テレビ視聴行動を再現するエージェントシミュレータの構築に向けた基礎分析

松田　裕貴[†,††]　榊原　太一[††]　真弓　大輝[†]　松田　裕貴[†]　安本　慶一[†]

† 奈良先端科学技術大学院大学　〒 630–0101 奈良県生駒市高山町 8916-5
†† 読売テレビ放送株式会社　〒 540–8510 大阪市中央区城見 1 丁目 3 番 50 号
E-mail: †{mayumi.daiki.mb9,yukimat,yasumoto}@is.naist.jp, ††{hiroki.matsuda,taichi.sakakibara}@ytv.co.jp

あらまし　インターネットに接続されたテレビから個人を特定しない形で視聴に関するログを収集する「非特定視聴履歴データ」が近年注目されており，近畿圏では約 400 万台分のテレビのデータを複数のテレビ局が収集している．個人を特定しない形式ではあるものの IP アドレスや視聴時刻等が含まれることから，改正個人情報保護法における個人関連情報に該当するため，一定期間でのデータ削除や外部公開の禁止といった運用が長期的な分析を妨げている．そこで本研究では，テレビの非特定視聴履歴データや視聴に関する統計情報を用いて，人・テレビ受像機・放送局のモデル化を行い，在阪 4 局の放送局における非特定視聴履歴データの再現を可能とするシミュレータの構築を目指している．本稿ではその設計にあたっての基礎的な分析結果について報告するとともに，シミュレータの設計について検討する．
キーワード　テレビ，ビッグデータ，視聴者行動，可視化，インターネット検索，IoT，シミュレータ

1 はじめに

近年，テレビをインターネットに接続して，データ放送コンテンツや動画配信サービスを利用する視聴者が増えている．インターネット接続されたテレビからは，視聴者がいつ，どの番組を視聴していたのかの情報を収集することが可能であることから，テレビ放送局やテレビ製造メーカがテレビの視聴履歴データを収集している．その中でも非特定視聴履歴データは，放送局が個人を特定しない形式で収集しているデータを指し，従来の視聴率などのデータに加えて，新たな価値を生み出すビッグデータとして，放送局のみならず，スポンサーや広告代理店からも利活用が期待されている．

しかし，非特定視聴履歴データは個人情報ではないものの改正個人情報保護法における個人関連情報に該当し，更に個人の趣味趣向を把握しうるため，取扱いの配慮が求められている．一般社団法人放送セキュリティーセンター（以下，SARC）は，「オプトアウト方式で取得する非特定視聴履歴の取扱いに関するプラクティス」[1] を作成し，放送局以外への第三者提供禁止やデータ保有期間を定め，期間を過ぎるとデータ削除する等のルール整備を行っている．これに伴い，読売テレビでは前述のルールに則る方法にて 2019 年 12 月から非特定視聴履歴データの収集を開始している．

現在，在阪 5 つの民間放送局では，非特定視聴履歴データの新たな価値創出に向けて，放送局間での非特定視聴履歴データの連携技術検証と連携データ利活用に向けた共同技術実験 [2] を実施している．本実験では，前述の SARC プラクティスに準じた取扱いルールを遵守するために，お互いに信用できる放送局の間でのみデータを交換することでデータの安全性を担保している．しかし，一定期間でのデータ削除や外部公開禁止といった条件が，長期的なデータ分析や利活用を妨げていることが課題となっている．

そこで本研究では，テレビの非特定視聴履歴データや視聴に関する統計情報を用いて，人・テレビ受像機・放送局のモデル化を行い，近畿 2 府 4 県の同一放送エリアを持つ 4 つの民間放送局における非特定視聴履歴データの再現を可能とするシミュレータ構築を目指す．本稿では，そのシミュレータ設計にあたっての基礎的な分析結果について報告するとともに，シミュレータを提案する．

2 関連研究

本研究では，視聴者のテレビ視聴行動とその視聴行動から生成される視聴履歴データの関係について検討する．そこで，テレビ視聴がどのように行われているかの関連研究や今回の検討の土台となる在阪視聴データ連携実験，最後に統計情報から個票データを合成する関連研究について紹介する．

2.1 テレビ視聴履歴データの分析

視聴履歴データの利活用について，様々な研究が行われている．菊池ら [3] は，放送局が収集している非特定視聴履歴データではなく，東芝製テレビに絞った分析を行っており，インターネットに結線された東芝製テレビ視聴者が番組ジャンル別にどのような視聴傾向を持っているのかを明らかにした．また，水岡ら [4] は，菊池ら同様に東芝製テレビに絞ったテレビの視聴パターンにより分類する手法を提案している．

放送局が収集している非特定視聴履歴データについては，筆者らの研究グループによって研究されている．松田ら [5] はテ

表 1　在阪 4 局の非特定視聴履歴データ保有期間

放送局	保有期間
A 局	23ヵ月
B 局	18ヵ月
C 局	60ヵ月
D 局	18ヵ月

レビ CM 視聴がその後のインターネット検索行動に与える影響について分析を行っている．また，吉村ら [6] は CM の完視聴率にどのような地域差が存在するのか分析している．このように利活用に向けた研究は行われているが，各放送局が個別に収集している本データ単体を分析したとしても，1 つの放送局におけるテレビ視聴状況しか把握出来ないため，新たな価値を生み出すには至っていないのが現状である．

2.2　在阪放送局によるテレビ視聴データ連携に関する共同技術検証実験

前節にある通り，放送局各社も 1 つの放送局におけるテレビ視聴状況しか分析できない場合，新たな価値を生み出すことは難しいと考えており，2021 年度に在阪の 4 つの民間放送局による初めての共同技術実験 [7] が行われた．しかし，非特定視聴履歴データは，各放送局が取得にあたりコストや仕様を検討し，各放送局の経営判断により取得が開始されているため，放送局ごとに取得方式やデータフォーマットが異なっており，放送局間のデータを統合することは難しい．これらのデータを統合する手法は，松田ら [8] により提案されており，実際に共同実験ではデータの統合に成功している．その結果，従来は単独の放送局における視聴状況しか把握できなかったが，放送局を横断した視聴状況の把握が可能となった．

2.3　非特定視聴履歴データの課題

各放送局が収集している非特定視聴履歴データは，IP アドレスを持っており，改正個人情報保護法における個人関連情報に該当する．その為，在阪の民放 4 局においてもそれぞれデータ保有期間を定めており，保有期間を過ぎたデータは速やかに抹消している．

また，SARC プラクティスにより放送局以外へのデータの持ち出しは認められておらず，放送局間でのデータ交換しか実施できない．このため，長期的なトレンドの分析であったり，広告代理店やスポンサーなどが直接分析することができないようになっている．

3　テレビ視聴行動を再現するエージェントシミュレータ

前述の通り，非特定視聴履歴データについては，分析により新たな知見の獲得が期待できるものの，個人関連情報に当たることから長期的なデータ保存・分析が難しいといえる．そこで本研究では，テレビの非特定視聴履歴データや視聴に関する統計情報を用い，視聴者とテレビのモデル化を行うことで，非特定視聴履歴データの再現を可能とするエージェントシミュレータを構築できれば，特定個人と関連する情報を含まないデータセットを構築可能と考えた．

この提案は，原田ら [9] の提供する「合成人口データ」に着想を得ている．合成人口データでは，国勢調査結果から各世帯の個票データの合成を行っている．統計情報に準拠した個票データではあるが，個人情報ではないため，第三者提供可能という特徴を持っている．

本研究では，非特定視聴履歴データを合成することで長期的なトレンド分析や放送局以外が直接分析できるデータセットの合成を目指す．

3.1　予 備 分 析

テレビ視聴行動を再現するエージェントシミュレータを構築するにあたり，後述のデータセットを分析することで実際の視聴行動分析と視聴行動から生まれるテレビ視聴履歴データの特徴について分析を実施した．

3.1.1　データセット

本分析では，在阪 4 局の非特定視聴履歴データ，テレビ放送実績データを利用する．以下では各データセットについての詳細について述べる．

在阪 4 局のテレビ非特定視聴履歴データ　本データセットは，2022 年度に新たに実施した，在阪 5 局の民間放送局で非特定視聴履歴データの連携技術検証と連携データ利活用に向けた共同技術実験 [2] により収集された．本分析では，近畿 2 府 4 県を放送エリアに持つ在阪の 4 つ [1] の放送局（毎日放送，朝日放送テレビ，関西テレビ，読売テレビ）が各々独自に収集している非特定視聴履歴データのうち，4 つの放送局すべてにおいて同一のテレビ受像機と推定できたものを使用する．各放送局が収集している非特定視聴履歴データは，サービス圏内である近畿 2 府 4 県にあるインターネット接続されたテレビのうち，オプトアウトしていないテレビが対象となっており，それぞれ約 300 万台の規模となる．データには IP アドレス，テレビデバイス ID，郵便番号，視聴開始時刻，視聴終了時刻，テレビメーカ ID などが含まれている．データ期間は 2022 年 10 月～2022 年 11 月（2ヶ月間）である．今回は，これらの非特定視聴履歴データのうち，松田らが提案した NNTM アルゴリズム [8] にて 4 つの放送局すべてにおいて同一テレビと推定できた約 120 万台を対象とする．

テレビ放送実績データ　本分析で使用するテレビ放送実績データは，実際にテレビ放送を実施している読売テレビから提供されたもので，読売テレビにおけるテレビ番組の開始・終了時刻やテレビ番組内で CM が放送された時刻が含まれている．また，各テレビ番組に関して番組名が含まれており，著者らによって番組ジャンルの情報を付与されている．データ期間は 2022 年 10 月～2022 年 11 月（2ヶ月間）である．

3.1.2　分析 A：テレビ受信機の状態遷移確率

在阪 4 局のテレビ視聴履歴データを分析することで各テレビ受信機の次の状態にどのような確率で遷移するのか分析を行っ

1：テレビ大阪については，大阪が放送エリアとなっており他と大きく異なることから本分析の対象外とした．

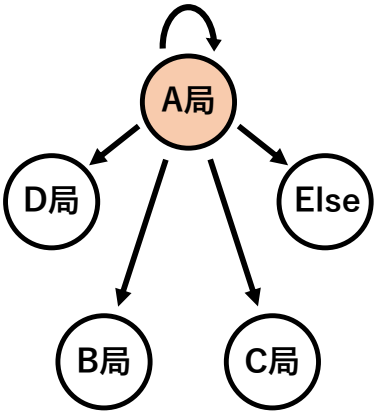

図 1　A 局中心の状態遷移図

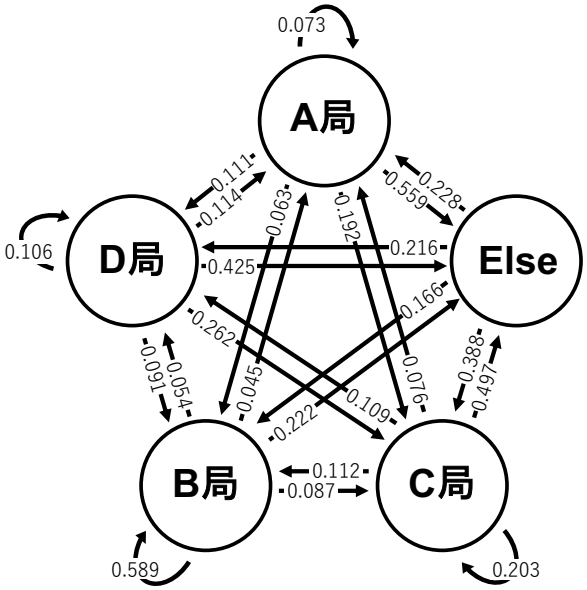

図 2　各放送局の状態遷移図（2022 年 10 月 21 日 19 時 30 分）

た．在阪 4 局のテレビ視聴履歴データからは，4 局のうちどこか 1 局を視聴している，もしくは 4 局のいずれも視聴していない，ということが分かる．

図 1 に A 局を中心とした状態遷移図を示す．A 局を視聴しているテレビは次の遷移先としては，B 局，C 局，D 局，それ以外，もしくは A 局を継続視聴する，のいずれかとなる．今回，実際の在阪 4 局のテレビ視聴履歴データを分析することでそれぞれの状態からの遷移確率を分析した．

本分析では，テレビ視聴履歴データを 1 分単位に区切り，各 1 分において次の 1 分間にそれぞれのテレビ受像機がどの状態に遷移しているのか分析し，それを確率として算出した．また，その分析を 4 局 [2]のそれぞれにおいて実施した．図 2 は，2022 年 10 月 21 日 19 時 30 分～31 分（1 分間）に，視聴者がどのように状態遷移したかの状態遷移図をそれぞれの状態を起点にして 1 つの図に示したものである．A 局からの状態遷移は，1 分後に A 局を見続けている割合は約 7.3 %であり，4 局以外の視聴であったり電源 OFF 状態である Else へ遷移するテレビが

2：Else は放送局からの離脱と流入先のため，自己ループするデータは存在しない．

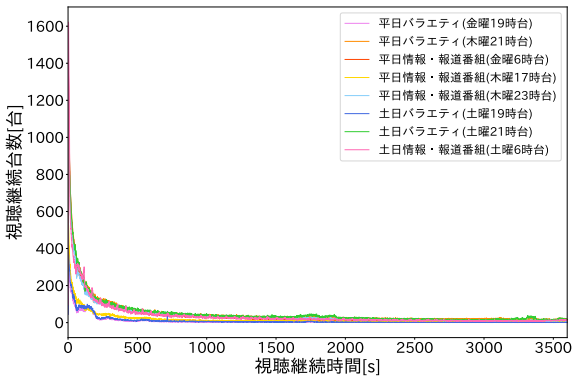

図 3　継続視聴時間別の視聴台数比較（60 分間）

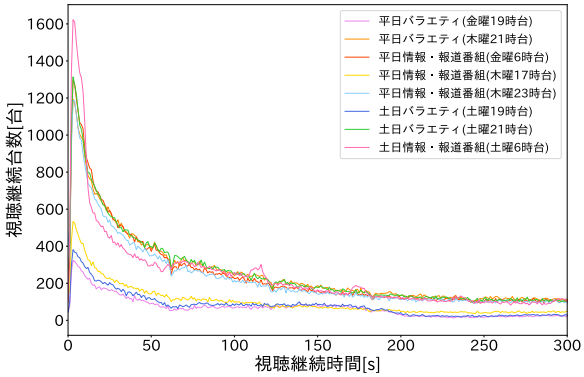

図 4　継続視聴時間別の視聴台数比較（5 分間）

55.9 %となっているということがわかる．各局からの遷移を比較すると，各局の状態遷移の傾向は類似しているものの，多少の違いが存在することが分かった．例えば，B 局については視聴継続が比較的多く，A 局と D 局は視聴継続が少ないことが分かる．

3.1.3　分析 B：番組ジャンル別の継続視聴分数

次に視聴継続時間を再現するための分析を実施した．特に時間帯やその時間帯に放送されている番組のジャンルにより継続視聴時間には差異が発生すると想定される．そこで読売テレビの視聴履歴データを用いて，平日と土日にそれぞれ放送されているバラエティ番組と情報・報道番組を対象に継続視聴分数を分析した．具体的には平日バラエティ番組 2 種類，平日情報・報道番組 3 種類，土日バラエティ番組 2 種類，土日情報・報道番組 1 種類の合計 8 種類の分析をしている．すべてのデータにおいて，4 回放送分の平均値を採用している．調査結果を図 3 に示す．図 3 では，横軸は各テレビの視聴継続時間を示している．そして，縦軸は視聴継続台数を示している．

全ての対象番組において，時間が経過するとともに視聴継続台数が減少していることが確認できる．

次に図 3 の最初の 5 分間を拡大したものを図 4 に示す．時間経過とともに減少していく傾向は同じであるが，番組ジャンルや時間帯によって傾向の違いがあることがわかる．

次にバラエティ番組だけを抜粋比較した結果を結果を図 5 に

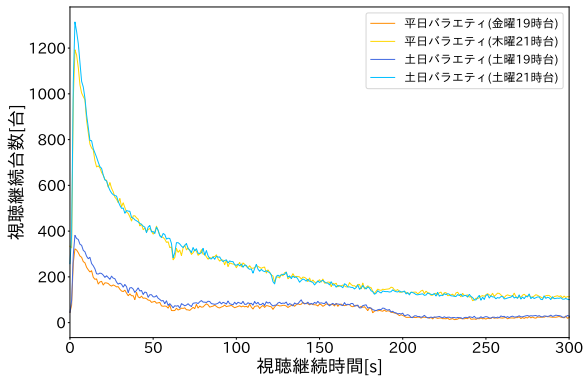

図 5　バラエティ番組における継続視聴時間別の視聴台数比較（5 分間）

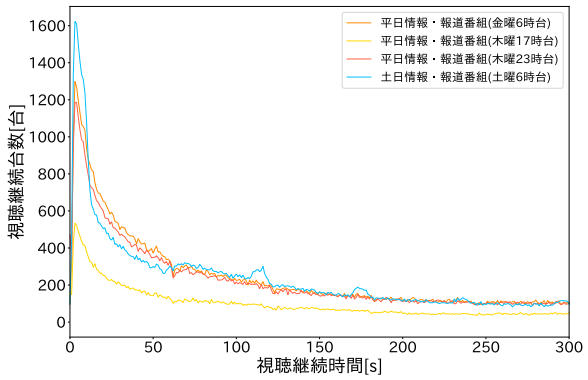

図 6　情報・報道番組における継続視聴時間別の視聴台数比較 (5 分間)

示す．

最後に図 5 では，平日 19 時台と土日 19 時台，平日 21 時台と土日 21 時台の 4 番組を比較している．19 時台と 21 時台で傾向が違うことが見てとれる．21 時台においては，50 秒以内の視聴が多く，19 時台では視聴継続時間が長くなっても台数減少が滑らかになっている．

次に図 6 では，情報・報道番組を比較した結果を示している．同じジャンルの番組であっても放送時間帯によって傾向に大きな差異があることが分かる．特に平日 6 時台と土日 6 時台では差が顕著に出ており，先ほどのバラエティ番組の傾向とは大きくかけ離れていることがわかる．

番組ジャンルや放送曜日・時間帯により継続視聴分数の傾向は違うので，それぞれ分けて検討をする必要があることが確認できた．

3.1.4　分析 C：IP アドレス変化期間の分析

テレビ視聴履歴データは，インターネット結線されたテレビ受像機を対象にデータを収集している．一般的にテレビ受像機は家庭のインターネット回線を経由して結線されており，グローバル IP アドレスは定期的に変化していく．テレビ視聴行動を再現するエージェントシミュレータでは，非特定視聴履歴データの再現を目指すので，IP アドレスの変化も実際のテレビ視聴履歴データから調査分析した．データは，読売テレビのテレビ視聴履歴データから 5,000 台を無作為抽出したものを利用し，調査期間は 1 週間とする．

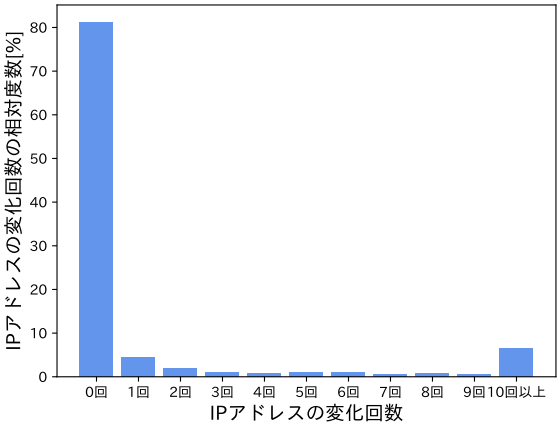

図 7　IP アドレス変化回数の相対度数分布

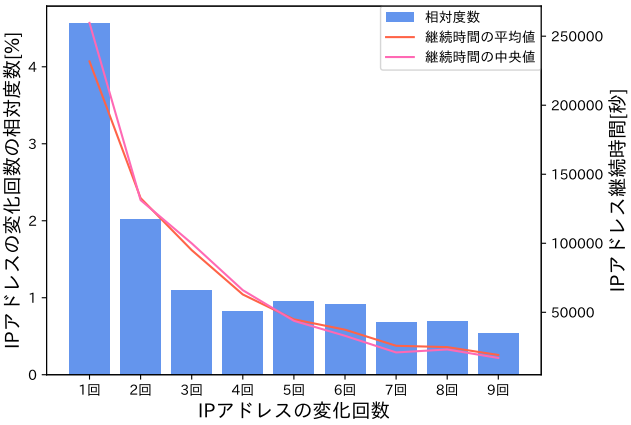

図 8　IP アドレス変化回数の相対度数分布と変化時間（1 回以上を抜粋）

まず，1 週間の IP アドレス変化回数の相対度数分布を図 7 に示す．全体の約 80 %のテレビ受像機において IP アドレスは変わっていないことがわかる．

次に IP アドレスが変わらなかった端末と 10 回以上変化した端末を除いたグラフを図 8 に示す．また，図 8 には変化回数別の変化までの IP アドレス継続時間の平均値と中央値も示している．IP アドレスは 1 回変化する台数が多く，その次に 2 回変化台数が続く，その後 3 回目以降はあまり変化がないことが確認できた．また，IP アドレスが変化するまでの継続時間は変化回数が減少するにつれて，同様に単調減少することが確認できた．

次に IP アドレスが 1 回以上変化するテレビ受像機が 1 週間のうち，どのタイミングで IP アドレスが変化しているのかを分析した．その結果を図 9 に示す．変化回数が多いと継続時間が短く，変化回数が少ないと継続時間が長くなっていることが確認できる．これらのように IP アドレスが変化する回数と周期に偏りがあることが確認できた．

3.2　テレビ視聴行動を再現するシミュレータの検討

前節のデータ分析結果から，視聴者のモデル化とテレビ受像

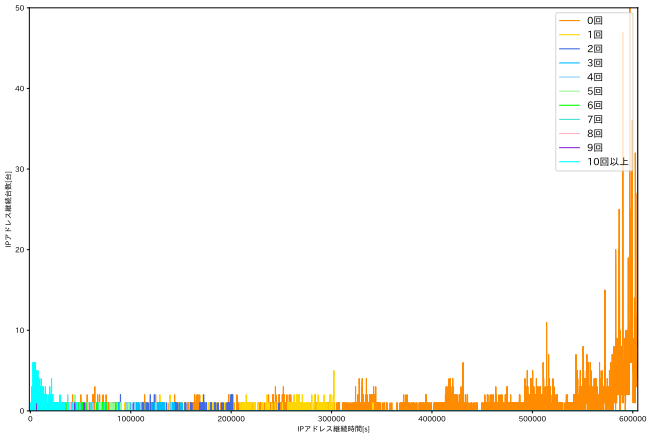

図 9　IP アドレス変化回数毎の経過時間（1 回以上を抜粋）

機のモデル化を行い，それらの組から成るエージェントを構成することで，疑似的な非特定視聴履歴データ合成の品質向上が図れると考えられる．ここでは，視聴者モデル・テレビ受像機モデルからなるエージェント，テレビ視聴行動をロギングするオブザーバ（本論文では，テレビ局各社を想定する），およびテレビ視聴に関する統計データ，国勢調査の統計データ，テレビ番組等のメタデータを用いて，各家庭に設置されているテレビ受像機の視聴に関する個票データを合成するシミュレータの実現方法について検討する．

3.2.1　シミュレータの全体構成

シミュレータの全体構成案を図 10 に示す．エージェントモデルは，視聴者モデルとテレビ受像機モデルから構成される．視聴者モデルは，非特定視聴履歴データと放送実績データ，国勢調査データから設計し，テレビ受像機モデルは非特定視聴履歴データから設計する．シミュレータを動かすために，国勢調査データ，放送メタデータと個人視聴率データを使う．まずは，エージェントモデルに国勢調査データを入力することで近畿地方を再現するエージェントを生成する．次に各エージェントの視聴行動を放送メタデータを入力することで生成する．この時，視聴に関する統計データである個人視聴率データを制約条件として利用する．生成された視聴行動を視聴履歴データの特徴をモデル化した放送局モデルに入力し，視聴履歴データを合成する．

個票データ生成にあたっては，視聴行動に関する状態遷移の制約として，合成視聴履歴データの集計結果が「個人視聴率データ」と合致することを条件として設定する．

3.2.2　データセット

本シミュレータに入力を想定している 3 つのデータセットについて説明する．

放送メタデータ　読売テレビの放送実績データだけでなく，株式会社エムデータが作っている在阪 4 局すべての番組・CM メタデータの入力を想定している．株式会社エムデータは，在阪 4 局の放送を 24 時間監視し，実際に番組が始まった時刻や終わった時刻だけでなく，CM 開始時刻や CM 終了時刻をメタデータ [10] として作成している．

個人視聴率データ　本研究で使用する個人視聴率データは，ビデオリサーチが作成しているデータを使っている．ビデオリサーチでは，近畿にある 1,200 世帯の約 2,100 人を対象に調査を実施している．本データは統計データであるが，調査対象世帯のテレビが 1 分単位でどのチャンネルを視聴していたか，もしくはテレビを視聴していなかったのかを統計化したデータである．

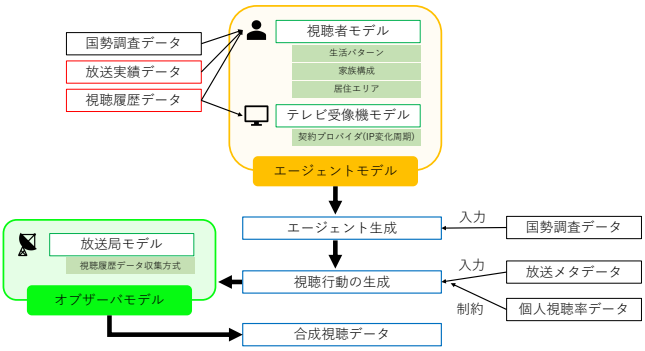

図 10　シミュレータ処理の流れ

表 2　各放送局の視聴履歴データ特徴

	A 局	B 局	C 局	D 局
方式	ビーコン	ビーコン	From-To	From-To
ビーコン間隔	60 秒	15 秒	-	-

国勢調査データ　本研究で使用する総務省統計データで，総務省が 5 年毎に実施している国勢調査のデータを使っている．本データ [11] では，近畿における人口を使っており，20,541,441 人となっている．データは 2020 年に調査されたものである．

3.2.3　視聴者モデルの検討

視聴者のテレビ視聴傾向は，3.1.3 項の分析結果から，番組ジャンルや放送曜日・時間帯によって変化する可能性が示されている．その為，視聴者モデルを生成するには，3.1.2 項項で在阪 4 局における非特定視聴履歴データから分析した状態遷移確率の時系列データおよびテレビ放送実績データを用いて，視聴行動ルールを作成する必要がある．これに加え，視聴者の属性（学生，労働者，高齢者，など）によって視聴行動パターンには大きな差異が生じると考えられるため，視聴行動ルールには視聴者の属性による変化が生じるように設計を行う必要がある．

3.2.4　テレビ受像機モデルの検討

3.1.4 項で実際の視聴履歴データを分析した通り，生成される視聴履歴データの IP アドレスは一定ではないことが分かっている．これは視聴者の契約しているプロバイダなどの諸条件により，様々な周期で変化している．

このテレビ受像機特有の特徴を再現するテレビ受像機モデルを構築し，前述の視聴者モデルと組み合わせることでエージェントモデルを構築することで，エージェント生成を行うことが可能となる．

3.2.5　放送局モデルの検討

本項では生成された視聴者行動から精緻な視聴履歴データを合成するために，放送局モデルの検討を実施する．放送局が収集する視聴履歴データは，データ放送プログラムを利用して収集しており，ビーコン方式と From-To 方式に分類され，表 2 の

ようになっている．ビーコン方式では，実際の視聴行動とそこから生成される視聴履歴データが一致しない場合がある．ビーコン方式は，視聴開始からビーコン間隔までの秒間のうち，ランダムでビーコン送信が開始される．これは，番組開始タイミングでサーバへの負荷を減らすために行っているが，ビーコン送信が開始されるまでにチャンネル遷移が行われると視聴履歴データが生成されないことになる．このような放送局毎の特徴を考慮したモデルを構築することで，精緻な視聴履歴データを合成することが可能となる．

4 まとめと今後の取り組み

本研究では，テレビ視聴履歴データに着目し，チャンネル遷移確率や IP アドレス変化周期，番組ジャンル別の視聴傾向の違いについて分析した．その結果，チャンネル遷移の確率や IP アドレス変化周期にはバラつきがあることを発見した．他に番組ジャンル別の視聴傾向を分析することで曜日・時間帯・ジャンルによりそれぞれ視聴傾向が異なることを明らかにした．これらの分析結果から得られたデータを活用したテレビ視聴履歴データを合成するシミュレータ設計を検討した．

今後は，テレビ視聴履歴データの更なる分析を行い，曜日・時間帯・ジャンルによる視聴傾向を分類し，その分類パターンを視聴者モデルに反映させる必要がある．また，テレビの視聴に関する統計情報である個人視聴率の割合を制約条件に，在阪 4 局のテレビ視聴履歴データから得られた状態遷移確率に沿って視聴者モデルを生成する際に齟齬が発生しないように調整する必要がある．これらを考慮したシミュレータ実装を今後の課題としたい．

謝　　辞

本研究にあたって多大なご協力をいただきました，株式会社毎日放送，朝日放送テレビ株式会社，関西テレビ放送株式会社の皆様に，この場を借りて深く感謝申し上げます．

文　　献

[1] 一般財団法人放送セキュリティセンター視聴関連情報の取扱いに関する協議会．オプトアウト方式で取得する非特定視聴履歴の取扱いに関するプラクティス（ver2.1）．https://www.sarc.or.jp/documents/www/NEWS/hogo/2021/optout_practice_ver2.1.pdf, 2021.

[2] 読売テレビ放送株式会社．「テレビ視聴データ連携に関する共同技術検証実験（2022 年度）」について．https://www.ytv.co.jp/privacy/experiments2022/index.html, 2022.

[3] 菊池匡晃, 坪井創吾, 中田康太．大規模テレビ視聴データによる番組視聴分析．デジタルプラクティス, Vol. 7, No. 4, pp. 352–360, 2016.

[4] 水岡良彰, 中田康太, 折原良平．大規模テレビ視聴データによる視聴パターン推移の分析．人工知能学会全国大会論文集, Vol. JSAI2018, pp. 1P203–1P203, 2018.

[5] 松田裕貴, 榊原太一, 木俣雄太, 鳥羽望海, 真弓大輝, 松田裕貴, 安本慶一．テレビ視聴における非特定視聴履歴データとインターネット検索データの関係性分析．第 14 回データ工学と情報マネジメントに関するフォーラム（DEIM ’22）, pp. 1–6. 日本データベース学会, 2022.

[6] 吉村啓, 水本旭洋, 榊原太一, 松田裕貴．テレビ視聴時の CM 離脱と地域傾向分析．人工知能と知識処理研究会, 第 121 巻, pp. 43–48, 2022.

[7] 読売テレビ放送株式会社．「テレビ視聴データ連携に関する共同技術検証実験」について．https://www.ytv.co.jp/privacy/experiments/index.html, 2021.

[8] 松田裕貴, 榊原太一, 松田裕貴, 水本旭洋, 安本慶一．放送局を横断する大規模テレビ視聴履歴データの統合手法の提案と実践．デジタルプラクティス, Vol. 4, No. 1, pp. 1–11, 2023.

[9] 原田拓弥, 村田忠彦．国勢調査結果を用いた全ての一般世帯と施設などの世帯を含む全世帯の合成．第 24 回社会システム部会研究会, pp. 17–23. 計測自動制御学会, 2021.

[10] 株式会社エム・データ．TV メタデータとは．https://mdata.tv/metadata/, 2022.

[11] 総務省統計局．令和 2 年国勢調査　調査の結果．https://www.stat.go.jp/data/kokusei/2020/kekka.html, 2020.