デジタルツインの魅力を引き出す音声デザイン

日本音響学会 2024 春 ビギナーズセミナー 「研究がプロダクトになるまで」

2024 年 3 月 5 日

株式会社サイバーエージェント AI Lab　吉本 暁文

CyberAgent AI Lab

Contents

1 自己紹介・事業部紹介・作品紹介

2 どのように始まったか

3 広告音声作りの旅

4 広告音声作りを振り返って

5 お台場スタジオや現在のチーム

自己紹介

吉本 暁文

趣味：気が向いたものを作ること、ヴァイオリン

経歴：　2007〜2011年　芝浦工業大学

2012〜2017年　奈良先端科学技術大学院大学

2017年〜現在　株式会社サイバーエージェント

AI Lab Audio チーム

@mulgray

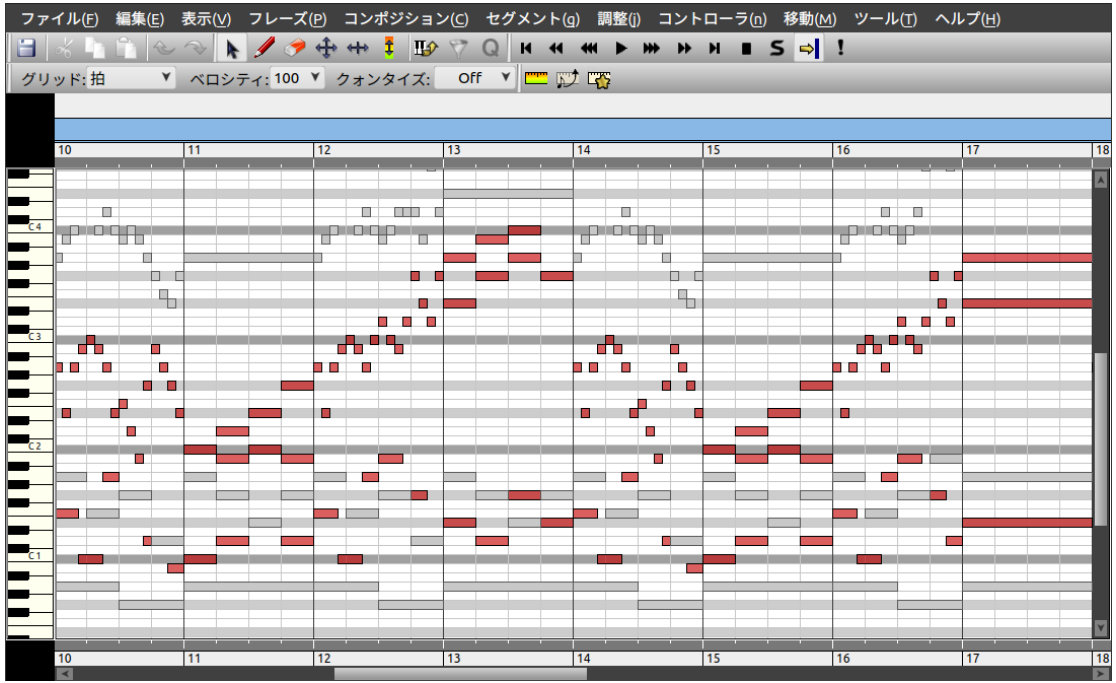
Developer Experts

AI Lab

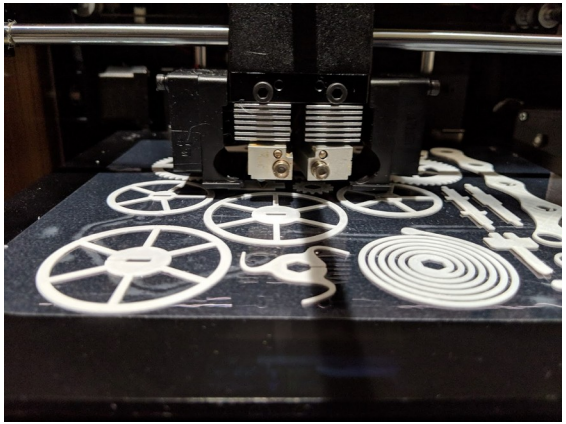
CyberAgent AI Lab

趣味補足

気が向いた時は作曲っぽいことをしていることもあります（MIDI 手打ちするタイプ）



他にも3Dプリント等色々



事業部紹介

メディア事業

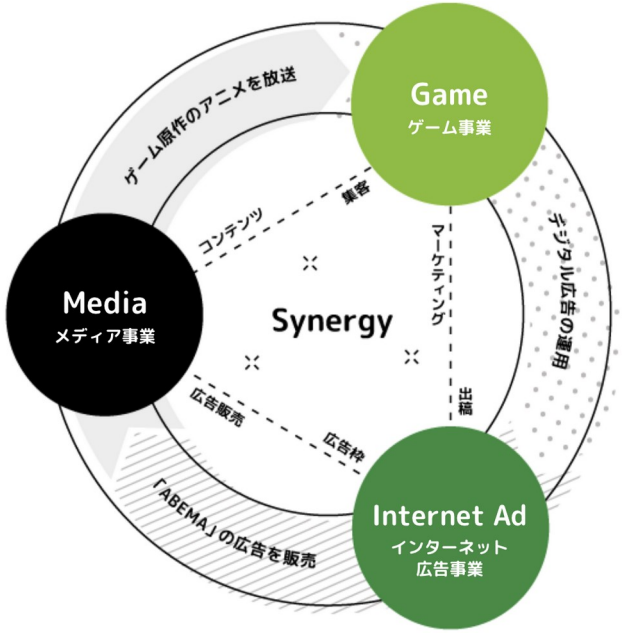
多彩なコンテンツを提供する新しい未来のテレビ「ABEMA」、国内最大級のブログサービス、マッチングアプリなど、インターネット産業の変化にあわせ、多くの方々にお楽しみいただけるサービスを提供しています。

インターネット広告事業

1998年の創業以来インターネット広告事業を展開しており、国内トップシェアを誇ります。 広告効果を最大化する運用力およびクリエイティブ力を強みに、AIを活用したアドテクノロジーなど総合的なソリューションを提供します。

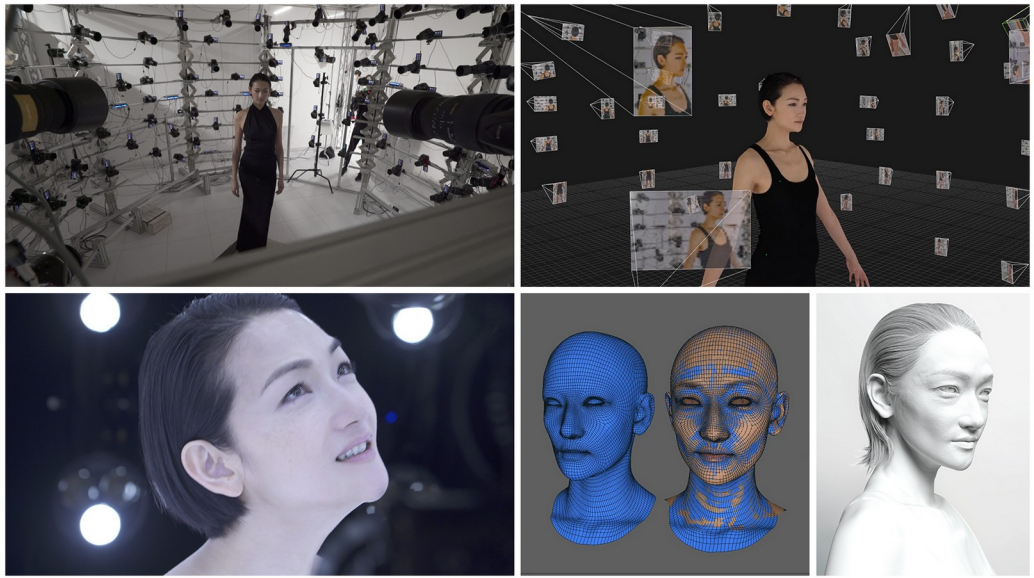
ゲーム事業

ゲーム・エンターテイメント事業を主軸とし、サイバーエージェントのゲーム事業に携わる12の子会社が所属。
バンドリ！ ガールズバンドパーティ！、NieR Re[in]carnationなど、多数のゲームをリリースしています。



作品紹介：冨永愛さんデジタルツイン

デジタルツインレーベル： CG や TTS/VC により人間を仮想化し、キャスティング
Cyber AI Productions と AI Lab で提供中

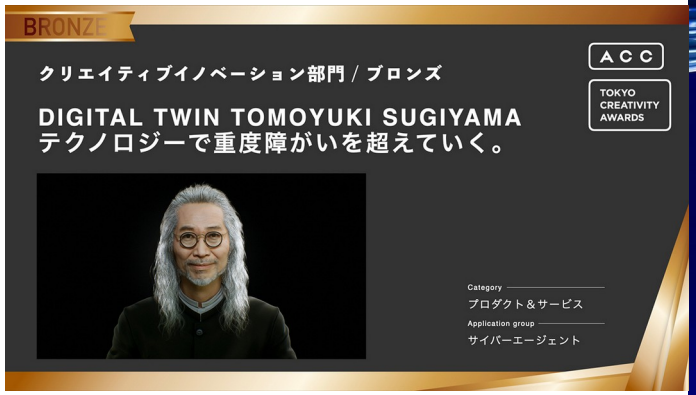


CYBER AI PRODUCTIONS

AI Lab

作品紹介：デジタルハリウッド杉山知之学長デジタルツイン

ALS で発話が困難だが、本人の魅力と表現力をできるだけ再現し、講演を実現



クリエイティブ業界で栄誉ある
ACC ブロンズと WIRED JAPAN 賞を受賞

どのように始まったか

背景

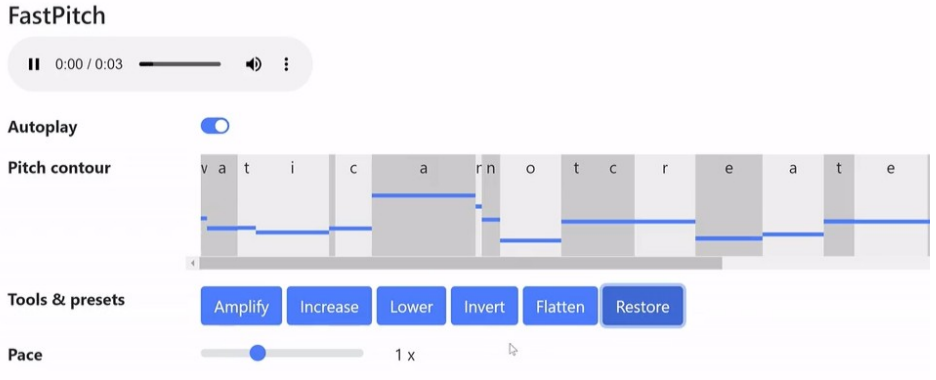
- 元々は自然言語処理 (NLP) が専門
- Web ページのテキストを分析して広告を出し分ける製品を担当していた
- 趣味で CG に興味があり、社内の CG 部署の動きもなんとなくチェックしていた

きっかけ

- ある日、CG 部署で音声をどこから持ってくるか議論した議事録を発見
  議事録には「かなり音に興味あればいける？（意訳）」
- 外から割り込んで「かなり音に興味あります」と書き込んだ
- 上司から「やってみなよ！（意訳）」と言ってもらえたので音声部署を作った

広告音声作りの旅

- この時、老舗の音声合成製品はアナウンスに向いているものが多かった
  - 感情表現などはあるものの、単調で飽きやすく、広告で使えるシーンは限られる
- 研究分野では、より自由な音声が出せるオープンな実装が出現していた
- 試しに Tacotron 等で TTS を作ってみたり、CycleGAN で VC を作ってみたり



その過程で、韻律が重要だった

VC は変換元話者の互換性に課題

当時 FastPitch を見て、これが UX 的には一つの解だと思った

今は TTS/VC は混ざりつつあるが...

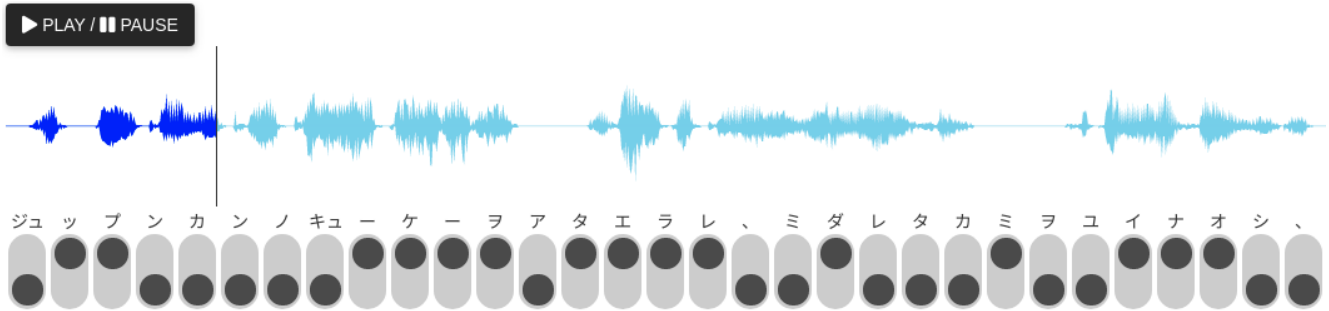
少しずつ音声系のお作法や ESPnet に慣れていく

- **オープン実装のいくつかはシンプルだったが、ボコーダと合わせると微妙な場合も**
  - normalization の計算法が異なっていて、合わせればそこそこの品質ではあったが…
- **ESPnet + ParallelWaveGAN で FastSpeech2 学習手順を説明通り試すと…**
  - **非常に音が良かった！（大事）**

- **duration や pitch の条件付けを変更すると狙い通りの発話に近づいた**
  - しかし Tacotron 教師だと duration の条件付けに忠実にならないことに気づく
- **duration 教師を Julius で与えたり DNN-HMM, HSMM で与えたりし始める**
  - **かなり忠実になった！**
- **そうこうしているうちに ESPnet がよくわかるようになり手放せなくなった**

実は制御だけを考えても一筋縄ではいかないことを知る

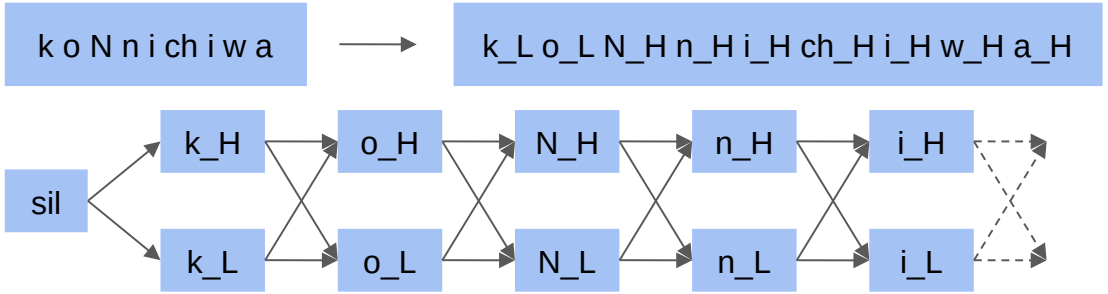
- **pitch と duration で制御できるのはいいが、結構難しい**
  - 複数人で作業すると必ず好みが分かれるので、基本指示はアクセントくらい荒い方が良い
- **アクセント条件付けを推論時に変えられるようにしても、効きが中途半端**
  - LH 式にして、学習時に正確なラベルで学習できるようにアノテーションし改善を狙う



  - ツールをスマホ対応して、電車でも風呂でも就寝前でもひたすらアノテーション

アノテーションはスケールしにくい & 技術が属人化してしまうので自動化へ

- 苦労の甲斐あってアクセント指示がかなり効くようになった
- 必要なデータがわかってきたので工数削減へ
  - LH 式アクセントラベルの自動化
    - ESPnet の音声認識を各音素の高低を選ぶビームサーチに改変する手法を提案（音響学会 2022 春）

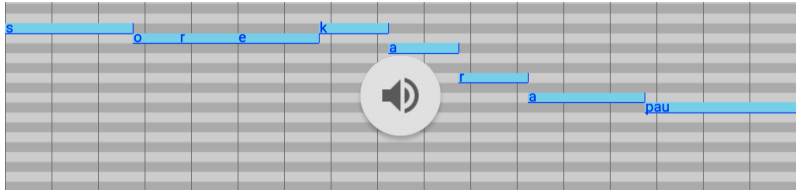


※ 文脈に過学習しがちなので
今はより音を重視する手法へ

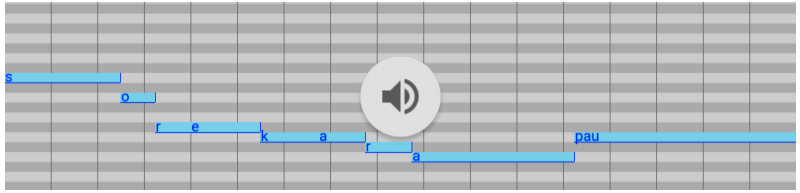
  - カナ読点認識も作り、台本修正とアクセントラベルを自動付与後に修正するフローへ
    - 修正しなくても動く

自然性はデータで大差がついてしまうので色々試行錯誤していた

- 最初に冨永愛さんの音声合成を試作した段階ではあまり「リアル」ではなかった

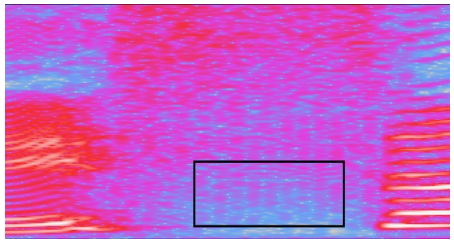


- 社内外の大勢に自然な文章を音素バランスも考慮して集めるアルゴリズムで作った台本を読んでもらい、そこから finetuning
  - 予測ピッチや表現の自然性が向上した

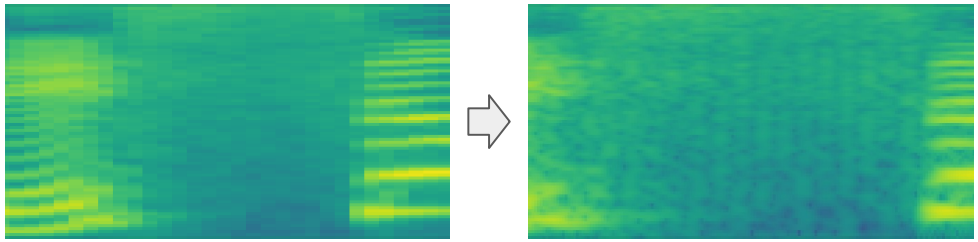


音質も適当ではいけなかった

- **ある日、プロダクト側から「す」の音がノイジーだと相談を受ける**
  - 感想１：無声音だしそもそもがノイズみたいなものだからな…
  - 感想２：ボコーダの upsample が原因か…？
- **Audacity でノイズを感じる箇所を眺めていると、確かに規則性が認められた**
  - mel loss で作っている spectrogram を見てみると、この規則性が見えなかった
  - hop size など調整すると同じ模様が見え始め、これで学習するとノイズが消えた



Audacity



元の mel は低解像度だった　　修正後

広告音声作りを振り返って

- **一旦、自分の感性を信じて作った**
  - 「ここはもうちょっと高らかに」みたいな要望がよくある → 音楽的デザインへ
  - デザインに重要だと感じたので反応速度にもこだわった
- **最初にリリースするものは一日でも早く出した**
  - それが生きている間に次のぎりぎり間に合いそうなインパクト最大の一手を考え続ける
  - 最初から真の理想を狙うと失速してしまう
- **段々に次に必要そうなこと・考え方がまた見えてくる**
  - 学習の工数削減と生成の工数削減
  - 大規模事前学習や周辺技術

デジタルツイン事業の直近の展望

- 専門的な撮影・収録ができる「極AIお台場スタジオ」が去年オープン
- 24 時間収録可能・複数名同時収録など想定
- まずは広告表現から、既に 1 万発話以上を収録

CyberAgent AI Lab

2020

2021

最初の音声プロジェクト
Audio チームを開始
戸田先生と連携開始

目的：デジタルツインのCGに音声をつける

声質変換・音声合成・認識の既存手法の再現を始める

AI SHIFT

AI Messenger Voicebot

AI Shift と連携開始
高道先生と連携開始
Audio チームが 2 名に

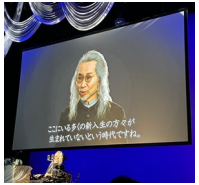
AI Shift の手がける自動電話の競争力向上のため、音声認識と音声合成を特化

2022

CA が音響学会賛助会員に
アクセント認識で研究発表
デジタルツイン冨永愛　リリース
郡山先生が入社される
Audio チームが 5 名に

ナレーションの範囲を超えるような、表現力豊かな声を音声合成で一旦形に

2023



ICASSP 論文採択
デザイナー向け音声合成
複数のチームや事業部と連携
Audio チームが 10 名に

音声強調や聴覚など、チームの領域が拡大

CyberAgent AI Lab

共同研究・採用・インターンシップなど
今後ともよろしくお願いいたします！