公開

# 研 究 終 了 報 告 書

## 「カクテルパーティ効果に着目したオンライン話者とオフライン話者の選択的聴取の支援」

研究期間: 2020 年 11 月～2023 年 3 月
研 究 者: 高木 健

### 1. 研究のねらい

遠隔地の人(オンライン話者)と会話をしつつも、時には自分の周りの人(オフライン話者)と会話をするというオンライン会議の形態ができつつある。これに伴い、ユーザはオンライン話者とオフライン話者の 音声の中から、興味のある内容だけを聞き取ることの重要性が増している。このためには、複数話者の音声の中から、ユーザが意識を向けた話者の音声を明瞭に聞き取りができること(選択的聴取)が必要である。

現状の音声インタフェースでは、特定の 1 つの音声をユーザにとってはっきり聞こえるようにすることはできる。しかし、このためにはユーザがどの音声に関心があるかについての情報を事前にインタフェースに与える必要があるため、聞き取る話者の切り替えがすぐにできない上、ユーザが意図しないタイミングで話しかけられたときには、そのことに気付くことができないという問題がある。

ここで、人間は複数話者の音声の中から、意識を向けた話者の音声を聞き取ることができるというカクテルパーティ効果に着目した。人間がカクテルパーティ効果を発揮しやすい状況は、①話者の音声の到来方向がずれている、②話者の音声が周波数軸上で被らない、③話者の間で音声の明瞭度が揃っているという状況であることが知られている。 したがって、本研究の目的は、オンライン話者とオフライン話者の音声に対してカクテルパーティ効果を最大限に発揮できるようにする音声変換システムを構築することである。

### 2. 研究成果

(1)概要

本研究ではカクテルパーティ効果に着目しその効果を享受しやすい音声処理を提案し、オンライン話者とオフライン話者の音声の選択的聴取をしやすくするシステムの実装を行った。

研究課題(i)研究課題(ii)では、人間がカクテルパーティ効果を発揮しやすくするための 3 条件を達成するための音声処理に取り組んだ。条件①オンライン話者とオフライン話者の音声の到来方向が被らないようにするため、(a) オフライン話者の空間的な音源位置を推定し、(b) オフライン話者とオンライン話者の知覚される音源位置について、空間的にどこに配置すれば選択的聴取性が上がるかについて明らかにした。条件②周波数軸上で音声が被らないようにするために、オンライン話者の音声に対する加工方法として、(a)ハイパスフィルタやロー

パスフィルタを音声にかける、(b)音声明瞭度を最大化するような畳み込みニューラルネットワーク、という手法に取り組んだ。条件③話者の間で音声の明瞭度が揃えることについては、条件②の信号処理の最適化の関数に話者間の明瞭度のバランスを取るようにすることで達成した。

研究課題(iii)では研究課題(i)(ii)の音声処理をユーザが実際に使用できる装置の試作を行った。研究課題(i)①のオンライン話者とオフライン話者の音声の到来方向が被らないようにするため、オフライン話者の音源位置を推定できる、装着可能なマイクアレイの製作を行った。

研究課題(iv)では、研究課題(i)を実施する中で、複数話者間のわずかな音量の差が聞き取りやすさに大きな影響を与えることがわかった。したがって、複数話者がいる中で、ユーザがより集中したい音声があるとき、そちらの音声に集中しやすくするようなインタフェースも必要となる。そこで、耳介の上で指をなぞった位置を連続的に推定することで、直観的に話者同士の音量調整を行えるインタフェースを提案した。

(2)詳細

**研究課題(i)オンライン話者音声をオフライン話者の音声から 3 次元空間中と周波数軸上で分離するフィルタの構築**

本研究課題では、人間がカクテルパーティ効果を発揮しやすくするための条件のうち、①話者間で音声の到来方向が被らない、②話者の音声が周波数軸上で被らないことを達成するようなフィルタを構築した。周波数軸上で音声が重複しないための音声変換に関しては様々な手法を実施し、本研究課題の主要な貢献となる。

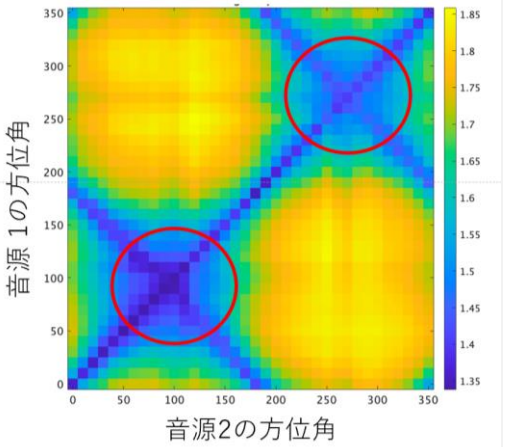


*図 1 音源 1 と音源 2 の明瞭度の和*

①オンライン話者とオフライン話者の音声の到来方向が被らないようにするため、(a) オフライン話者の空間的な音源位置を推定することと、(b) オフライン話者とオンライン話者の知覚される音源位置が重複しないようにする必要がある。

(a)リアルタイムでオフライン話者の位置を推定するために、必要な分離性能に応じてマイクの間隔を自由に変えられるマイクアレイを製作した。そして、MUSIC アルゴリズム(Multiple signal classification)によって複数話者の音源位置を推定できることを確かめた。

(b)次に、空間的にどの位置におくことで、選択的聴取性が高くなるかについての評価を行った。選択的聴取性の高さの指標については、それぞれの音声の明瞭度の和とした。図 1 に 2 つの音源の方位角と客観的明瞭度の和の関係を示す。方位角は、正面を 0 度として、時計回りの方向を正とする。赤い丸で囲った箇所に示すように、右耳の方向である 90 度や、左耳の方向である 270 度に音源が集中すると明瞭度が下がることがわかった。また、片方の音源が 90 度にあり、もう片方の音源が 270 度の角度にある場合の明瞭度が高いことが分かった。

②周波数軸上で音声が被らないようにするために、オンライン話者の音声に対する加工方法として、(a)ハイパスフィルタやローパスフィルタを音声にかける、(b)音声明瞭度を最大化するような畳み込みニューラルネットワーク、ということに取り組んだ。

(a)最初の実験として、2 つの音声のうちの 1 つにローパスフィルタ、もう 1 つの音声にハイパスフィルタをかけた上で足し合わせ、客観的明瞭度を計算した。その結果、ローパスフィルタをかけた方の音声の明瞭度は上がり、ハイパスフィルタをかけた方の明瞭度は下がり、明瞭度の和は低下した。このようにパラメータが入力した音声によって変わらないフィルタでは、話者や語音の種類によって選択的聴取性のしやすさが異なり、選択的聴取性があまり上がらないと考えられる。

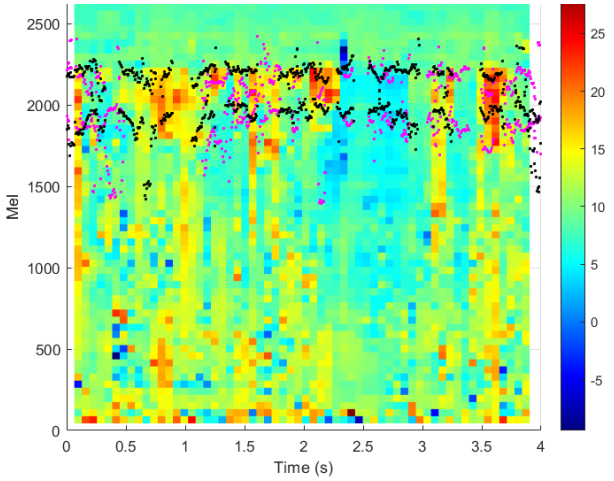


*図 2 提案手法により強調された箇所を示すスペクトログラムと、第 3 第 4 フォルマントの位置の例。*

(b)次に、(a)のフィルタを汎用的にするにあたり、話者の個人性と発話内容を保持しながら選択的聴取性を向上させる畳み込みニューラルネットワークを作成した。学習のコストとしてはそれぞれの音声に対して，選択的聴取性を示す指標となる客観的明瞭度 Extended short-time objective intelligibility (ESTOI) (Jensen & Taal, 2016) と入力音声と出力音声の間の Multi-scale spectrogram loss を計算し，その和を用いた． その結果，入力音声の ESTOI の和の平均値は，1.103，出力音声の ESTOI の和の平均値は 1.119 となり，明瞭度は有意に向上した($p < 0.01$)．図 2 に提案するネットワークで処理する前後のメルスペクトログラムの差を色で示した例を示す．2000 メル付近のピンク印は 1 人目の話者の音声の第 3，第 4 フォルマントを表し，黒印は 2 人目の話者の音声の第 3，第 4 フォルマントを表す．図 2 より，2 人の話者の音声の第 3，第 4 フォルマントが強調されるようなフィルタが学習されたのではないかと考えらえる。なお、音声の第 3、第 4 フォルマントは、人間が話者認識するために用いられていると考えられている。

**研究課題(iii)オフライン話者とオンライン話者の音声の選択的聴取を支援するデバイスの実装・評価**

研究課題(i)(ii)の音声処理をユーザが実際に使用できる装置の試作を行った。研究課題(i)①のオンライン話者とオフライン話者の音声の到来方向が被らないようにするため、オフライン話者の音源位置を推定できる、ウェアラブルのマイクアレイの製作を行った。今後は研究課題(i)(ii)音声処理をリアルタイム動作できるようにする。そしてシステムの効果を評価するために、オフライン話者とオンライン話者の音声を同時に呈示し、試行ごとに注目する音声を変えながら内容の書き取りをしてもらい、その正解率をみる実験や、本システムを日常生活で実際

に使ってもらい、認知負荷、他の話者とのインタラクションの変化を評価する予定である。

**研究課題(iv)直観的なオフライン話者とオンライン話者間の音量調整のための耳介インタフェースの検討**

研究課題(i)を実施する中で、複数話者間のわずかな音量の差が聞き取りやすさに大きな影響を与えることがわかった。したがって、複数話者がいる中で、ユーザがより集中したい音声があるとき、そちらの音声に集中しやすくするようなインタフェースも必要となる。そこで、耳介の上で指をなぞった位置を連続的に推定することで、話者同士の音量調整を行うインタフェースを提案した。本手法では指の位置に依存した骨伝導の音漏れの変化を観測することで、耳介上の指の位置を連続的に推定できる。そのため、必要である微妙な音量調整が可能である上に、画面等を見て調整する必要がないため、目の前にいるオフライン話者とのコミュニケーションを妨げることがないというメリットがある。

3. 今後の展開

今後は、本課題で取り組んだ選択的聴取性向上のアルゴリズムを実際に使用できるものにするために、学習データセット内の音声だけでなく、ユーザの周りの話者に対応できるようにする。さらに、小型端末においてリアルタイムで動作するように改良を行う予定である。そして本研究成果の社会実装という意味では、5 年以内に、実際のハイブリッド会議で使用できるシステムを実装することが考えられる。

4. 自己評価

選択的聴取性向上は、2 人の話者の聞き取りやすさを同時に上げるというゼロサムゲームに近いものがあり、非常に挑戦的なものであった。

また、研究計画当初は予想できなかった方向性により、選択的聴取性向上へのアプローチをすることができた。領域会議におけるある先生の講演や、ディスカッションがなければ生じなかった発想と考えている。それをきっかけに 2 人の選択的聴取性向上というゼロサムゲームに対する一つの突破口を見出すことができた。

また、当初設定した目標である音声処理のリアルタイムシステムを完成させるところまで達成するとはいかなかった。そのため、研究成果が実際に社会や経済に貢献するためには、実環境に適用させるための研究が必要となる。

ほかにも、ACT-X のほかの研究者に対して貢献ができたと考えている。ACT-X 1 期生で北九州市立大学の研究の音響応用にあたり、測定についてのディスカッション、測定設備の検証などを行った。そしてこの結果を第 66 回システム制御情報学会 研究発表講演会に共著として発表した。このようにして理論研究を得意とする藤本先生に対して、自身の実測経験を活かし、理論の実応用の検証に貢献できたと考えている。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：1件

| |
|---|
| 1. Yuke Zhang, Ken Takaki, Hiroaki Murakami, Takuya Sasatani, and Yoshihiro Kawahara. Poster Abstract: Toward Continuous Finger Positioning on Ear Using Bone Conduction Speaker. SenSys '22: Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems. To appear. |
| 骨伝導ワイヤレスヘッドセットなどの小型音響ウェアラブルデバイスが普及している。しかし、このような小型の機器では、タッチスクリーンなどの大きく、直感的な入力装置を搭載できないため、機器操作が不便である。本論文では、骨伝導スピーカとマイクで測定した音響特性を用いて、耳をタッチ入力インタフェースとして利用する手法を提案した。耳の異なる部位に指を置くと、耳の音響放射特性に影響を与え、漏れる音が変調することを発見し、この効果を利用することでタッチ位置を推定することができる。 |

（2）特許出願

なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. 高木 健, 齋藤大輔, 川原圭博, "畳み込みニューラルネットワークによる複数話者音声の選択的聴取性向上, " 電子情報通信学会ソサイエティ大会, A-5-4, オンライン, Sept. 2021.
2. 北九州市立大学ひびきのキャンパス, "ディープラーニングによる聴覚拡張の研究", 2021/12/23