常, 映像情報(Video, 以下”V”と書く)と音声情報(Audio, 以下”A”と書く)が共存するケースがほとんどである．Home Video を編集する際には，両情報を使って行うのが一般的である．ところが，特に屋外で撮影されたビデオには雑音が重畳されている場合が多く, ある程度大きな音にしないと聴くことができない．さらに，外出先（屋外）で編集を行おうとしたときに，周囲の雑音の影響により，自端末の音を大きくしないと聴くことができない（図 1）．結果，自端末のボリュームを大きくせざるを得なくなるため，周囲に迷惑をかける可能性が生ずる（これはイヤホンを付けている場合も同様である）. また, イヤホン等を付ける手間さえも省略したい．そこで，画面上の情報だけで，すなわち，目（視覚）だけで編集ができたらより簡単に編集ができるようになるはずである（なお，筆者らは，目（視覚）と片手を使うだけで簡単に編集をできるようにすることを目標としている.）

以上から，筆者らは，”A”情報を画面上に視覚化することができれば，それは動画像編集支援につながると考えた．最も簡単な視覚化方法は，図 1 のようにタイムライン上に波形を示すことである．ところが，雑音が多く含まれていると，同図のようにすべての箇所がほぼ同じに見えてしまい，特に携帯端末のような小さい画面では意味をなさない．そこで，各時刻でどのような情報が含まれているのかを”A”の情報から分析する必要がある．

[†] KDDI 研究所
KDDI R&D Labs. Inc.
[††] 東京工業大学情報理工学研究科
Graduate School of Information Science and Engineering, Tokyo Institute of Tech.

Vol.2010-AVM-69 No.10
2010/7/16

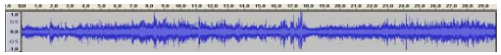

図 1　Can we use this wave data as it is for video authoring?

ところで，”A”の情報を視覚化するにあたって，タイムライン上の各時刻の音がそれぞれ何を意味しているのか等の既存研究（indexing, annotation, classification）は数多く存在する．それらはすべて，

１．　”V”とは独立に”A”の情報を評価し，Video に付与したもの

２．　”A”と”V”の両方が存在したものを対象とし，両者とも同等に扱われたもののどちらかに相当する．それに対し，筆者らが本稿で対象とするのは，”V”はそのまま提示する一方，”A”のみ何らかの形に加工して提示するケースである．すなわち，”V”がそのまま閲覧できるときに，それに対しどのような”A”を付加することが，ユーザにとって本質的なのかである．

以上から，筆者らは，この問題に対する回答を模索するために，予備実験を行う．具体的には，被験者に，１つのコンテンツに対し”V”のみの場合と”V+A”の場合の両方を提示し，それらをどのように編集するかについて評価実験を行う．その差異から，”A”の情報として本質的に必要な情報は何であるかを分析する．その後で，その“本質的な情報”を”A”の情報から分析した上で携帯端末上に動画像と一緒に表示する（以下“visA”と書く）ための方法について提案し，本提案手法により構成される視覚情報のみの画面で編集をする場合と，”V+A”の情報をそのまま使って編集する場合について，その差異をコンテンツごとに評価する．

図 2　Impossible to listen to audio information due to an ambient noise.

2.　関連研究

短尺化には２つの方法が考えられる．１つは全自動で行う方法，もう１つは何らかの形で indexing を行ったうえで，それを頼りにユーザが手動で行う方法である．

従来，前者，すなわち，全自動短尺化に挑戦した研究は多く存在する 0．ところが，実際のところ，これらはユーザにより異なることが想定される．つまり，全員に共通したハイライトを抽出するのは極めて困難である．特にホームビデオの場合はなおさらである．

一方，後者，すなわち，短尺化（編集，閲覧も含む）の補助のために，何らかの形で indexing を行う方法も多数検討されている．例えば，文献 0 では，”A”の情報を使って indexing を行い，”V＋A”の編集を行うことを目的としている．放送向けの映像中に流れる音声と楽曲の区間を検出し，映像に索引付けしている．同研究では，放送用編集で使用されるルールを用いている点もポイントとなっている．また，”A”の情報と”V”の情報を一緒に使ってこれを行うことを目的とした既存研究も多数存在する．短尺化のみの目的ではないが，”V”と”A”を同時に使用する方式をサーベイした結果は，文献 1)に掲載されている．ただし，これら既存研究では基本的に”A”のみの情報に特化して，もしくは”A/V”を同一レベルで扱っている．例えば segmentation の際に，”A”のみで，もしくは”A/V”を混合して切れ目となるであろう個所を探索し，そのまま提示している．

それに対し，本稿では，”A”を聴かずに（聴覚を使用せずに），視覚情報だけを使用して短尺化を行うケースを対象とする．よって，”A”により”V”を補完する場合のみを抽出できればよいはずである．また，被験者ごとに得られる“V＋A”の結果に“V＋visA”の結果をどれだけ近づけることができるかを対象とする．この場合，短尺化の結果がユーザにより異なるという前提で，ユーザが動画像を短尺化する際の支援になればよいはずである．

3.　“A”が”V”を補完するケースとは？

前節で述べた通り，本稿で必要となるのは，”V”だけからは得られない情報で，かつ”A”として本質的な情報は何であるかを見極めることである．例えば”V”だけからは得られない情報として，ビデオに映し出されていない物体が発する音の情報はそれに相当するが，それでも，すべての音の情報が必要であるとは限らない．そこで，本節では”V のみ”，”V+A”のそれぞれについて実際に短尺化を行い，その差異を観測する予備実験を行う

3.1　予備実験

本予備実験において，被験者は複数のコンテンツに対する短尺化を行う．実験条件

は表 1 の通りである．具体的に被験者は，
「自分が重要だと考える箇所を含め，その duration がオリジナルシーケンスの 20～30%の長さになるように切り出してください（短尺化してください）」
という依頼に対し，N(=5)種類のシーケンスを評価する．さらに，それぞれのシーケンスに対し，“V のみ”，および，“V+A”の情報提示に基づく短尺化処理を行う．すなわち，同一シーケンスに対して２度の短尺化処理を行うことになる．なお，あるシーケンスに対して，“V のみ”の場合を評価する前に“V+A”の場合を評価してしまうと，被験者は“V のみ”を評価する際に，記憶された”A”の情報をもとに判断してしまう，すなわち，“V+A”との差異が認められなくなる恐れがある．そこで，あえて両者の差異を明確にするために，“V のみ”の情報を提示した後に“V+A”の情報を提示することとする．また，連続して同一シーケンスを評価することによる飽きを起こさせないようにするため,同一シーケンスは連続して提示しないようにする.以上を踏まえて，シーケンス提示順は図 3 の通りとする．

なお，各被験者に提示するシーケンスはそれぞれ部分的な重なりはあるものの，異なるものとする．

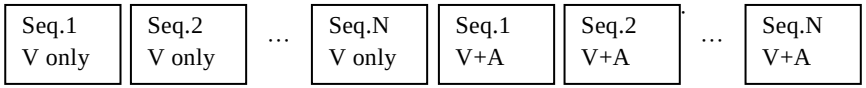

図 3　Sequences presentation order (Preliminary experiments)

表 1　Experimental setup

	予備実験（３節）	本実験（５節）
提示シーケンス数／被験者数	５種類×２通り (“V only”, “V+A”)	５種類×３通り (“V only”, ”V+visA”, ”V+A”)
被験者数	16（20～40 代・男女）	16（20～40 代，男女）
シーケンス仕様	携帯端末で撮影できるものを対象とする duration：30 秒～2 分 file format：MP4 フォーマット（3g2/3gp） Video：H.264, 15fps, QCIF～QVGA，64～256kbps Audio：AMR-NB (8kHz, mono, 12.2kbps)	
シーケンス内容	一般の方により撮影されたムービー(Home Video) 例：公園での散歩，食事の風景，ペットとの戯れ，ドライブ，ショッピング，ランドマークへの旅行など	

3.2 結果および考察

実験の結果，“V のみ”の場合と“V+A”の場合で，当初の予想通り，差異が生じることが確認された．

以上の結果を細かく分析するために，提示したシーケンスに対し，あらかじめ Audio indexing を手動で行っておき，本実験により得られた結果と照合した(図 4)．結果を表 2 に示す．本節で Audio index とは，各セグメント（１秒ごと）に対し表 2 の index に相当する情報が存在するか否かを重複を許して示したものである．“V のみ”の結果に対し”V+A”の結果で挿入された箇所(Ins.)，削除された箇所(Del.)，およびどちらにおいても選択された箇所(Co-corr.)を求め，各 Audio index に対し，そこに存在しているラベル数をセグメント（1 秒）ごとにカウントした．

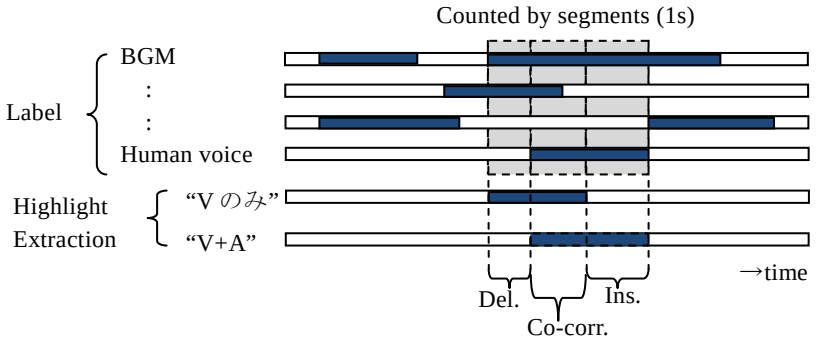

図 4　Verification method of highlight extraction results between “only V” and “V+A.”

同表より，“V のみ”の場合と比較して，“V+A”の場合に抽出されたセグメントに含まれている情報として,「人の声」が挙げられることがわかる. 特に, 人の声では”Ins”が非常に多くなっており，音声を聞くと「人の声」の部分を含めたくなることがわかる. 一方, 両者の差異が大きいものとして「突発的な音」が挙げられる. これは Ins., Del. ともに多くなっている．突発的な音は Highlight として意味があるケースもあるが，逆に突発な音が入ることを嫌うケースも考えられる．それゆえ，このような結果になったと想像できる．なお，逆に，オーディオを用いた短尺化の際に，一般的に最も注目される歓声の部分であるが，ここでは抽出の対象として，その割合があまり高くなっていないことがわかる．この理由として，通常の放送用の映像に対し，ホームビデオでは歓声を意図してとらえないこと(逆に他の音声を意図してとらえようとすること)，Co-corr.の数を見てもわかる通り，歓声があがっている箇所は”V”だけからもわかるためであることなどが考えられる．いずれにせよ，今回の目的に置いては不要であることがわかる．

以上，「人の声」「突発的な音」の２つの情報は他の情報とは突出して変化があった項目である．よって，これらの情報を効率的に検出し，”V”と一緒に提示することが

できれば，当初の目的を達成することができると期待できる．

表 2　Verification results of highlight extraction between “only V” and “V+A.”

index	Description	# segments	Ins.	Del.	Co-corr.
Human voice	人の声	1316	234	62	259
Artificial noise(Bell,etc.)	機械的に作られた音（ベルなど）	248	22	32	51
Vehicle (w/o bell)	乗物から発せられる音	315	10	12	52
Animal	動物から発せられる音	143	8	2	13
Cheer	歓声	621	21	23	101
Applause	拍手	321	18	12	71
BGM	背景に流れる音楽（除楽器）	1437	65	105	72
Musical instrument (w/o BGM)	楽器の音全般	312	12	2	23
Water, wave	水に関連する音	102	2	0	9
Wind	風の音	82	0	3	8
Life sound	人のざわつき，足音	1050	13	12	35
Impulse sound	モノがぶつかる音等	612	39	41	42

4. 本質的に必要となる”A”情報の抽出および提示

本節では，前節の予備実験の結果を受け，本質的に必要な”A”の情報を効率的に求める方法，およびその提示方法について論ずる．

4.1 必要となる”A”情報の検出

前節で述べた通り，本質的に必要となるは「人の声」「突発音」である．基本的に，これらの検出は，既存の方式で実現可能であることが想定されるが，具体的にどの方式を組み合わせるのかがポイントとなる．

4.1.1 人の声

人の声の検出に関する既存研究は多く存在し，一般的に VAD(Voice Activity Detection)と呼ばれている．ところが，人の声にはいろいろと種類があり，それらをすべて検出することが望ましいとは限らない．特に，今回の目的，すなわち，短尺化のために意味を持つ「人の声」とは一体どのようなものであるかを考慮すべきである．そこで，前節の予備実験の結果を観測したところ，ある程度大きな声で明瞭に発声されている部分が多く抽出されていることが確認できた．逆に，声があまり大きくない場合，もしくは他の音にかき消されてしまっている場合はその対象とならないことが多い．

以上から，本節では，「人の声」の抽出方法として，文献 4)にならい，以下の３つの方法を用いることとする a.

- 振幅レベル
  雑音に対する耐性は低いが，VADを実現するためには最も簡単な方法である．一定長のフレームごとに信号にハミング窓を適用した値$\{x_n : n=1, \ldots, N\}$を用いて
  $$(1/N)\,\Sigma \log(x_n). \quad (1)$$
  で表現する．
- ZCR
  ZCR(ゼロクロス比)も VAD ではよく用いられる方法である．一定長のフレームごとに，レベル 0 をクロスする比率を求める．
- GMM 尤度に基づく分類
  GMM(Gaussian Mixture Model)も音声検出にはよく用いられる値である．ここでは入力フレームに対する speech と noise の GMM 尤度比を以下を用いて計算する．
  $$\log(p(v_t|\Theta_s)) - \log(p(v_t|\Theta_n)) \quad (2)$$
  ただし，v_tはGMMのための音響ベクトル，Θ_s，Θ_nはそれぞれ音声とノイズのモデルパラメータセットである．音響ベクトルはMFCC12 次元，およびそれらの 1 次微分，2 次微分の計 36 次元に Δpower, Δ Δpowerを加えた計 38 次元のベクトルを求め使用した．ノイズのGMMはノイズ特性の多様化をカバーしないといけない．そこで， 表 1 内の３節で使用した音源の中でhuman voiceとラベルされなかった箇所を雑音の学習用に用いて雑音モデルを生成する．

ただし，上述の方法は音声の明瞭性を考慮に入れていない．そこで，上記に加え，明瞭に発声されている箇所が重要であることから，音声明瞭度指標(AI)を用いてこれを判別することとする．

- AI
  ここでは簡単のため the-count-the-dot 方式 5)を適用する．すなわち周波数ごとに重みづけられた dot の数に基づき，対象となる音がそれらをどのくらい含ん

a 文献 4)ではスペクトル尺度，すなわち，各サブバンドごとの音声と雑音の S/N による評価に関する言及もあるが，N（音声でない）区間を自動的に事前に抽出することは難しいため，ここでは対象外とする．

でいるかにより表現される指標である．

以上，振幅レベル，ZCR，および AI についてはフレーム長を 100msec，GMM についてはフレーム長を 25msec とし，それぞれ 50%オーバラップさせながらシフトするものとする．また，それぞれ，各セグメントに存在するフレームに対する値の中で最大となる値を適用する．さらに，以上の各特徴量（x とする）に対するダイナミックレンジはそれぞれ異なるため，文献 4)同様，シグモイド関数

$$f(x) = \left[1 + \exp\left\{-\frac{4}{\sqrt{2\pi\sigma}}(x-\mu)\right\}\right]^{-1} \quad (3)$$

を用いて値域を 0 ～ 1 に正規化する．なお，その際の平均 μ ・分散 σ^2 は表 1 の予備実験で用いたムービーから求める．

また，本稿では簡単のため，以上 4 つの値を並列に提示することとする。

4.1.2 突発音

突発音の検出についても既存研究がいくつも存在する．例えば，文献 6)では，野球において打球音が発せられている箇所をパワーの盛り上がりから盛り下がりまでの時間が特定の間隔内にあるとし，さらに人の声として取られることを回避する為，周期成分の除去を行っている．また，文献 7)では銃声と他の音との分類をベイジアンネットワークを使って行っている．

以上，特定の音を扱おうとすると，上述のように具体的な方策の適用が必要と考えられるが，ここでは突発音を一括して扱えばよいため，以下の方法を取ることとする．突発音では，そのパワーは時間的に急激に大きくなる．そこで，サブバンドb（全部で 5 サブバンドに分割）ごとの短時間スペクトル $s_b(i)$（ただし，フレーム長は 16msec とし，50%オーバラップシフトである）の時間方向変化 $\Delta s_b(i)$ を求め，各セグメント（1 秒）内で最大となる値 $\max \Delta s_b(i)$ を求める．ただし単なる変動のみの場合，声などの周期性のある音も同時に検出されてしまうため，それを避けるために上記同様にサブバンドごとの周期性 bp(b) を

$$bp(b,k) = \frac{\sum_{i=0}^{M-1} s_b(i) s_b(i-k)}{\sqrt{\sum_{i=0}^{M-1} s_b{}^2(i)} \sqrt{\sum_{i=0}^{M-1} s_b{}^2(i-k)}} \quad (4)$$

$$bp(b) = \max_k |bp(b,k)| \quad (5)$$

で求める．なお，周期性が高い箇所を低く見積もるために 1－bp(b)を求め，この値を前述のスペクトルの変動にかける．そして，4.1.1 同様に(3)式のシグモイド関数を適用し，正規化して提示することとする．

4.2 Audio情報提示方法

“A”の情報の提示は以下の通り行うこととする．タイムライン上に，4.1 節の方法で「人の話し声」および「突発音」として検出されたセグメント（＝1 秒単位）について，それぞれ図 5 のようにマーカを提示する．このマーカは 4.1 で求められた値をセグメントごとに 256 階調に量子化し表現したものである．0 のとき最低階調，1 のとき最高階調となる．被験者は”V”の情報に加え，このように表示された情報を参考にしながら編集（短尺化）を行う．

図 5 GUI for evaluation (with audio information visualization)

5. 実験および考察

前節の検討の有効性を評価する為に以下の実験を行う．

“V のみ”のデータに 4.2 節で述べた方法で 4.1 節で求めた音声情報を付加したもの(以下“V+VisA”と書く)を作成し，被験者に提示する．そして，3 節同様，”V+A”のデータを提示した場合と比較を行う．提示の順番は，3 節同様”V+A”のデータに対する結果を ground truth とするため，”V のみ”，”V+visA”のデータの後に”V+A”のデータを提示することとする．

以上の結果を表 3 に示す．同表における Ins, Del., Co-corr は 3 節で述べたものと同じである．すなわち”V+A”の場合と比較し，”V のみ”および”V+visA”で得られた結果がどの程度接近しているかを示すものである．

表 3　Verification results of highlight extraction segments

index	# segments for “V+A”	Ins.	Del.	Co-corr.
V のみ	889	321	305	568(=63.9%)
V+visA		76	85	813(=91.5%)

同表より，同一の被験者に対し，”V のみ”の場合と比較し”V+visA”の方が”V+A”の結果に大きく近づけることができることがわかる．特に”V+visA”の場合，9 割以上のセグメントで”V+A”と同一の検出結果が得られており，本方式の有効性が確認できる．

また，取得されたデータを細かく分析した結果，大きな雑音が重畳していない音については，人の声の検出にパワーの情報が最も有用であるが，雑音が大きく重畳した音については，図１にも掲げた通り，時間方向にパワーは変化しない．そのような場合にも人の声であると判別できた理由として，ZCR や雑音 GMM，明瞭度の利用が挙げられる．実際に少なくとも ZCR，雑音 GMM の尤度，明瞭度の内の１つは，その値が高くなっていることを確認した．

また，”V のみ”の場合，突発音のみが提示されている箇所で，３節における予備実験同様に Ins.，Del となるケースが散見された．それに対し，”V+visA”では”V+A”とほぼ同等の結果が得られることを確認した．

以上から，”人の声”，および，”突発音”の箇所を視覚的に提示することにより，短尺化の処理支援に大きく貢献できることが確認できた．

また，ユーザに全体の感想を求めた結果，音声であることを示すと，誰の声なのかわかるようにしてほしい（音声識別），かつ，可能であれば何としゃべっているのかもわかるようになる（音声認識）とうれしいという意見が多く聞かれた．本研究ではできるだけ簡単に，かつ携帯端末上に表示できるレベルで行うことを主眼に置いているため，ここまでは対象外であったが，発話内容を間違いなく，かつわかりやすく表示することができれば，さらに使い勝手の良いものになると考える．これらは今後の課題である．

6. おわりに

本稿では，携帯端末上で視覚情報のみで home video の編集（短尺化）を行うことを目的とし，同 video に含まれる音声情報を視覚化する方法について検討を行った．

まず，home video の映像を無音にして見た場合と，音声を聴きながら見た場合の両者において，短尺化を行った結果の差異について検証した．その結果，音声を聴きながら見た場合には「人の声」の部分および「突発音」の部分がそれぞれ多く含まれていることを確認した．

そこで，「人の声」および「突発音」に相当するセグメントに対し，その情報を視覚化してビデオ情報と一緒に提示することにより，視覚情報のみで効率的に短尺化を行うことが可能であると考え，これらの音が存在する箇所のみを自動的に検出し，タイムライン上に提示した．

以上を用いて短尺化を行った結果，ビデオのみ（音声なし）で短尺化を行った場合と比較して，ビデオ・音声の両方を参照して短尺化を行った場合に大きく近づくことを確認した．

なお，本研究における，それぞれの項目の寄与度を測定する為には，（１）音の情報を視覚化した場合とそうでない場合の比較，（２）各音の情報がどのくらい寄与しているのか？を測定する必要があったが，これらの主観評価実験を複数のシーケンスに対しすべて行おうとすると，１人の被験者ごとに数時間を要する必要があるため，本稿では（１）のみとした．そこで，今後は，短尺化における人の声の情報および突発音の寄与度のさらなる分析を行う．また，短尺化のための音声（人の声）情報の自動検出部の性能改善を行う予定である．

参考文献

1) 例えば，Cees G.M. Snoek, Marcel Worring, "Multimodal Video Indexing: A Review of the State-of-the-art," Multimedia Tools and Applications, Volume 25, Number 1, p.p. 5-35, 2005.
2) Truong, B. T. and Venkatesh, S., “Video abstraction: A systematic review and classification,” ACM Trans. Multimedia Comput.Commun. Appl. 3, 1, Article 3 (Feb. 2007).
3) K. Minami, A. Akutsu, H. Hamada, Y. Tonomura, “Video handling with music and speech detection,” J. IEEE Multimedia, Vol. 5, Issue 3, p.p. 17-25, 1998.
4) Yusuke Kida, Tatsuya Kawahara, “Evaluation of voice activity detection by combining multiple features with weight adaptation,” Proc. INTERSPEECH, pp.1966--1969, 2006.
5) H. Gustav Mueller and Mead C. Killion, “Aneasy Method For Calculating the Articulation Index,” Hearing Journal, Vol 43. No. 9, Sept. 1990.
6) 三上　弾，紺谷　精一，森本　正志, “突発音検出と教師なし動きクラスタリングを用いた野球映像からの投球イベント検出,” 信学論 D, Vol.J90-D , No.2, pp.526-534, 2007
7) Aggelos Pikrakis, Sergios Theodoridis, “GUNSHOT DETECTION IN AUDIO STREAMS FROM MOVIES BY MEANS OF DYNAMIC PROGRAMMING AND BAYESIAN NETWORKS,” Proc. IEEE ICASSP2008, pp.21-24, Mar. 2008.