Vol.2012-HCI-149 No.13
2012/7/20

音声区間自動検出技術を用いた変速再生方式による映像の高速鑑賞システムの検討

栗原一貴† 佐々木洋子† 緒方淳† 後藤真孝†

本論文では，多くの映画のように会話やナレーションが主体として構成されている映像に広く適用可能な高速鑑賞システムを提案する．具体的には「音声箇所は聴取理解可能な速度で再生し，非音声箇所はさらに高速な速度で再生する」という変速再生方式を採用し，その前処理として必要な，対象映像中の音声区間と非音声区間の分離を自動化する認識器を構築する．市販の字幕付き映画 DVD のデータセットを用いて，字幕表示区間は音声区間とみなし，MFCC を特徴量とした Gaussian Mixture Model による機械学習を行うことにより，実用的な性能を実現した．さらに映像の高速鑑賞を PC やスマートフォンなどの多様なデバイスから行えるよう，Pogoplug を用いたクラウドベースのシステムとして実装した．

Discussion on a System for Watching Videos at Very High Speed using Two-level Fast-forwarding based on Automatic Speech Detection

KAZUTAKA KURIHARA† YOKO SASAKI†
JUN OGATA† MASATAKA GOTO†

In video content such as feature films, the main themes and messages are often sufficiently conveyed through dialogue and narration. Here we propose a system for watching such videos at very high speed while ensuring that speech is still comprehensible. Specifically, we employ a purpose-built automatic speech detector to realize two-level fast-forwarding for a wide variety of video content: very fast during segments without speech, and understandably fast during segments with speech. In our experiments, practical performance was achieved by frame-by-frame audio classification using Gaussian mixture models trained on subtitle information from 120 commercial DVD movies. In addition, we used the Pogoplug service to implement a cloud-based prototype that incorporates our speech detector, which allows end users to watch videos at very high speed on a wide variety of devices such as PCs and smart phones.

1. はじめに

現代において我々個人が扱う可能性のある情報は飛躍的に増加している．映像メディアについてもそれは例外ではない．Youtube には 1 分間で 600 本もの動画が投稿されており，その総時間は 25 時間に及ぶと言われている[1].

増大する情報に対応し個人の情報消費を支援するため，映像を要約する技術はこれまでに多く研究されている[11][12]．それらはニュースや監視カメラ映像，スポーツ映像などの要約に対象を絞ったものが多い．これらの映像は決まった繰り返し構造を持っていたり，背景映像の種類が限定されていたりしており，ハイライトとそうでない箇所の識別が比較的容易である．研究が盛んに進められた背景には，このような理由があったことが挙げられるだろう．

本研究では，そのような特徴のない，映画のような映像を対象にした情報の高速消費の支援を行う．娯楽のために映画を見るのに，その高速消費にニーズがあるのかという疑問が沸き起こるかもしれない．確かに与えられた余暇の時間をリラックスするためだけに映画を見るのであれば，高速消費は必要ないかもしれない．しかし，我々に与えられた時間は有限で細切れであるのに対し，興味をそそられる映像作品，さらにはそれほど興味は無いが話題を呼んでいる流行の映像作品なども増加の一途をたどっているため，それらの作品を効率的に閲覧する技術も必要になってくる．たとえば友人との共通の話題の確保のために，さして興味のない流行のドラマや古い名作映画を見る場合を想像するとよいだろう．

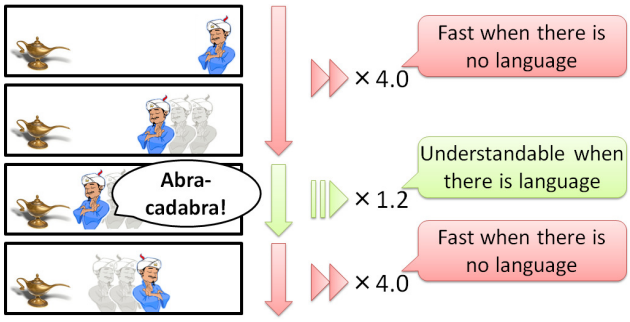

図 1 映像の高速鑑賞システム「CinemaGazer」における変速再生の概要[2]

映像の高速消費支援方法には要約と高速再生の 2 つの手段があるが，我々は高速再生のアプローチをとる．[2][8][13]などで指摘されているように，要約よりも「ひと通り見た」という満足感の得られる高速再生の方がユーザ

† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

は好ましいと考えている．特に映画のような対象では，ハイライト抽出に基づく要約は一般化が難しく，適さないと考える．

我々は広い対象の映画の高速再生を実現するため，言語情報に注目する．小説を原作とした映画が多いことからも言えるように，多くの映画は会話やナレーションなどの言語情報を主体として物語が構成されており，その理解がコンテンツ鑑賞において重要な位置を占めている．言語情報は映像コンテンツ中に字幕や音声の形で与えられるため，その抽出に特定のドメインの知識が必要なく，単一の解析技術を適用可能であるという汎用性がある．一方で非言語的な表現である CG の映像美や，カンフーのような詳細な身体表現，音楽表現などは提案手法では相対的に軽視される限界がある．

Kurihara は[2]において，このアプローチにより映像を高速に鑑賞するシステム「CinemaGazer」を提案した．これは市販 DVD などの字幕付き映像を対象として，字幕表示区間はその字幕が読めて理解できる速さで再生し，字幕のない区間はさらに高速に再生するという変速再生方式を実現するものであった（図１）．

しかし，変速再生方式の実現には字幕や音声などの言語情報が映像中のどこで提示されるかが明示されたアノテーション情報が必要であり，[2]ではその応報が予め得られる環境に応用範囲を限定している点が問題であった．そのため，字幕情報が独立したファイルとして抽出可能な市販映画 DVD などでは変速再生が実現できたが，任意の映像に適用することはこれまでできなかった．

本論文では，この「言語情報が映像中のどこに存在しているか」という情報について，映像中の音声を入力として自動的に推定することで，より多くの映像の高速鑑賞を実現する．

我々は Amazon.com の DVD ベストセラーランキングからすべてのジャンルにわたって合計 160 本のコンテンツを収集した．これを学習用と評価用に分割し，字幕付与区間は音声区間，それ以外は非音声区間とみなし GMM(Gaussian Mixture Model)による機械学習および性能評価を行い，実用性を確認した．さらに映像の高速鑑賞を PC やスマートフォンなどの多様なデバイスから行えるよう，Pogoplug を用いたクラウドベースのシステムとして実装した．

本論文では，まず本論文の関連研究を概説し，その後発話区間検出のアルゴリズムについて記述する．それに続いて認識器の構築および評価実験について述べ，構築した認識器を応用した任意の映像の高速鑑賞システムの実装について触れ，考察する．

2. 関連研究

本研究は入力された一連の音響信号に対し「どの部分がなんの音か」を求める音響ダイアライゼーション[3]と呼ばれる問題のシンプルな応用例である．音響ダイアライゼーションは音声認識の高度化と高性能化を支える基盤技術として研究が活発であり，これまで対話コンテンツを対象として「誰がいつ話したか」を推定したり[4]，ポッドキャスト中の注目箇所として笑いや相づちを検出するシステム[5]などが提案されている．コンテンツの要約に応用した例では，Divakaran らや Peker らがスポーツ映像や監視カメラ映像のハイライト抽出のための特徴量の一つに導入している[11][12].

コンテンツの高速鑑賞については，[2][8][13]などの対象の全区間を高速で再生するものと，[11]などの重要な箇所を抽出し「切り貼りの基づく要約」を行うもの，[10]などの並列して複数のストリームを同時に鑑賞するものに大別される．本研究は最初のカテゴリに属する．

メディアの高速再生に関しては，一般的なメディアプレーヤーで既に単純な再生速度の調整は広く実装されているが，約２倍以上の速度では音声の理解が難しくなる．Foulke らは，音高を変えない音声の高速再生が理解の上で有効であることを示した[6]. Vemuri らは音声情報の高速再生時に，その音声の音声認識結果のテキスト情報を提示することによるユーザの情報処理速度の向上の試みを検討した[7].

Cheng ら[8]や Peker ら[13]は映像の変化率に合わせて再生速度を自動調整するアプローチを提案している．我々も類似のアプローチをとるが，我々は映画を扱う上でより重要かつ汎用的である言語モダリティに注目する．本研究においては特に高速再生時の複数モダリティ間の理解可能速度の違いの検討が重要である．

3. 発話区間検出アルゴリズム

本論文で用いる発話区間検出アルゴリズムは，入力信号を音声(voice)かそれ以外(other)の２つのカテゴリに分類するものである．事前学習として，各カテゴリでフレーム（250ms 幅，10ms シフト）ごとに求めた音響特徴量を GMM(Gaussian Mixture Model) でモデル化し，認識時にそれらのモデルに対する尤度を比較することで音の種類を判定する．特徴量には 12 次元 MFCC，正規化した対数パワー，Δ12 次元 MFCC，Δ対数パワーの 26 次元の特徴量を用いる．認識には一般的なビタビデコーダを用い，音の種類およびその区間を同時に推定する．実装には HTK[14]を用いた．

4. 認識器の構築と評価実験

4.1 DVD データセットからの認識器の構築

我々は機械学習のための学習データおよび評価用のデータとして，多数の市販 DVD からなるデータセットを用意した．これは Amazon.co.jp の DVD のベストセラーランキング(2011/12/06 時点)から，洋画，邦画，アニメの各サブ

ジャンル（ドラマ，SF 等）すべてにおいて，上位 4 位までを選択したものである．ここからジャンルが偏らないように 120 本(総時間約 230 時間）と 40 本(総時間約 77 時間)に分割し，前者を学習用データ，後者を評価用データとした．なお，1 つの DVD で複数言語の音声情報が利用可能なものもあるため，それらは別の同じジャンルの別のコンテンツとして扱った．

次に DVD からの字幕情報の抽出を行い，各字幕の開始時刻と終了時刻を得た．これは CUI のものでは VSRip(http://sourceforge.net/projects/guliverkli/files/VSRip/)，GUI のものでは DVDFab (http://www.dvdfab.com/)などを用いることで自動化が可能である．機械学習は，データセット中の字幕表示区間を音声(voice)，それ以外の区間を other の正解例とみなして行った．これにより，実際的なあらゆる声質と雑音を含む音声区間を学習に反映させることが可能である．構築された GMM の混合数は 20 であった．

ここで一般的な映画 DVD における字幕にどのような種類のものがあるかを列挙すると，以下の 3 種類に分けられる．

1. 登場人物のセリフ
2. 背景音の記述（ドアのノック，風の音など）
3. 映像中の文字や文章の母国語訳

我々は音声言語の特徴量に基づく識別を行うので，上記のうち正例として学習したいのは 1.についてであり，2．は不適切である．また，言語情報の検出という意味では 3.も抽出したいが，本研究では音声言語の特徴量を用いて学習を行うので，3．の区間を正例に含めるのは不適切である．本来ならばこれらの不適切な字幕は排除すべきだが，本研究ではそのまま正例として扱い，学習を進めた．

また逆に，字幕の付与されていない箇所についても，学習に影響を与える可能性のある映像箇所として以下のものが挙げられる．

1. 音声を伴うバックグラウンドミュージック
2. 街の喧騒など，字幕の付与されていない音声
3. 登場人物のセリフのうち，叫び声などの字幕の省略された箇所

これらは音声を含む箇所にもかかわらず負例として学習されるので学習性能への影響が予想されるが，本研究ではそのまま負例として扱った．

4.2 評価実験

構築した音声区間自動認識器を，評価用データに適用し性能を評価した．評価は[9]に習い，E_{miss}，E_{fa} および Dialization Error Rate (DER)を応用して行った．これらは以下のように定義される．

$$E_{miss} = T_{miss}/T_{total}$$
$$E_{fa} = T_{fa}/T_{total}$$
$$DER = E_{miss} + E_{fa}$$

ここで T_{total} は総時間数，T_{miss} は音声区間にもかかわらず認識出来なかった総時間，T_{fa} は音声区間でないにも関わらず音声と認識してしまった区間の総時間を示す．これらはそれぞれ検索システム評価における指標である Recall, Precision, および F 値に類似の概念であるが，時系列コンテンツの分類問題の評価へと拡張されており，分母に全コンテンツ長（時間）を用いて計算を行う点が異なる．この特徴により，特定のラベルの認識誤りが起こっても，全コンテンツ長に対する相対的な誤り時間が短ければ影響が少ないという事実を性能評価へと反映可能である．

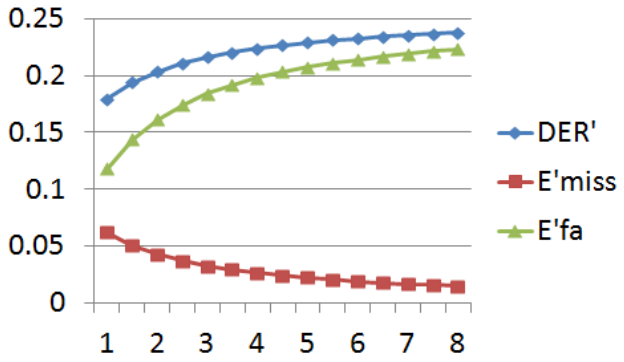

図 2　S_m/S_s を変化させた時の E'_{miss}, E'_{fa}, DER'

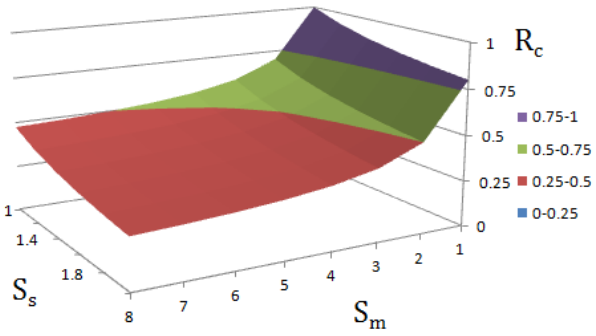

図 3　S_m と S_s を変化させた時の R_c

さらに，我々が採用する変速再生においては，voice と認識された区間は S_s 倍の速度で，またそれ以外(other)と認識された区間は S_m 倍の速度で再生される(通常 $S_m \geq S_s$ である)ため，実際にユーザがコンテンツを鑑賞する圧縮された時系列における E'_{miss}, E'_{fa}, DER'はそれぞれ以下のように変形される．

$$T_{total} = T_m + T_s$$
$$T'_{total} = T_m/S_m + T_s/S_s$$
$$E'_{miss} = T_{miss}/S_m/T'_{total}$$
$$E'_{fa} = T_{fa}/S_s/T'_{total}$$
$$DER' = E'_{miss} + E'_{fa}$$

ここで T_s は全コンテンツ中の音声区間と認識された合計時間，T_m は非音声区間と認識された合計時間を指す．さらにこの時，コンテンツが元の時間に対しどの程度圧縮されたかを表す圧縮率 R_c は以下のように表せる

$$R_c = T'_{total}/T_{total}$$

なお，評価では[9]に習って，正解ラベルに対し前後 250ms

までのずれを許容した．

評価用映画 40 本における S_s および S_m を様々に変化した際に，S_m/S_s を横軸にとった時の E'_{miss}, E'_{fa} および DER'について図 2 に示す．またそれぞれの S_m と S_s の組み合わせについて，その時のコンテンツ圧縮率 R_c の値を図 3 に示す．S_m および S_s の範囲については先行研究[2]を参考に決定した．これらの値の解釈については 6 節で議論する．

5. 任意の映像の高速鑑賞システム

本節では構築した発話区間認識器を応用した映像の高速鑑賞システムのプロトタイプについて述べる．

プロトタイプシステムのインタフェースはパーソナルクラウドシステム Pogoplug 上で実装されている（図 4）．これにより，Pogoplug クライアントさえ準備すれば，鑑賞用デバイスへのシステムのインストール作業を伴わずに映像の高速鑑賞が可能となる．

ユーザが任意の動画を web インタフェース等の Pogoplug クライアントを通じてアップロードすると，クラウド側でシステムが検知し，発話区間の認識および[2]による変速再生映像の生成，およびエンコードを行う．ユーザは web インタフェースや iPhone などのモバイルデバイス上からストリーミング再生を鑑賞することができる（図 4）．音声区間および非音声区間における再生速度の指定，もしくは総鑑賞時間の指定はアップロードするフォルダ名として指定する．たとえば，非音声区間が 4 倍，音声区間が 2 倍で再生する場合は"4_2"という名前のフォルダ，また総鑑賞時間が 30 分で音声区間を 2 倍で再生する場合は"30min_2"という名前のフォルダを用いる ．

また，PC 上で専用のメディアプレーヤーを用いることにより，音声区間と非音声区間の再生速度のリアルタイムな調節，および音声を聞きのがしてしまった際に数秒だけ遡って通常速度で再生し直す"skip back "機能を実現し，より快適な高速鑑賞を可能とした．

図 4　Pogoplug 上でのストリーミング再生による高速鑑賞

6. 考察

6.1 認識器の性能

4 節の評価実験について，得られた性能数値がどのような意味を持つかを考察する．

全体的な結論としては，言語やジャンルに依存しない実用的なよい性能が得られたといえる．図 2 で総合的な性能指標である DER'に注目すると，E'_{miss} が低いものの E'_{fa} が比較的高いため，0.18 から 0.22 という値になってしまっている．しかし，本研究の応用局面ではまずは E'_{miss} を重視すべきである． 我々の高速映像鑑賞システムは「言語提示区間を言語理解可能な速さで再生し，それ以外の区間をさらに高速に再生する」というアルゴリズムであり，コンテンツ中のすべての言語情報をユーザに理解可能な速さで提供することが映像鑑賞のエンタテイメント性の保存の上で非常に重要な要求仕様である．E'_{miss} が低くなければ，より多くの言語箇所が高速で再生されてしまうため理解不能になり，5 節で実装した skip back 機能の使用頻度が増え，ユーザの満足度を低下させる要因になる．しかし現状で E'_{miss} はほとんどの範囲（$S_m/S_s \geq 1.5$）で 0.05 以下の非常に低い値であり，充分実用的な性能であるといえる．

一方で E'_{fa} が比較的大きい，すなわち誤検出区間について本来あるべき S_m 倍ではなく比較的低速な S_s 倍での再生になってしまう問題については，圧縮率 R_c が最終的に低くなれば問題ない．ここで比較対象として市販のビデオレコーダを考える．これらの上で再生速度を 1.2 倍から 2.0 倍に調整できることは，R_c が 0.83 から 0.5 程度であることを意味している．図 3 を見ると，S_s が 1.2 倍から 2.0 倍の範囲であっても広い範囲の S_m で圧縮率が 0.5 を下回っており（赤い領域），効率的な鑑賞時間の圧縮が実現されていることがわかる．

本認識器を用いて評価用映像 を実際に高速鑑賞してみた結果，音声区間を 1.8 倍，非音声区間を 4 倍で変速再生したところ，体感ではほぼすべての音声情報がストレスなく理解可能であり，確かに非音声区間が誤認識され 1.8 倍の低速で再生される箇所は発見されたが，それは鑑賞のストレスとはならなかった．非音声区間はより高速での再生であったが，（映像の細部はわからないものの）シーン全体としての意味合いを理解する上では概ね問題なかった．そして全体として 63%程度の鑑賞時間の削減が可能であった．このことから，構築された認識器は実用的な性能を持っていることが示唆された．よりフォーマルなユーザスタディは今後の課題である．

また 4.1 節 で述べたように，（風の音に字幕が付与されたり，街の喧騒に字幕がなかったりといった）一部の正例・負例の与え方が不完全である点による性能低下の可能性がある．より正確な認識器構築および性能評価のために，人力によりそれらの不適切な正例・負例を排除する必要がある．これは今後の課題である．

6.2 映像の高速鑑賞の限度について

より高速に映像を鑑賞したい場合は，対象映像中の音声に対し字幕が付与されている方が望ましい．図 5 は先行研究[2]における，音声，字幕，主たる映像の各モダリティの鑑賞速度と鑑賞可能人数比率の関係を再掲したものである．

これによると音声，字幕，主たる映像の順に鑑賞可能な限度の速度が上がっていき，個人差もこの順で大きくなった．

注目すべきは，音声の平均鑑賞可能限界が 1.55 倍なのに対し，字幕の平均鑑賞可能限界が 5.91 倍と大きな値であることである．この実験で用いられた映像コンテンツの性質や実験協力者のバイアスがこれらの値にはかかっていることを踏まえても，通常人々は音声よりも文章の方を高速に理解できるという点には異論は少ないであろう．

したがって鑑賞対象のコンテンツの音声区間に字幕が付与されていれば（我々は音声で言語区間を抽出するので，字幕は映像に埋め込まれていてもよく，DVD のように陽に抽出できる必要はない．），より高速に映像を鑑賞できる可能性が高まる．海外映画など，外国語で作られた映像コンテンツは，正確に各発話に対応する字幕が付与されていることが多いため，この仮定は現実的なものである．言語情報が音声と文字の２つのモダリティで与えられることには冗長性があるが，現状の提案システムにとっては有効である．

一方で，映像中から字幕提示箇所を検知する画像処理のアプローチと統合することにより，より有効なシステムとすることも考えられる．これは今後の課題である．

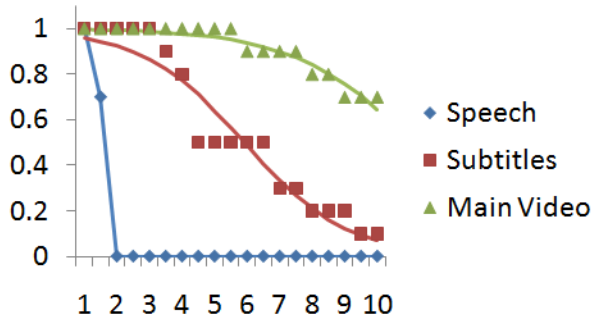

図 5　鑑賞の速度倍率（横軸）と実験協力者の鑑賞可能人数比率（縦軸）の関係[2]

7. まとめ

本論文では，多くの映画に適用可能であると考えられる高速鑑賞方法として変速再生方式を採用する際の前処理に必要な，対象映像中の音声区間と非音声区間を自動的に分離する技術ついて，市販の字幕付き DVD160 本のデータセットを用いて，MFCC を特徴量とする GMM による認識器として実装，評価した．その結果実用的な性能をもつことが示され，この認識器を組み込んだクラウド型の映像の高速鑑賞システムのプロトタイプを実装した．

謝辞　本研究の一部は科研費(23700155)の助成を受けた．

参考文献

1)　INFOGRAPHIC: What Happens Online Every 60 S. http://www.scribbal.com/2011/06/infographic-what-happens-online-every-60-s/
2) Kazutaka Kurihara. CinemaGazer: a System for Watching Videos at Very High Speed. *In Proc. of AVI'12*, pp.108-115, 2012.
3) D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. *In Proc. of ICASP'05*, pp.V-953-956, 2005.
4) M. Kotti, V. Moschou, and C. Kotropoulos. Review: Speaker segmentation and clustering. *Signal Processing*, 88(5):1091-1124, 2008.
5) K. Sumi, T. Kawahara, J. Ogata, and M. Goto. Acoustic event detection for spotting hot spots in podcasts. *In Proc. of ISCA'09*, pp.1143-1146, 2009.
6) Foulke, W., and Sticht, T.G. "Review of research on the intelligibility and comprehension of accelerated speech," *Psychological Bulletin*, 72, pp.50-62, 1969.
7) Vemuri et al. "Improving speech playback using time-compression and speech recognition," *In Proc. of CHI'04*, pp.295-302, 2004.
8) Cheng et al., "SmartPlayer: User-Centric Video Fast-Forwarding," *In Proc. of CHI'09*, pp. 789-798, 2009.
9) Diarization Error Rate. http://www.xavieranguera.com/phdthesis/node108.html.
10)　Manfred et al., "Instant video browsing: a tool for fast non-sequential hierarchical video browsing," *In Proc of USAB'10*, pp. 443-446, 2010.
11)　Divakaran, A., and Otsuka, I., "A video-browsing-enhanced personal video recorder," *In Proceedings of IEEE International Conference of Image Analysis and Processing Workshops (ICIAPW)*, pp.137-142, 2007.
12)　Peker, K. A., and Divaskaran, A., "An extended framework for adaptive playback-based video summarization," *SPIE Internet Multimedia Management Systems IV 5242*, pp.26-33, 2003.
13)　Peker, K.A., Divakaran, A., and Sun, H., "Constant pace skimming and temporal sub-sampling of video using motion activity," *In Proceedings of IEEE International Conference on Image Processing (ICIP)*, Vol.3, pp.414-417, 2001.
14)　Hidden Markov Model Toolkit. http://htk.eng.cam.ac.uk/