

SymPhonicAudio: 複数のスマートフォンを同期し 現実空間への音像定位を動的に行う手法の提案

金子翔麻^{†1} 渡邊恵太^{†1}

概要: AR において、視覚による AR だけでなく、空間音響によって現実世界を拡張する「音の AR」も重要である。ARKit や ARCore のリリースによって、スマートフォンによる AR を手軽に実現できる環境が普及した。スマートフォンによる AR はその手軽さや 1 人 1 台近く持っている状況から AR 空間を複数人で共有することに価値が見出せる。しかし、「音の AR」の主流であるヘッドホンによるバイノーラル方式ではコミュニケーションの取りにくさがあり、体験を共有しにくい。そこで本研究では、「音の AR」による体験の共有を目的とし、複数のスマートフォンを同期し現実空間への音像定位を動的に行う手法 SymPhonicAudio を提案する。評価実験では本手法とバイノーラル手法の 2 つの場合で比較した。結果、多人数で協力して遊ぶ「音の AR」を用いたゲームにおいて、音の方向性のわかりやすさはバイノーラルの方が優れていたが、コミュニケーションの取りやすさや一体感があるという点において本手法が優れていた。

1. はじめに

ARKit[1]や ARCore[2]のリリースによって、スマートフォン内蔵のカメラとモーションセンサのみで AR 体験をメーカーレスに実現できる環境が爆発的に普及した。ARKit アプリの潜在的なユーザー数はリリース時で 3 億 8,100 万人、2017 年末には 5 億 500 万人、2020 年には 8 億 5,000 万人になると推測されている[3]。AR において、空間音響によって現実世界を拡張する「音の AR」も重要である。特に画面領域が小さいスマートフォンにおいては、画面外の「見えないものへのインタラクション」を補助するものとして有効である[4]。「音の AR」を実現する手法はヘッドホンによるバイノーラル方式が主流である[5][6]。そのほかに指向性スピーカを用いた「音の AR」手法[7]や、360 度動画や VR 動画で用いられているアンビソニックス[8]が挙げられる。また、空間音響を実現する手法として、サラウンド方式[9][10]や点スポット再生[11]などが挙げられる。

スマートフォンによる AR は HMD のような没入感がない代わりに、その手軽さや 1 人 1 台近く持っているという状況から AR 体験の共有に適している。周囲の人と共有できる「音の AR」により、AR 空間での体験共有をより強化し、周囲の人との一体感を生み、コミュニケーションを促進させる効果が期待できる。しかし、ほとんどの空間音響手法がその性質上、特定の人のみが聞こえるパーソナライズした音を提供するものか、狭い範囲でのみ空間音響を聞くことができるものである。バイノーラルでは擬似的に音の共有を実現できるが、ヘッドホンをしていることから、コミュニケーションが取りにくい、認識の齟齬が生じやすいという問題がある。

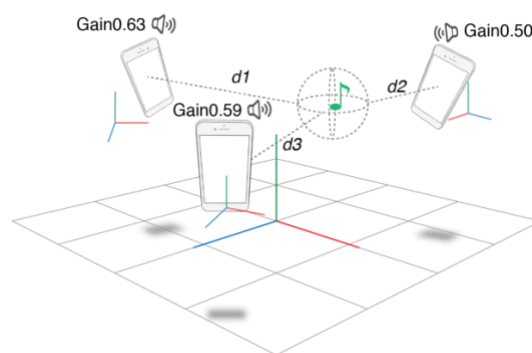


図 1 SymPhonicAudio: 各スマートフォンを各チャンネルに見立て、サラウンド音響を動的に構築する。

2. SymPhonicAudio の提案

そこで本研究では、「音の AR」による体験の共有を目的とし、複数のスマートフォンを同期し現実空間への音像定位を動的に行う手法 SymPhonicAudio を提案する。SymPhonicAudio では、時刻同期と位置キャリブレーションを行ったスマートフォンに DBAP (Distance Based Amplitude Panning) [10]を適用し、各スマートフォンを各チャンネルに見立て、サラウンド音響を動的に構築することで「音の AR」を実現する (図 1)。

SymPhonicAudio の特徴は次の 3 つである。

- **体験の共有:** その場にいる人たちは同じ場所から同じ音が鳴っているのを聞くことができる。
- **多人数に対応:** 多数のスマートフォンで同時に参加できる。
- **手軽さ:** ARKit または ARCore に対応したスマートフォンが複数あれば利用できる。

^{†1} 明治大学総合数理学部先端メディアサイエンス学科
Department of Frontier Media Science, School of Interdisciplinary
Mathematical Sciences, Meiji University

そして SymPhonicAudio は次の 3 点を満たす手法である。

1. スマートフォンで「音の AR」を実現する。
2. ヘッドホンを使わずに「音の AR」を実現する。
3. 複数人で同時に聞ける「音の AR」である。

SymPhonicAudio は、複数人が持つスマートフォンを使い、その場にいる人たちに「音の AR」を提供する。ヘッドホンなしに、その周囲の人たちは同じ場所から同じ音が鳴っているように聞こえるため、「音の AR」による体験を周囲の人たちと自然に共有しやすい。スマートフォンの画面を使った視覚による AR と組み合わせることが容易であり、音だけでなく、視覚も合わせた AR 空間を共有することができる。その場合、画面外の“動き”を「音の AR」で補うことで、画面で追って見ていなければ、何が起きているのかわからないという状況を防ぐことができる。

仕組みは各スマートフォンのスピーカを各チャンネルに見立てたサラウンド方式である。繋がっているすべてのスマートフォンのスピーカから同時に音源を鳴らし、その音量バランスによって音像定位を行う。スマートフォン自体が音響設備の要素となるため、そのほかの音響設備を一切要求しない。そのためネットワークに繋がったスマートフォンが複数あれば、様々な場所で使うことができる。

2.1 関連研究

サラウンド方式には、チャンネルベースとオブジェクトベースの 2 つがある。「音の AR」ではインタラクティブなシステムを必要とするため、オブジェクトベースになる。

オブジェクトベースの場合、その基本は Pukki ら (1997) の VBAP (Vector Based Amplitude Panning) [9], もしくは Lossius ら (2009) の DBAP (Distance Based Amplitude Panning) [10] のアルゴリズムに従う。

VBAP では聴取者を囲む同一球面上にスピーカを配置する。スイートスポットは球の中心に固定される (図 2)。

DBAP は VBAP よりも簡易なアルゴリズムであり、スピーカの位置と仮想音源の位置のみで計算する (図 3)。DBAP ではスピーカを配置する位置に制限はなく、不規則に配置しても問題がない。DBAP を用いた例として空間音響をスマートテーブル上で行う研究がある [12][13]。

DBAP は VBAP よりも音像定位の精度が低い。しかし Kostadinov ら (2010) は、VBAP は聴取位置を中心の狭い範囲に仮定するという制限があり、計算負荷・測位精度・柔軟性・シンプルさの点で DBAP が VBAP を上回ることを示している [14]。

DBAP によって、位置が既知である複数のスマートフォンのスピーカを用いて空間音響を作ることが可能である。井納ら (2016) は音響処理によって複数のスマートフォンの同期と位置を推定する手法の提案し、複数のスマートフォンのスピーカを用いて空間音響を作ることが可能であることを示している [15]。DBAP はスピーカ位置が動的だと

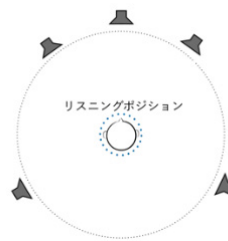


図 2 VBAP

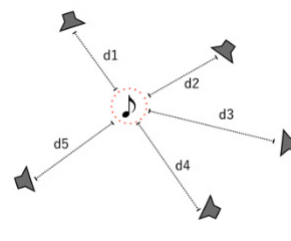


図 3 DBAP

しても、その位置がトラッキングできれば適用可能である。つまり、スマートフォンの AR によるポジショントラッキングを用いて、スマートフォンでの「音の AR」をヘッドホンなしに動的に実現できる。しかし、複数のスマートフォンが動き回る中で動的に DBAP を適用した例はなく、また AR 体験の共有という HCI の観点で DBAP を用いた「音の AR」の効果について調べた研究はなされていない。

2.2 想定する利用シーン

SymPhonicAudio は多人数での体験共有を目的としており、多人数で遊ぶゲームに使うことができる。特にイベント会場や宴会場などで行われるパーティゲームに使えば、一体感を演出でき、皆で楽しめる。このような場面で、バイノーラル再生で、ヘッドホンをつけながらの対面コミュニケーションは違和感がある。しかし SymPhonicAudio ではヘッドホンを使わずに「音の AR」での演出ができ、スマートフォンに囲まれた空間なら、3 次元のかつ動きのある仮想音源を再生できる。使うのはスマートフォンのみであり、音響機材が一切ない会場であっても実現可能である。

SymPhonicAudio は場所の制限が特にないため、屋外でも使える。例えばスマートフォンをラケットに見立てて、振ると「音の AR」でボールが打てるスポーツゲームで使うことができる。

そのほか、エンターテインメント施設で来場者のスマートフォンを使って、音の演出を行うという利用方法がある。

3. SymPhonicAudio の実装

実装はすべて WebAR[16]上で、Web 技術のみを用いて実装している。WebAR とは、Google 社が実験的に開発と提供をしている開発環境であり、ブラウザを通して ARKit や ARCore をクロスプラットフォームで利用できるフレームワークを想定している。サーバサイドは Node.js で実装し、クライアントサイド（スマートフォン側）は JavaScript で実装した。オーディオの再生には WebAudioAPI を用いる。サーバとクライアント間の通信はすべて WebSocket 通信で行う。

3.1 音像定位

SymPhonicAudio は DBAP[10]による振幅パンニングによって音像定位を行う。DBAP は各スピーカ位置と仮想音源位置の 2 つの関係のみで音量を決定する。DBAP で決定し



図 5 実験の様子

されるため、参加者には現実空間の同じ位置から音が鳴っているように聞こえ、「あそこにゴーストがいる」という体験の共有ができる。

GhostHunting は WebAR[16] 上で動作する、本手法 SymPhonicAudio による音の AR 機能を実装した。またバイノーラル再生には Resonance Audio[18] を用いた。2 つの再生方式は切り替えることができる。

4.2 実験内容

参加者は GhostHunting を SymPhonicAudio (スピーカ) とバイノーラル(イヤホン) の 2 つの手法の場合を体験した。参加者は正常な聴力を持つ 20 人の大学生 (男性 15 人、女性 5 人、19～23 歳) であり、本研究と関わりのない人たちである。実験は 4 人 1 グループで行い、参加者内計画で行った。体験の順番はランダムとした。実験は静かな教室で行い、実験中は参加者 4 人と実験者だけが教室内にいた。教室には机で囲まれた障害物のない 4m 四方のプレイエリアを用意した (図 5)。実験で使用したスマートフォンは iPhone8 (2 台)、iPhone8Plus、iPhoneX の 4 台である。イヤホンはそれらのスマートフォンに付属している標準のイヤホンを使用した。通信はすべてローカルネットワーク内で行った。

1 ゲーム 3 分間に設定し、その間ゴーストは 4m 四方のプレイエリア内を常に動き回る。ゴーストは水平方向のみに動き、垂直方向には動かない。ゴーストの動きによって差が出ないようにするために、あらかじめ用意した動きデータを用いた。バイノーラル手法の方では点対称に動きを変換したデータを使った。

本実験では体験の共有への影響を調べることを目的としているため、参加者は協力してグループの点数を稼ぐことを目標とし、声を出したり指を指したりして教えあっている。また実験中はプレイエリア内を自由に移動していい。

スマートフォンに方向を入力できる UI を設け、参加者はゴーストがいると思う方向を入力した。参加者が正しい方向 (30 度以内) を指している間、点数を加算するようにした。SymPhonicAudio のときは音量レベルを同じに揃え、バイノーラルでは各個人で好みの音量に設定した。参加者は事前練習で、音の動き方と入力方法の確認を行った。

4.3 結果

各参加者の点数 (ゴーストのいる方向を 30 度以内で正しく入力できていたとき加算する) と、7 点を最高点にし

たリッカート尺度で答えたアンケートの結果を示す (M: Mean, SD: Standard Division)。

各参加者の点数の結果はバイノーラル (M: 1730, SD: 1416)、本手法 (M: 2059, SD: 754) となった (図 6)。本実験は体験の共有という観点での効果を目的に設定し、他の参加者とゴーストの位置を教え合うことや、場所の移動をしてよいと教示したため、直接に音の方向性の認識精度を表すものではないことに注意する必要がある。

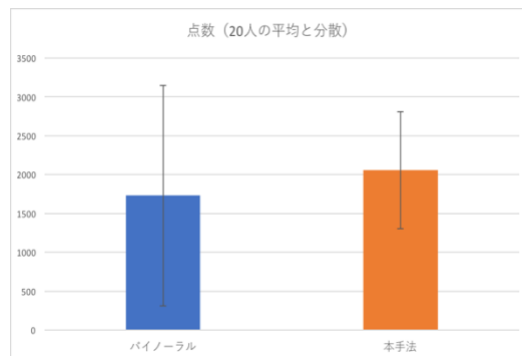


図 6 点数の結果

音についての質問は Q1～Q5 である。結果を図 7 に示す。5% の有意水準で有意差があったのは Q1 である。

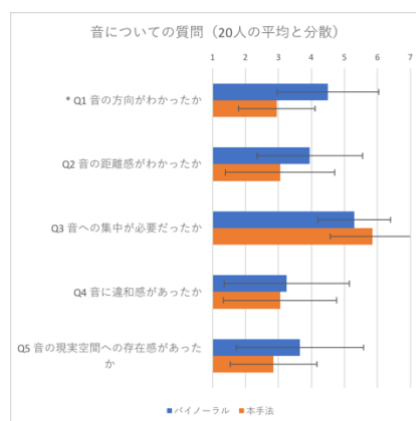


図 7 回答結果 (音について)

ゲームについての質問は Q6～Q9 である。結果を図 8 に示す。5% の有意水準で有意差があったのは Q7・Q9 である。

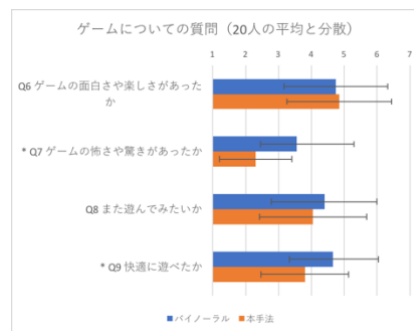


図 8 回答結果 (ゲームについて)

体験の共有についての質問は Q10～Q15 である。結果を図 9 に示す。5%の有意水準で有意差があったのは Q11・Q12・Q13 である。

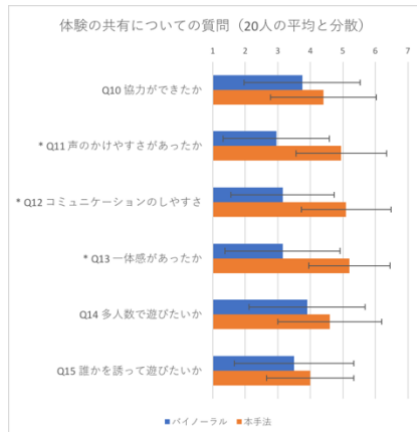


図 9 回答結果（体験の共有について）

4.4 考察

ゲームの面白さ・楽しさ (Q6) やまた遊んでみたいか (Q8) という項目がどちらの手法の場合でも得点が高いことから、まずゲームとして体験の設計がうまく機能していたことが確認できた。

他の参加者への声のかけやすさ (Q11)、コミュニケーションのとりやすさ (Q12)、他の参加者との一体感 (Q13) において、本手法がバイノーラルより優れていた。イヤホンのあるなしが、複数人で遊ぶゲームにおいて体験に影響すると示唆される。実験の様子の観察からも本手法でゲームをしていたときの方が、参加者同士で声をかけ合っていた。

バイノーラルの方が本手法より音が鳴っている方向がわかる結果 (Q1) になった。しかし、点数に有意差はなく、わずかではあるが本手法の方が高い結果になった。この結果は本手法では、コミュニケーションのとりやすさによって、ゴーストの位置がわかる人の「ここ (そこ) にいるよ」という言葉がすぐに伝わるからだと考察する。

バイノーラルの方がゲームの怖さや驚きがあったか (Q7) という項目で有意差があった。実験の様子の観察から、バイノーラルではゴーストが近くに来了時にびっくりする参加者が多く見られた。バイノーラルでは耳元で音になるため、臨場感があったと示唆される。

快適に遊べたか (Q9) という項目について、バイノーラルの方が優れていた。音の方向がわかりやすいことや、イヤホンで聞くので他の参加者の声に遮られずに音をしっかり聞くことができることなどが影響したと思われる。

音の方向が認知できるかについては全体的に個人差が大きい。どちらかの手法に限らず、全く方向がわからなかったという参加者がいた。同じグループ内でもバイノーラルの方がわかりやすかったという人もいれば、本手法の方がわかりやすかったという人もいた。

5. 議論・考察

本研究を通じて得られた結果から、音の AR を体験の共有という方向性で設計することには意義があるといえる。例えば実験用ゲームのような多人数で一緒に遊ぶゲームでは、本手法による音の AR を使った方が体験の設計として良い。一方、少人数や 1 人で遊ぶ場合はヘッドホンによるバイノーラルの方が、音の方向がわかり、臨場感もあるため適している。そもそも本手法は複数のスマートフォンがあることを前提としており、1 つまたは 2 つのスマートフォンだけではほとんど意味をなさない。ゲームやアプリ設計者は、体験の設計意図に基づいて、音の AR 手法を使い分けるのが良いだろう。

SymPhonicAudio の音像定位可能な領域はスピーカ (スマートフォン) の凸包内に限る。そのため、ユーザがスマートフォンを持つという想定では、ユーザの動き方によって音像定位可能な領域が変動する。人数 (スマートフォン) が多ければ、自然と人が空間に散らばり、音像定位可能な領域は大きく確保できる。ただ、アプリケーションによっては、ユーザが 1 箇所に集まってしまう、狭い範囲でしか音像定位ができないということが起こり得る。したがって、本手法を使ったアプリケーションを設計するときには、ユーザの動き方も考慮した設計を求める。

本手法は位置がわかっている静的なスピーカを組み込むことが理論上可能である。スマートフォンだけでなく、静的なスピーカを追加して、音像定位可能な領域を広げるということもできる。例えば、エンターテインメント施設において、来場者のスマートフォンを使い、本手法による「音の AR」を構築するときに、施設内の静的なスピーカも使うことで音像定位可能な領域を安定して広く確保することができる。

他の手法と比べて本手法の特徴は、各ユーザがスピーカ (スマートフォン) を手に持っているということである。そのため、AR の演出において、音 (もしくは音を伴ったオブジェクト) が別の場所から自分に対して向かってくるという演出が効果的にできる。またスマートフォンのスピーカの振動が手元に伝わることも相まって、音による体験に自分も参加しているという感覚を強めていると考察ができる。ヘッドホンによるバイノーラルでも自分に対して向かってくる音の表現が可能である。しかし、本手法ではヘッドホンによって分離された体験とは違って、一体化した体験に自分も加わるという部分に他の手法にはない面白さがあるといえる。

本手法は音像定位の精度に課題が残る。ずれを生み出す要因を挙げると、時刻同期のずれ・通信処理の遅れ・再生処理の遅れ・スピーカの音量レベルの差・スマートフォンの角度による音量の差・位置キャリブレーションのずれ・ポジショントラッキングのずれと数多くある。それぞれの

要素のずれがわずかでも、積み重なって大きなずれとなっている可能性がある。今後、1つ1つの精度の検証を行い、問題があればその対策を立てていく必要がある。

本手法は iPhone8 以上なら安定して動作することが確認できたが、それよりも前の世代の iPhone だと安定性がなかった。iPhone6 や iPhoneSE でも ARKit は対応している。しかし AR 機能を使うと、処理が追いついていない様子があり、フレームレートが極端に低くなる・ポジショントラッキングが追いつかないということがあった。そのような状況で同期再生を行ったとしても、再生処理のもたつきでタイミングがずれてしまう。事前実験によって、そのような端末が含まれている状態で本手法を使うと、音がずれて二重に聞こえ、体験が破綻してしまうことがわかった。

ARCore を使った Android 端末については本手法が動作することは確認できたが、OS の違いや音質の違いによる精度への影響については未確認である。

本手法で同時に適用可能なスマートフォンの台数上限については調べられていない。DBAP のアルゴリズム自体はスピーカのチャンネル数に上限はない。ただ、台数が多くなってくると、計算や通信処理が重くなることや、音像定位がばやけてしまうことが起こる懸念がある。台数の違いが体験にどう影響を与えるのかは今後調査が必要である。

6. おわりに

ARKit や ARCore のリリースにより、スマートフォン内蔵のカメラとモーションセンサのみで AR 体験をマーカレスに実現できる環境が普及した。スマートフォンによる AR は拡張空間を他人と共有し、一体感やコミュニケーションを楽しむという方向性に利点がある。しかし、「音の AR」を実現するための空間音響手法のほとんどがその性質上、パーソナライズした音もしくは狭い範囲での提供であり、主流であるバイノーラル方式ではヘッドホンをしていることから、コミュニケーションの取りにくさや認識の齟齬が生じやすいという問題がある。

本研究では、「音の AR」による体験の共有を目的とし、複数のスマートフォンを同期し現実空間への音像定位を動的に行う手法 SymPhonicAudio を提案し、その実装を行った。そして、本手法とバイノーラルを比較する評価実験を行った。その結果、多人数で協力して遊ぶ「音の AR」を用いたゲームにおいて、音の方向性のわかりやすさはバイノーラルの方が優れているが、本手法の方が、他の参加者への声のかけやすさやコミュニケーションの取りやすさ、ゲームを通じての一体感があった。

参考文献

- [1] Apple, ARKit. <https://developer.apple.com/arkit/> (参照 2017-11-29)
- [2] Google, ARCore. <https://developers.google.com/ar/> (参照 2017-11-29)
- [3] ARtillery, How Big Will ARkit Really Be? (Unit Forecast). <https://artillery.co/2017/08/01/how-big-will-arkit-really-be-unit-forecast/> (参照 2017-11-29)
- [4] Lam, Jonathan, et al. "Sound localization on a horizontal surface: virtual and real sound source localization." *Virtual Reality* 19.3-4 (2015): 213-222.
- [5] Møller, Henrik. "Fundamentals of binaural technology." *Applied acoustics* 36.3-4 (1992): 171-218.
- [6] Algazi, V. Ralph, et al. "The cipic hrtf database." *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*. IEEE, 2001.
- [7] Nakagaki, Ken, and Yasuaki Kakehi. "SonalShooter: a spatial augmented reality system using handheld directional speaker with camera." *ACM SIGGRAPH 2011 Posters*. ACM, 2011.
- [8] Gerzon, Michael A. "Ambisonics in multichannel broadcasting and video." *Journal of the Audio Engineering Society* 33.11 (1985): 859-871.
- [9] Pulkki, Ville. "Virtual sound source positioning using vector base amplitude panning." *Journal of the audio engineering society* 45.6 (1997): 456-466.
- [10] Lossius, Trond, Pascal Baltazar, and Théo de la Hogue. "DBAP—distance-based amplitude panning." *ICMC*. 2009.
- [11] Ochiai, Yoichi, Takayuki Hoshi, and Ippei Suzuki. "Holographic Whisper: Rendering Audible Sound Spots in Three-dimensional Space by Focusing Ultrasonic Waves." *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017.
- [12] Collins, Karen, et al. "Audio air hockey: A pilot study in using audio-based games for the measurement of loudspeaker placement preferences for smart tables." *Audio Engineering Society Conference: 41st International Conference: Audio for Games*. Audio Engineering Society, 2011.
- [13] Kapralos, Bill, et al. "Sound localization on tabletop computers: A comparison of two amplitude panning methods." *Computers in Entertainment (CIE)* 12.2 (2014): 4.
- [14] Kostadinov, Dimitar, Joshua D. Reiss, and Valeri Mladenov. "Evaluation of distance based amplitude panning for spatial audio." *ICASSP*. 2010.
- [15] 伊納洋佑, 石川佑樹, and 中祐介. "複数の携帯端末の同期・位置推定手法による閉鎖空間の音響環境構築" 電子情報通信学会応用音響研究会 115.424 (2016): 19-26.
- [16] Google, WebARonARKit. <https://github.com/google-ar/WebARonARKit> (参照 2017-11-30)
- [17] Haas, Helmut. "The influence of a single echo on the audibility of speech." *Journal of the Audio Engineering Society* 20.2 (1972): 146-159.
- [18] Google, Resonance Audio. <https://developers.google.com/resonance-audio/> (参照 2017-12-12)