合成音声と生音声の自然なクロスフェード手法および エンターテインメントへの応用

伏見 遼平[1,a)]

概要：話者性を消す効果のあるフォルマントフィルタを用いた人間の音声と合成音声の自然なクロスフェード手法を提案する. この手法は人間と人工知能が協調するハイブリッドな対話エージェントの実現や, これを用いた「話している相手が人間なのか機械なのかわからない」という体験を中心としたエンターテイメントにも応用できる. 制作したインスタレーション作品「ある声について」の展示の様子や感想を中心に紹介する.

Crossfade between TTS voice and raw voice using a formant filter

Ryohei Fushimi[1,a)]

Abstract: A novel method for natural crossfading between human voice and voice generated by text-to-speech systems is presented. Authors also played performance art "About a voice" in an exhibition and participants' impressions were collected.

1. はじめに

様々な SF 作品の中では, 高度な音声合成や音声認識技術を持った自動対話エージェントが人間社会の中に溶け込み, 人間側も他の人間と同じようにエージェントと対話しているような描写が多く見られる.

音声合成 (TTS: Text-To-Speech) の技術は近年大きく進歩し, 近年で自動対話エージェントに搭載することで一般ユーザもエージェントとの対話を行うこともできるようになったが, 依然として, エージェントとの対話は, 人間相手の対話と同じようになされるわけではない. 人間と機械の協調を行うことができるエージェントシステムは多く提案されているが, テキストを用いたものが中心である.

本稿では, 人間と機械が協調したハイブリッドな音声対話エージェントシステムの実現のため, 人間が話す生の音声と, 事前に録音した音声, TTS を用いて合成した音声を自然に合成する手法を提案し, この手法を用いて制作したパフォーマンス作品の構成や体験者の感想についてまとめる.

生音声・録音音声・合成音声をつなぎ合わせる方法としてクロスフェード, ボイスモーフィングなどが知られているが, 合成音声特有の声質の不自然さを解決するアプローチはなされておらず, 結果としてつなぎあわせた部分は容易に判別できる.

本稿で提案する手法は, 声の話者性 (音声に含まれる情報のうち話者特有の成分) や録音環境の違いなど, 音声が人間であるか機械であるかを判別する手がかりとなる成分が消えるようなフィルタを, それぞれの音声に一律に適用することにより, つなぎ目を目立たせなくするというものである. 今回制作した作品では, フィルタとしてフォルマントフィルタ (ヴォコーダー) を用いた.

本稿で提案した手法は人間と機械の協調を進めていく際に応用できると考えている. 具体的には, コールセンターなどの音声コミュニケーションを行う現場で現在人間が行う業務に機械の力を借りて省力化を行ったり, 逆に Siri や Cortana などの人工知能を用いた音声対話エージェントシステムにおいて機械では繊細な対応できない部分について

[1] 東京大学 学際情報学府
The University of Tokyo, 7-3-1 Bunkyo, Tokyo 103–0032, Japan
[a)] fushimi@cyber.t.u-tokyo.ac.jp

図 1 1 枚目のグラフィック. 東京藝術大学油絵科に所属する協力者が『夢』というテーマを与えられ, 自由にモチーフを挙げて平面構成を行い制作したもの.

図 2 2 枚目のグラフィック. フランク・ステラのコンポジションをトレースした上に, コンピュータでランダムに生成した配色の指示通りに筆者らが塗り直した.

人間によるコミュニケーションを行ったりなどが応用先として考えられる.

2. 提案手法

本稿で提案する手法の発想は, 「TTS など音声合成技術や, それらのモーフィング技術に限界があり, 人間の音声と合成音声の合成が難しいのであれば, 人間の側を機械の合成音声に合わせてクオリティを下げてあげることで, 自然に合成を行うことができるのではないか」というものである. 本手法では, 声の話者性 (音声に含まれる情報のうち話者特有の成分) や録音環境の違いなど音声が人間であるか機械であるかを判別する手がかりとなる成分が消えるようなフィルタを, それぞれの音声に一律に適用することにより, つなぎ目を目立たせなくする. 今回制作した作品では, フィルタとしてフォルマントフィルタ (ヴォコーダー) を用いた.

3. パフォーマンス作品の構築・設計

提案するシステムを用いたパフォーマンス作品『ある声について』制作し, 第 17 回東京大学制作展 [1] にて発表を行った. この章ではその作品の構築・設計について記載する.

3.1 掲示したグラフィック作品

額装した A4 サイズの平面作品を 4 つ掲示した. それぞれ, 『人間が自由に描いた作品』『人間がコンピュータプログラムの指示に従って描いた作品』『人間が関与せず, プログラムの指示に従って描いた作品』『プログラムが人間が描いた作品を加工した作品』を用意した. これは音声ガイドの中で, それぞれの作品について「作品の作者は誰か?」という問いかけを行いながら解説を進めていくためである. それぞれの作品を図 1,2,3,4 に示した.

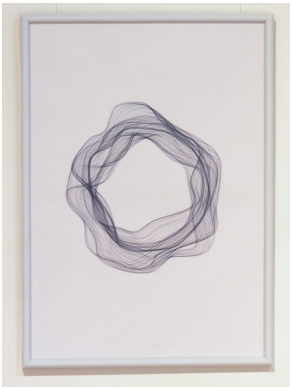

図 3 3 枚目のグラフィック.[2] に掲載されていた processing のサンプルコードをアレンジしたものを用いた.

図 4 4 枚目のグラフィック.Deep Learning による画像認識に用いるニューラルネットワークを用いたフィルタ Deep Dream[3], [4] を用いて, 協力者の書いた絵を加工したもの.

3.2 「音声ガイド」システム

作品のために, 音声ガイド風の端末と, 端末に音声を送信するコンピュータを用意した. 音声ガイド風の端末は, スマートフォン (Apple iPhone 5S) にネックストラップが付いたケースを装着したものを用いた. ケースは体験者に画面が見えないように, フタ付きのものを選んだ.

端末に音声を送信するコンピュータは, ガイドの進行に応じて後述するようにオペレータ (操作者) が実際に話した音声に加え, 事前に録音した音声, 合成した音声, リアル

タイムでタイピングした音声などをクロスフェードし, さらにフィルタを掛けて端末に送信することができる. また, スマートフォンのカメラに写っている映像を確認し, 音声ガイドに対応して

音声・映像の送出伝達には Skype を用いた. スマートフォンでは

3.3 体験の流れ

本作品『ある声について』は, 「人間と機械の境界について考える, 4 つのグラフィック作品とインタラクティブな音声ガイド」として展示され, Web サイトやパンフレット等にもそのような記載を行った. (参加者は, 実際の人と話す部分があるということにき気付くきっかけは与えられなかった.)

キャプションには音声ガイドのネックストラップを首にかけると, 自動的に音声ガイドが始まるように指示が書かれており, オペレーターはカメラを通じて最初の作品の前に立っていることが確認できたら体験を始めるようにした.

音声ガイドは 4 パートに分かれている.

(1) 人間のオペレーターが実際に話すパート (以降, 生音声パートと呼ぶ). 原稿通り話す部分と, オペレータからの問いかけに対して自由に返答を行う部分があった. このパートでは音声ガイドそのものの説明や図 1 の作品の解説を行う.
(2) オペレーターが, 事前に録音しておいたパート. 事前に決めた内容をそのまま再生したが, オペレータからの問いかけに対する返答が YES/NO どちらであるかに応じて返答の音声を変えた部分があった. このパートでは図 2 の作品の解説を行う.
(3) オペレータの音声に似た合成音声. Mac OSX に付属する say コマンドで合成した音声を用いた. このパートでは図 3, 4 の作品の解説を行う.
(4) 最後に人間のオペレーターが「ところでいま話している私は, 人間だと思いますか, 機械だと思いますか?」と問いかける. 返答によらず,「そうかもしれませんね」という事前録音された音声を再生する.

再生されるすべての音声には, Mac OSX の付属ソフト Garageband により, フォルマント・フィルタをかけて体験者にフィードバックされた.

音声ガイドのパートによる音声の変化は, 音声ガイドの対象の作品を作りだしたの主体が次第に人間とも機械とも言いがたいものとなることにも対応している. ただし, この相関に気付き事後アンケートで報告した参加者は少なかった.

4. 参加者の反応

終演後に任意の聞き取りアンケートを行った.

4.1 声について

(1) 声の変わり目には気付かなかった. (多くの参加者)
(2) 問いかけの中で, 参加者に考えさせる質問が多かったので, 聴かれた内容について考えているうちに声の主が変わっていた. 喋っている内容のアクセントが間違っていることがきっかけで気付いた.
(3) ぎこちない音声になったこともあったが, スマートフォンの形をした機器を渡されたので, 通信の不具合だと思った.
(4) 単なる偶然かもしれないが, 会話がうまく噛み合ったため, 最後まで全く機械であることに気付かなかった. アクセントのおかしさに関しても, 会話がスムーズれあれば気にならなかった.

生音声と録音音声の切り替わりにはほとんどの人が気づかず, 合成音声との切り替わりは解説を聞いている途中のアクセントの微妙なズレなどによって気付くことができたという参加者が多かった.

合成音声の不自然なアクセントに気付いたものの, 他の解釈を行うことで話し相手が人間であるという思い込みを維持していた参加者がいた (3,4). 例えば不安定な通信状況を再現するノイズ等を入れるなど, このことをより意識的に用いたシステムを用いれば, より気付かれにくくなると考えられる.

4.2 その他

(1) それぞれの作品の ”作者とは誰か” について考えた.
(2) 機械が作品の作者になることができるのかどうかについて考えた.
(3) 音声ガイドのコンテンツそのものがおもしろく, 声の仕掛けよりもそちらに集中して聞いていた.
(4) 人間は意外と話している相手が機械でも人間でも気付かないということがわかった.
(5) 脱出ゲームや参加型ゲームのような体験だった. アトラクションなどに使えそう.
(6) コールセンターなど非対面のコミュニケーションで使えそうだと思った.
(7) 仕組みを知った後は声に対する感覚が敏感になった. 電話相手の声とかが信じられなくなるかもしれないと思った.
(8) 今後, 自動音声の少しの不自然さにも敏感になってしまうかも知れないと思った.

作品体験として, 音声ガイドの内容だけでも楽しめるものを用意したため, 音声のしかけよりも内容が印象に残ったという参加者も多かった (1-3). また, 仕組みを話したあとは, 考えられる有用な応用先や, 体験による影響について話した参加者もいた (5-8).

最後に展示の光景を図 5 に示す.

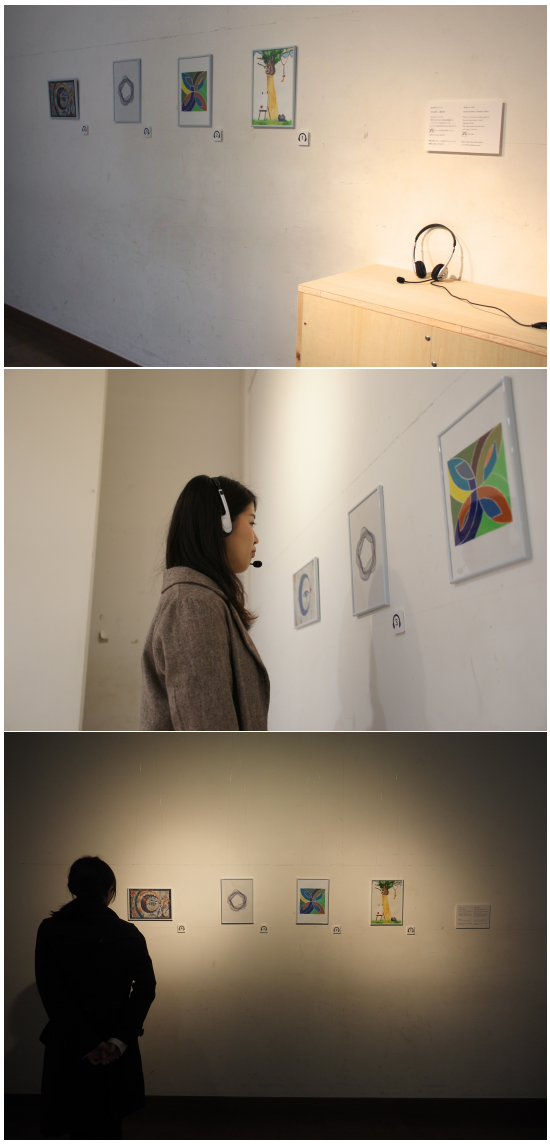

図 5 展示の様子

5. エンターテイメント以外の応用先

本稿で提案した手法は, 特に音声コミュニケーションの領域について, 人間と機械の協調を進めていくために用いることができる. ここではその協調の 2 つの方向性を示す.

1 つ目は, 現在は人間しか行うことができないと考えられているコミュニケーション領域に機械による支援を行うという方向性である. 例えば, コールセンターで行われている商品の使い方のサポート業務を例に上げる. 多くの利用者は, 定型的な Q&A で対応できる質問に答えていくことでサポートを受けることになるが, 例外的な状況が発生したり, 回答に責任が伴う場合 (将来の対応や返品等のサービスを約束する場合) には, 定形外のコミュニケーションが発生する.

このようなコミュニケーションについて, 本手法を用いたハイブリッドなエージェントシステムを用いれば, 定型的な対応については録音音声が, 準非定型な (自動生成したテキストの読み上げで対応できる範囲の) 対応については合成音声が, オペレータによる個別対応が必要な範囲の対応はオペレータの音声が担うことで, オペレータの負担を減らすことができる. 理想的には, オペレータが同時に複数のカスタマーとの対応業務を並行で担当することもできる.

当然ながらこのようなシステムは本手法を使わず生音声・録音音声・合成音声の切り替わりが明確に判別できる形でも実装できるが, 利用者の視点からすると話す相手がパートによって変わっているように感じ, ぎこちなく感じてしまうかもしれないのが問題である.

2 つ目は, 現在コンピュータによって実現している対話エージェントシステムに, 人間による支援を行うという方向性である. 例えば Apple の Siri などの対話エージェントや, 公共交通などで用いられている案内システムの対話エージェントなどが行う業務のうち人間が対応したほうがよりきめ細かな対応ができるような業務については, 人間が分担することもできる.

このように, 本手法を用いれば, 人間が得意としている判断や責任, 利用者に合わせた対応を必要とするコミュニケーションと, そうではないコミュニケーション, 場面に応じた人間と機械の協調がより容易になることが考えられる.

6. 謝辞

本研究及び作品『ある声について』共同制作者の三輪桃子さん, および東京藝術大学大学の伊東五津美さん, 東京大学の片山健さんをはじめとする作品の制作に協力いただいた協力者の方々に感謝します.

参考文献

[1] 東京大学大学院学際情報学府：第 17 回東京大学制作展アーカイブサイト.

[2] Bohnacker, H., Gross, B., Laub, J. and Lazzeroni, C.(eds.): *Generative Gestaltung: Entwerfen. Programmieren. Visualisieren*, Schmidt Hermann Verlag (2009).

[3] Generator, D. D.: Deep Dream Generator. `http://deepdreamgenerator.com/`.

[4] Mordvintsev, A., Olah, C. and Tyka, M.: Google Research Blog - Inceptionism: Going Deeper into Neural Networks (2015). `https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html`.