

オンライン会議において相手に与える印象を 変化させるためのリアルタイム語尾ピッチ変換システム

西原 宗太郎¹ 渡邊 拓貴² 寺田 努¹ 塚本昌彦¹

概要：

新型コロナウイルス感染症の流行による職場や教育におけるリモートワークや遠隔作業の推進に伴い、オンライン会議システムが普及しつつある。オンラインでのコミュニケーションは話し手の感情、雰囲気といった音声と映像のみでは伝達が難しい情報が存在するため、聞き手が話し手の感情を理解しづらいという課題がある。話し手の心理状態を理解する手段として、音量の強弱や声の高さといった周辺言語が重要視されているが、オンライン会議においてはそれが伝わりづらい。周辺言語の中でも特に語尾のピッチ変化は話し手の性格と印象を決定する重要な要素であるため、オンライン会議において聞き手に与える印象を変化させるために、語尾のピッチを変化させることが有効である可能性が高い。そこで本研究では、オンライン会議において話し手の意図する印象を聞き手に与えるためのオンライン会議システムを提案する。具体的には、オンライン会議時の発話から語尾を検出し、検出した語尾に対してピッチ変換を行う。語尾のピッチ変換の有効性を確認するため、提案システムの使用状況を再現した実験を行った。語尾ピッチを上昇させた場合、聞き手は話し手に対して元気である印象を抱く傾向が有意に増加し、提案システムが印象変化に対して有効であることが示唆された。

1. はじめに

新型コロナウイルス感染症の流行により人と人との直接的な接触が制限され、オンライン会議システムが普及しつつある。オンライン会議システムは在宅勤務時の会議や授業、飲み会など、ビジネスだけでなく様々な場面で用いられている。オンライン会議では対面でのコミュニケーションと異なり、マイクやカメラを通じた音声や映像でコミュニケーションを行うため、マイクでは取得しづらい細かい息遣い、声の強弱や、カメラでは記録できない全身を用いたジェスチャといった情報の伝達は難しい。オンラインでのコミュニケーションは対面で行うコミュニケーションと違い、話し手の感情、雰囲気といった音声と映像のみでは伝達が難しい情報が存在するため、聞き手が話し手の感情の理解がしづらい [1]。

このように、オンライン会議では音声情報の重要性はますます高まっている。音声によるコミュニケーションは、話の内容そのものを示す言語情報と発話音声のアクセントや音量の強弱、声の高さ、間、発話の速さといった発言内

容以外の情報を示す周辺言語の2つの情報で構成される。周辺言語は具体的な事項を伝達する言語情報と違い、話者の感情、雰囲気、態度といった言語化の難しい情報を伝達する。対話中に発話者の周辺言語によって表れる、怒っている、楽しんでいるなどの感情と言語情報の内容が矛盾した場合、聞き手は93%で周辺言語によって表される情報を信頼するという研究がある [2]。このように、周辺言語の聞き手に与える影響は大きい。また対人コミュニケーションでは、発言内容の強調、発話衝突の回避を周辺言語の中でも特に声の高さ、大きさ、間、発話の速さを変化させることで行っており、話の要点の強調や、発言権の自然なやり取りを可能にしている [3]。周辺言語の中でも特にアクセント、声の大きさ、語尾ピッチは話し手の性格と印象を聞き手に類推させる役割を果たす [4,5]。そのため、オンライン会議において実際の話し手の心理状態を理解する手段として、音声と映像に含まれる周辺言語は重要視される。

既存のオンライン会議システムには、映像に関する支援技術は複数提案されている。例えば、外見を補正することで聞き手に与える印象を変化させる仮想背景や顔の明るさを調整するフィルタ機能といった、外見補正機能が実装されている。外見補正機能を使用することで、顔の明るさ、色調を調整し、聞き手にシステムユーザが伝えたい雰囲気の伝達を補助できる。一方で、音声加工による補正機能は

¹ 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University

² 北海道大学大学院情報科学研究院
Graduate School of Information Science and Technology,
Hokkaido University

音量自動調整、ノイズ抑制が存在するが、システムユーザの発話音声自体の補正を行う機能は、筆者の知る限り存在しない。感情伝達を行う周辺言語は発話語尾に集中しているという調査結果 [6] があるように、発話語尾の音声のピッチを上昇、下降させることで周辺言語を補正し、話し手の感情伝達を補助できる可能性がある。外見の補正と同様に、音声も補正が可能になれば、オンライン会議での周辺言語による情報伝達の補助につながり、聞き手により良い印象を与えることが可能になると推測できる。

そこで本研究では、オンライン会議においてシステムユーザが自身の語尾のピッチを自由に補正できるようにすることで、システムユーザの意図する印象を聞き手に与えるためのオンライン会議システムを提案する。本論文ではオンライン会議時の語尾を検出し、語尾のピッチ変換を行うシステムを提案した。提案手法の有効性を評価するため、提案システムの使用状況を再現した映像を被験者に視聴してもらい、システムユーザから受ける印象を評価させた。評価の結果、提案システムと同じ条件で語尾ピッチを上昇させた場合、システムユーザに対し元気である印象を抱く傾向があるという結果が得られた。

本論文は以下のように構成されている、2章で関連研究を紹介し、3章では提案システムについて述べる。4章で提案システムを再現した映像視聴による印象評価実験と考察を行い、5章で実験結果と考察をふまえた議論を行う。最後に6章で本論文をまとめる。

2. 関連研究

2.1 周辺言語の役割

コミュニケーションにおける周辺言語の影響を調査する研究が行われている。Banziger らは、様々な感情が含まれる音声の定量的な韻律分析を行い、発言に含まれる基本周波数の振れ幅、平均を調べることで発言にどのような感情が含まれるかを分類できると述べている [7]。鶴谷は、録音音声の分析を行うことにより、日本語における丁寧な発話に現れる韻律的特徴として、ゆっくり話す、低めに話す、語尾のピッチを次第に上げる傾向にあると述べている [8]。Dahlback らは、被験者の居住地域特有のアクセントとそれとは異なるアクセントの2種類の英語を使用した場合では前者の方に類似性を感じ、信頼する傾向にあると評価した [9]。市川らは、非対面で行われた会話データに含まれる声の大きさ、高さ、間、発話の早さからなる抑揚情報を解析することにより、会話相手の抑揚情報に合わせて自身の抑揚情報を操作することで発言権、話題の転換のやり取りを制御していると述べている [3]。郡は、日本語における語尾のアクセントを分類し、それらを再現した合成音声聞いた際に受ける印象の調査を行い、合成音声でも語尾のアクセントの違いにより聞き手の受ける感情が異なることを確認した [4]。これらの先行研究から、オンライン通

話での入力音声に補正を施すことにより聞き手が受ける感情に影響を与える可能性がある。発話に含まれる感情に応じて周辺言語を変化させるシステムは、聞き手が受ける印象を変えるために有用であるが、オンライン通話上でも同様の効果が得られるかを調査した研究は、筆者の知る限り存在しない。

2.2 音声加工による補助

音声信号処理技術の発達により、ソフトウェアなどを用いて発話音声のピッチやイントネーションを変化させることが可能になっている。常島らは、発話における助詞や語尾のピッチ変換によって聞き手が受ける印象に影響を与えるかについて調査を行い、語尾と助詞の基本周波数を下降させた音声は通常の音声と比べて、「経験が豊富である」、「説得力がある」などの印象をより強く与えることを示した [10]。しかし、この研究では、あらかじめ録音した音声の語尾と助詞を加工したものを使用しており、日常のコミュニケーションへ応用することを想定するとリアルタイムで加工されることが望ましい。亀岡は、音声を構成する基本周波数を作り出す役割をもつ器官である甲状軟骨のメカニズムを模擬した制御物理モデルによる基本周波数パターンの高精度な推定を行うことにより、発話者特有の基本周波数を求め、入力音声のアクセントを自由に操作することを可能にした [11]。外川らは、携帯電話での受信音と周囲雑音の SN 比に応じた利得の制御による音声強調技術、無音区間の短縮と声質を変えない音声の引き延ばし技術による話速変換技術を用いて聞こえやすさ支援をした [12]。この研究では、発話音声全体を加工することで、発話内容の伝達を補助をする技術が実装されているが、発話音声を加工することによる感情表現を制御するものではない。

2.3 オンライン会議支援システム

オンライン会議でも聞き手が周辺言語から受ける影響は大きい。玉木らは、オンライン会議上で聞き手の手と頭の動きに関する周辺言語を認識しすることで、次に会話を始めようとする人を予測し、システムユーザに提示することで発話の衝突を回避するシステムを提案した [13]。井上らは発言中の人にクローズアップした映像を共有することで、発言者の周辺言語をより多く伝えるシステムを実装した [14]。shaburov らは、聞き手の顔の表情、声の特徴を分析し、否定的なものであった場合はシステムユーザに知らせることで円滑なコミュニケーションを図るためのシステムを実装した [15]。伊藤らは、仮想空間上にアバタを導入し、アバタの表情を操作可能にすることでカメラ越しに話者が表現することが難しい感情を表現することや、アバタの立ち位置を操作することで自身の発言意欲を示すことが可能なオンライン会議アプリケーションを提案している [16]。これらの研究では、オンライン会議を使用する際

に聞き手に出力する映像を変化させることによって、聞き手にオンライン会議上では伝わりづらい周辺言語を適切に伝達する仕組みがなされていたが、聞き手に対し新たに周辺言語情報を与えるものではない。本研究ではオンライン会議での発話音声の語尾のみを加工することで、聞き手に与える周辺言語情報を付加する。

3. 提案システム

本研究では、聞き手により良い印象を与えるために、システムユーザが自身の語尾ピッチを自由に補正できるシステムを構築することを目指す。

3.1 システムの概要

提案システムの流れを図1に示す。システムユーザの発話はユーザ側のPCのマイクで取得されている。PCでは、リアルタイムで入力された音声データを蓄積し、無音状態を判定する。システムユーザの発話から語尾を自動検出し、システムユーザの意図に沿ったピッチ変換を語尾に適用し、聞き手側のPCに送信する。入力音声を蓄積することにより、語尾の検出に必要な発話終了タイミングを検出し、検出した語尾に対してピッチを変換する。

3.2 語尾判定手法

リアルタイムで入力された音声データの語尾判定を行うために、発話終了タイミングを検出し、語尾の長さを定義する。本研究における発話終了タイミングと語尾の長さの定義について図2に示す。語尾検出のためには、発話終了タイミングを検出する必要がある。発話終了タイミングの検出には、発話終了後の無音区間を用いる。無音状態が一定時間続いた時の無音区間の開始時点が発話終了タイミングとする。発話終了タイミングから語尾の長さだけ遡った範囲を語尾として、ピッチを変換する。語尾の長さとは無音区間の長さを定義する必要があるが、語尾の長さとは発話終了後の無音区間の長さは使用する語尾の種類、話す速度によって変化するため一意に定めることは困難である。したがって本研究では、実験的に無音区間と語尾の長さを定めた。無音区間の長さの制定には、発言中に文節で区切ったり息継ぎしたりすることで音声途切れる瞬間との誤認識を防ぐための適切な値にする必要がある。1対1で行われた1分程度の会話を4種類録音し、録音音声の分析を行った結果、考え込むなどの無音区間を除いた、会話に含まれる短い無音区間は平均0.53秒であった。この値にマージンを持たせ、本研究では、0.7秒間無音が続いた場合を発話終了と定義する。録音音声の語尾の長さの平均は0.23秒であった。分析を行った録音音声には「です」、「ます」、「よ」、「ね」、「か」の5種類の語尾が含まれていた。語尾区間では語尾開始時から次第に音量が小さくなっていくため、語尾終了時付近の音声を除いた語尾の長さを0.2秒と

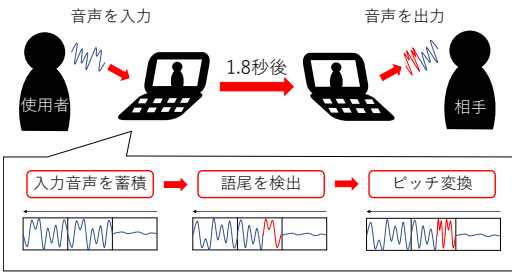


図 1: システムの流れ

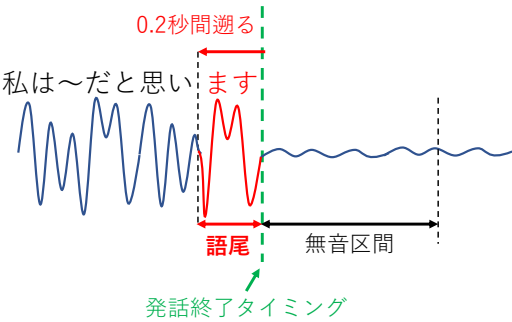


図 2: 語尾ピッチ変換手法

	効果	アクセント
強調型上昇調	ぜひわからせたい気持ちを込める	分かり ^{ます}
急下降調	気づいて分かったことを示す	分かり ^{ます}
疑問型上昇調	答えを求める・反応を待つ	分かり ^{ます}
上昇下降調	気づかせ、わからせたい気持ちを込める	分かり ^{ます}

図 3: ピッチ変換により得られる効果とアクセントの種類

定義する。したがって、0.7秒間の無音区間があった場合、直前の0.2秒間の音声データを語尾としてピッチ変換の対象とする。

3.3 ピッチ変換手法

提案システムでは強調型上昇調、急下降調、疑問型上昇調、上昇下降調 [4] を補正する。提案システムのピッチ変換により得られる効果とアクセントの種類をまとめたものを図3に示す。本研究ではこれらの中から最も基本的である強調型上昇調、急下降調の2種類に限定して調査を行った。オンライン通話に接続中にアプリケーションを操作し、システムユーザが伝えたい感情に合わせてピッチ変換の種類を切り替える。

3.4 実装

3.2節の設計をもとに、入力音声の語尾のピッチ変換をリアルタイムで行うシステムを実装する。実装にはPython [17]を用いる。PC内の処理を図4に示す。まず、PCは入力音声に対して無音状態の判定を行い、ピッチ変換範囲を決定する。決定したピッチ変換範囲に対してピッチ変換を行い、仮想ミキサによってピッチ変換後の音声を

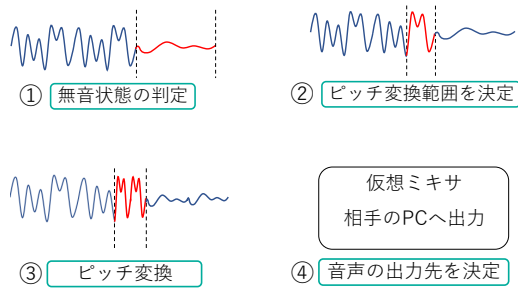


図 4: PC 内の処理

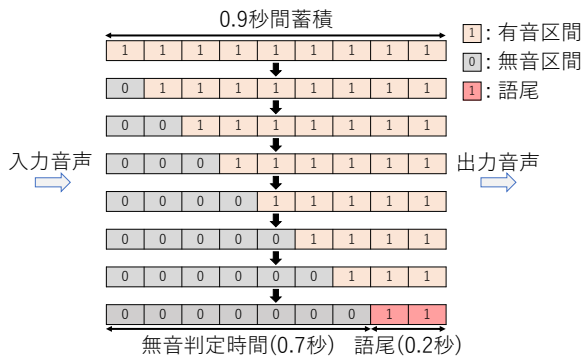


図 5: 音声蓄積の流れ

聞き手に出力する。音声蓄積のイメージを図 5 に示す。発話終了タイミング検出に要する 0.7 秒と、語尾の長さである 0.2 秒の合計 0.9 秒間の音声を蓄積し、蓄積した音声にピッチ変換を施した音声を聞き手側へ出力する。また、無音区間の検出には入力音声信号の強さが必要になる、入力信号の強さ $L[\text{dB}]$ は 16bit で量子化された入力信号の絶対値 S を用いて以下のように求められる。

$$L = 20 \log_{10} S$$

提案システムでは、無音状態を入力信号の強さが 60dB 以下の時とする。0.7 秒以上の無音区間がなければ蓄積した音声を取り出してピッチ変換を行わず聞き手側へ出力し、0.7 秒以上の無音状態を確認した際には、語尾と判定された 0.2 秒間の範囲のみをピッチ変換した音声を聞き手側へ出力する。出力音声は仮想ミキサ VoiceMeeter Banana [18] を用いて聞き手に送信する。本実験では試験的に語尾のピッチを上昇させる場合は語尾に含まれる基本周波数の 1.2 倍の周波数に、語尾のピッチを下降させる場合は語尾に含まれる基本周波数の 0.8 倍の周波数にする。ピッチ変換は TD-PSOLA 方式 [19] を用いて行う。

提案システムの出力音声は 1.8 秒の遅延が生じており、遅延の要因は語尾検出遅延と処理遅延である。語尾検出遅延とは、語尾を検出するタイミングを決定する蓄積音声内の無音区間の検出に要する 0.7 秒間である。処理遅延とは、蓄積音声内の語尾の長さである 0.2 秒に加え、PC が音声の入出力、ピッチ変換、音声蓄積の処理に要する時間を加えたもので合計は実測値で 1.1 秒であった。聞き手側に出

力される音声と映像のズレが 480ms 以上になると通話が困難になることが報告されている [25]。したがって、本研究では映像を 1.8 秒遅らせて出力することで、音声と映像のズレを軽減させる。映像の遅延には Open Broadcaster Software [20] の仮想カメラ機能を使用する。ビデオ通話中に発生する往復遅延時間が 1050ms を超えるとビデオ通話体感品質に影響を与える [21]。提案システムでの PC での処理速度があがれば、提案システムによる音声の遅延を音声蓄積の長さである 0.9 秒に近づけることができ、ビデオ通話体感品質に影響を抑えることができる。

4. 評価実験

オンライン会議において提案システムを利用した際、聞き手が受ける印象の変化について評価実験を行った。

4.1 実験の設計

本稿では、オンライン会議における語尾ピッチ補正が、システムユーザの印象を変化させるのかをまず明らかにするため、手動で語尾ピッチ補正を行った映像を用いた評価実験を行った。オンライン会議において発話語尾のピッチ変換が聞き手の受ける印象に影響を与えると期待される状況を想定し、1 対 1 で会話を行う映像を用意する。システムを利用した場合と利用しない場合とでビデオを見た人がシステムユーザに対して受ける印象についてそれぞれ評価を行う。対話において聞き手が受ける印象のうち、提案システムの音声補正が影響を与えられる点は以下の通りである。

- 自信がある
- 丁寧である
- 元気がある
- 説得力がある

先行研究より、対人コミュニケーションにおいて語尾ピッチを下降させることで「丁寧さ」、「言い切り」の印象が強くなることが報告されており、「言い切り」は説得力の向上につながると考えられる [8, 22] また、語尾ピッチを上昇させることで「力強さ」の伝達が可能になる [23]。「力強さ」を伝達することは元気、自信があることにつながると考えられる。したがって、本研究ではピッチ上昇により「自信がある」、「元気がある」印象に影響を与え、ピッチ下降により「丁寧である」、「説得力がある」印象に影響を与えると考え、上記した 4 つの印象を評価対象とした。

4.2 実験用映像

上記の印象が重要になると予想される内容を取り入れた 7 種類の映像を用意した。対話の内容は映像 A、映像 B、映像 C、映像 D の 4 種類とし、映像 A、映像 B、映像 C については 2 者の発言内容を入れ替えた 2 通りの映像をそれぞれ作成した。映像の作成にあたり、会話内容はあらかじめ



図 6: システムユーザと聞き手が映っている映像



図 7: システムユーザと資料が映っている映像

め指定した。映像 A-1–C-2 では図 6 のようにシステムユーザと聞き手が表示される画面構成とした。映像 D では図 7 のようにシステムユーザと共有資料が表示される画面構成とした。7 種類の映像の概要は以下になっている。会話内容と求められる印象の関係を表 1 に示す。また、実際の会話内容を付録に示す。

映像 A-1：旅行に行きたい場所

旅行に行きたい場所についての対話で、システムユーザが話題提供を行う。娯楽に関する話題であるため、元気であることが聞き手により良い印象を与えることにつながると予想される。

映像 A-2：旅行に行きたい場所 (話者交代)

システムユーザが提供された話題に対して返答をしている。映像 A-1 と同様に元気であることが聞き手により良い印象を与えることにつながると予想される。

映像 B-1：分からないことについての相談

システムユーザが聞き手に対して相談するという内容である。相談をする際には丁寧な口調であることが聞き手により良い印象を与えることにつながると予想される。

映像 B-2：分からないことについての相談 (話者交代)

システムユーザが聞き手から相談を受けるという内容である。相談を受ける際には自信、説得力があり、丁寧な口調であることが聞き手により良い印象を与えることにつながると予想される。

映像 C-1：ディベート

システムユーザと聞き手とで 1 つの議題に対してディベートを行う。システムユーザの主張に対して聞き手が反論をするという内容である。発言内容に自信、説得力を持たせ、なおかつ丁寧な口調であることが聞き手により良い印象を与えることにつながると予想さ

表 1: 会話内容と求められる印象の関係

	自信	説得力	元気	丁寧
映像 A-1			○	
映像 A-2			○	
映像 B-1				○
映像 B-2	○	○		○
映像 C-1	○	○		○
映像 C-2	○	○		○
映像 D	○	○		○

れる。

映像 C-2：ディベート (話者交代)

聞き手の主張に対してシステムユーザが反論をするという内容である。映像 E と同様にディベートの場において発言内容に自信、説得力を持たせ、なおかつ丁寧な口調であることが聞き手により良い印象を与えることにつながると予想される。

映像 D：プレゼン

システムユーザが資料を共有しながらプレゼンを行う。プレゼンでは発言に説得力・自信を持たせる必要があり、さらに丁寧な口調であることが聞き手により良い印象を与えることにつながると予想される。

7 種類の映像それぞれについて、「加工無し」、「語尾ピッチ上昇」、「語尾ピッチ下降」の 3 種類のピッチ変換パターンを用意し、計 21 種類の映像を用意した。提案システムを再現した映像の作成手順として、まず提案システムユーザ側の音声と映像を 1.8 秒遅らせて聞き手に出力して行われた会話の音声と映像を録画し、録画した音声を 3.2 節での提案内容と同じ条件で加工した。音声加工には Audacity [24] を使用した。映像の遅延には Open Broadcaster Software [20] の仮想カメラ機能を使用した。

4.3 実験方法

実験開始時、被験者にシステムユーザ側が語尾のみのピッチ補正をしていることを伝える。その後、被験者は椅子に座った状態で 4.2 節で述べた 21 種類の映像をランダムに視聴した。各映像を視聴するごとにアンケート調査を行った。アンケートでは被験者が映像のシステムユーザに対して抱く感情について、4 つの調査項目 (Q1–Q4) を用意し、システム使用側に対して当てはまるものを回答させた。調査項目は以下の内容で、それぞれ 5 段階のリッカート尺度で回答させた。

- Q1「自信があるように感じるか」
(5: 感じる–3: どちらでもない–1: 感じない)
- Q2「説得力があるように感じるか」
(5: 感じる–3: どちらでもない–1: 感じない)
- Q3「元気があるように感じるか」

(5: 感じる- 3: どちらでもない- 1: 感じない)

● Q4「丁寧であるように感じるか」

(5: 感じる- 3: どちらでもない- 1: 感じない)

実験終了後、自由記述式のアンケートを用意し、任意で回答を得た。被験者は20代の学生7名である。

4.4 実験結果と考察

各アンケート結果の平均値を図8、図9、図10、図11に示す。エラーバーは標準偏差を示す。動画内容、ピッチ変換に被験者内2要因分散分析を行った結果、Q1、Q2、Q3に動画内容とピッチ変換の交互作用は有意差なしとなり、Q4が有意差ありであった($F(12, 72)=1.96, p<0.05$)。以下でQ1、Q2、Q3では動画内容、ピッチ変換の主効果を、Q4では単純主効果を見る。

Q1: 自信があるように感じるか

動画内容の主効果が有意なし、ピッチ変換の主効果が有意であった($F(2, 12)=5.80, p<0.05$)。ピッチ変換について Bonferroni 法を用いて多重比較を行ったところ、語尾ピッチ上げの平均値が語尾ピッチ下げの平均値より有意に大きかった($p<0.05$)。このため、提案システムを用いて自信を高める効果を求める場合、語尾ピッチ下げではなく語尾ピッチ上げを用いることが効果的であると考えられる。加工無しと語尾ピッチ上げとの間に有意な差は無いが、アンケート結果において加工無しに対して語尾ピッチ上げの平均値が高くなる傾向にあるため、提案システムを用いて自信を高める効果がある可能性がある。

Q2: 説得力があるように感じるか

動画内容、ピッチ変換ともに主効果に有意なしであった。今回の結果からは、提案システムによる説得力を高める効果は確認できなかった。

Q3: 元気があるように感じるか

動画内容の主効果が有意なし、ピッチ変換の主効果が有意であった($F(2, 6)=44.78, p<0.01$)。ピッチ変換について Bonferroni 法を用いて多重比較を行ったところ、語尾ピッチ上げの平均値が加工無しの平均値より有意に大きく($p<0.05$)、語尾ピッチ上げの平均値が語尾ピッチ下げの平均値より有意に大きく($p<0.05$)、加工無しの平均値がピッチ下げの平均値より有意に大きかった($p<0.05$)。この結果から、提案システムによって元気を高める効果を求めることは有用であると考えられる。また、加工無し、語尾ピッチ上げに対して語尾ピッチ下げの平均値は有意に小さいため、提案システムの語尾ピッチ下げを用いることで、元気でない印象を高められる可能性がある。

Q4: 丁寧であるように感じるか

動画内容、ピッチ変換の単純主効果を検定した結果、動画 A-2 におけるピッチ変換は有意であった($F(2,$

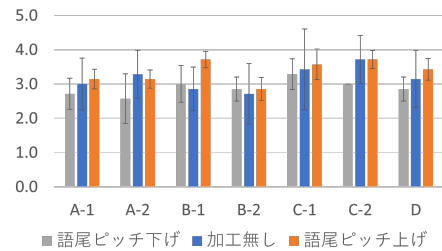


図8: アンケート Q1 の平均値と標準偏差

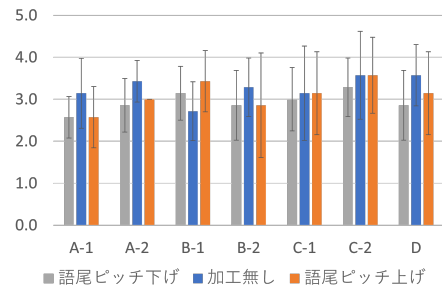


図9: アンケート Q2 の平均値と標準偏差

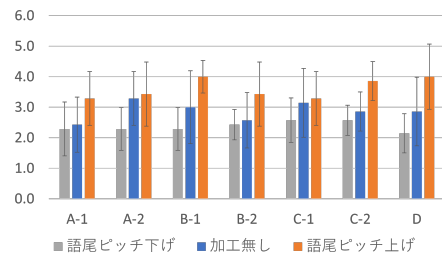


図10: アンケート Q3 の平均値と標準偏差

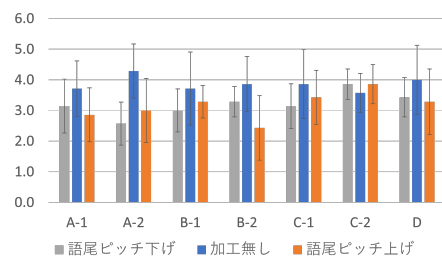


図11: アンケート Q4 の平均値と標準偏差

12)=9.75, $p<0.01$)。動画 A-2 について Bonferroni 法を用いて多重比較を行ったところ、A-2 における加工無しの平均値がピッチ下げの平均値より有意に大きかった($p<0.05$)。また、A-2 における加工無しの平均値がピッチ上げの平均値より有意に大きかった($p<0.05$)。これは映像 A-2 にてシステムユーザの話し方が加工無し状態で十分に丁寧な口調であったため、システムユーザが丁寧である印象を与えたと推察される。今回の結果からは、提案システムによる丁寧な印象を与える効果は確認できなかった。

Q1-Q4 のそれぞれで、動画内容の主効果、単純主効果は映像 A-2 を除いて有意なしであった。そのため、提案システムは自信、説得力、元気、丁寧さを高める効果を求める場合において会話の内容に依存しないと推察できる。また、Q1, Q2, Q4 において語尾ピッチ下げの平均値は有意に小さかった。提案システムの語尾ピッチ下げにより聞き手により良い印象与えることは有用ではない可能性がある。また、被験者から得られたアンケートから動画の種類によっては、ピッチ変換度合いが大きすぎる、音質が落ちているなど、システムユーザに対してかなり違和感を感じるといった意見が得られた。

5. 議論

5.1 ピッチ変換パラメータの検討

システムユーザの発話音声に対して違和感があったという意見が得られた。違和感が生じた要因として、ピッチ変換度合いが適切でなかったことが考えられる。今回の実験では、語尾区間の音声データの周波数をピッチを上昇させる場合は語尾の基本周波数の 1.2 倍、ピッチを下降させる場合は語尾の基本周波数の 0.8 倍として一律な加工を施したが、違和感を減少させるには個人の感覚に合わせたパラメータ設定が必要である。今後は個人差を考慮した適切なピッチ変換度合いについて調査を行う。

5.2 提案システムにより影響を与えるその他の印象

今回の実験において提案システムの使用によりシステムユーザに対する「自信がある」、「丁寧である」、「元気である」、「説得力がある」の 4 つの印象が変化すると仮定した。結果として「丁寧である」、「説得力がある」の印象は、提案システムの使用による変化が見られず、「自信がある」の印象は、提案システムの使用により変化を与える可能性はあるが、有意差は見られなかった。「元気である」の印象では提案システムによる語尾ピッチ上昇により有意差が見られた。そのため、提案システムの使用により「喜んでいる」「悲しんでいる」といった、今回の実験で調査を行っていないその他の感情に対して効果的である可能性があるため、今後より詳しい調査を行う。

5.3 提案システムユーザによる効果の違い

今回の実験に用いた映像では、全ての映像でシステムユーザが同一人物であった。映像ではシステムユーザと対話相手は敬語で会話をしていた。実験結果では「丁寧である」印象は提案システムの使用により評価が低下していた。これはシステムユーザが加工無しの状態ですでに丁寧であったため、提案システムが影響を及ぼさなかったと推察される。システムユーザの性格の違いによって今回の実験結果と異なる結果となる可能性があるため、今後被験者の性格と提案システムの関係についてより詳しい調査を

行う。

6. まとめ

本研究では、オンライン会議においてシステムユーザが自身の語尾ピッチを自由に補正することで、システムユーザの意図する印象を聞き手に与えるためのオンライン会議システムを提案した。特に本論文では、オンライン会議での会話中に提案システムを使用した状況を再現した映像を、被験者が視聴した際にシステムユーザに抱く印象の評価を行った。結果、提案システムと同じ条件で語尾ピッチを上昇させた場合、システムユーザに対し元気である印象を抱く傾向が見られた。今後の課題として、提案システムの使用により聞き手に違和感を感じさせないような適切なピッチ変換度合いの調査を行う必要がある。また、聞き手に違和感を感じさせない適切なピッチ変換度合いを明らかにするとともに、ピッチ変換提案システムによって影響を与えることができる聞き手が受ける印象を明らかにすることで、聞き手により良い印象を与えるための音声補正を行うシステムの構築を目指す。

謝辞

本研究の一部は、JST CREST(JPMJCR18A3) の支援によるものである。ここに記して謝意を表す。

参考文献

- [1] 馮 晨, 池田悠平, 菅谷みどり: 遠隔地コミュニケーションにおける感情伝達方法の比較, マルチメディア, 分散, 協調とモバイルシンポジウム 2017 論文集, pp. 149-153 (June 2017).
- [2] A. Merabian: Communication Without Words, *Psychological Today*, Vol. 2, pp. 53-55 (1968).
- [3] 市川 薫, 佐藤伸二: 対話理解に対する抑揚情報の役割, 情報処理学会研究報告音声言語情報処理, Vol. 57, pp. 51-58 (July 1994).
- [4] 郡 史郎: 間投助詞のイントネーションと間投助詞的イントネーション—型の使い分けについて—, 言語文化研究, Vol. 42, pp. 61-84 (Mar. 2016).
- [5] 今井田亜弓: 若い日本人女性のピッチ変化に見る文化的規範の影響, 言語文化論集, Vol. 27, No. 2, pp. 13-26 (Mar. 2006).
- [6] 前川喜久雄, 北川智利: 音声はパラ言語情報をいかに伝えるか, 認知科学, Vol. 9, No. 1, pp. 46-66 (Oct. 2008).
- [7] T. Banziger and K. R. Scherer: the Role of Intonation in Emotional Expressions, *Speech Communication*, Vol. 46, pp. 252-267 (July 2005).
- [8] 鶴谷千春: 丁寧さを表現するために日本語母語話者が用いる韻律的特徴, 国立国語研究所論集, No. 11, pp. 167-180 (July 2016).
- [9] N. Dahlbäck, Q. Wang, C. Nass, and J. Alwin: Similarity Is More Important than Expertise: Accent Effects in Speech Interfaces, *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*, pp. 1553-1556 (Apr. 2007).
- [10] 常島杏菜, 吐師道子: 語尾上げ・助詞上げ発話の「ST としての専門職らしさ」の印象への影響, 音声言語医学, Vol. 58, No. 4, pp. 339-345 (Mar. 2017).

- [11] 亀岡弘和: 音声のイントネーションとアクセントを分析, 合成, 変換 (特集 未来への羅針盤としてのコミュニケーション科学), NTT 技術ジャーナル, Vol. 27, No. 9, pp. 10–12 (Sep. 2015).
- [12] T. Togawa, T. Otani, and K. Suzuki: Speech Enhancement Technology, Speech Rate Control Technology, *the Journal of the Institute of Electronics, Information and Communication Engineers*, Vol. 96, No. 11, pp. 874–881 (Nov. 2013).
- [13] 玉木秀和, 東野 豪, 小林 稔, 井原雅行, 岡田謙一: 遠隔会議における発話衝突低減手法, 情報処理学会論文誌, Vol. 53, No. 7, pp. 1797–1806 (July 2012).
- [14] 井上智雄, 岡田謙一, 松下 温, 井原雅行, 岡田謙一: テレビ番組のカメラワークの知識に基づいた TV 会議システム, 情報処理学会論文誌, Vol. 37, No. 11, pp. 2095–2104 (Nov. 1996).
- [15] V. Shaburov and Y. Monastyrshyn: Emotion Recognition in Video Conferencing, *U. S. Patent 9576190B2* (Feb. 2017).
- [16] 伊藤冬子, 廣安知之, 三木光範: 雰囲気視覚化機能とアバター利用による合意形成のためのオンライン会議システム, 人工知能学会全国大会論文集, Vol. 6, p. 119 (2006).
- [17] Python Software Foundation: Python, <https://www.python.org/>.
- [18] VB-AUDIO Software: Voicemeeter Banana, <https://vb-audio.com/Voicemeeter/banana.htm>.
- [19] E. Moulines and F. Charpentier: Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones, *Speech Communication*, Vol. 9, pp. 453–467 (1990).
- [20] OBS Project: Open Broadcaster Software, <https://obsproject.com>.
- [21] 神永 睦, 内海哲史, 高原 円, 中川和重: Skype によるビデオ通話におけるネットワーク遅延と会話のしやすさの関係, 第 80 回全国大会講演論文集, Vol. 2018, No. 1, pp. 301–302 (Mar. 2018).
- [22] 小山哲春: 文末詞と文末イントネーション, 文法と音声, pp. 97–119 (1997).
- [23] 森山卓郎: 文の意味とイントネーション, 講座 日本語と日本語教育, Vol. 1, pp. 172–196 (1989).
- [24] The Audacity Team: Audacity, <https://www.audacityteam.org/>.
- [25] M. Hashimoto: Effect of Intermodal Desynchrony on Audio-Visual Speech Perception, *the Japanese Journal of Ergonomics*, Vol. 29, pp. 482–483 (1993).

付録

動画のスク립トを以下に示す。() は語尾を指す。

旅行に行きたい場所

最近旅行とかいけてないですよ (ね)。そうです (ね)。今コロナとかで難しいと思うんですけど、卒業したら旅行に行きたいところあります (か)。国内だと東京に行きたい (です)。東京です (か)。旅行とかじゃなくても行けそうですけど (ね)。まだ一回も行ったことがないから行ってみたいんです (よ)。それに、スカイツリーにも一度登ってみたいです (し)。確かに僕も行ったことがないので行ってみたい (です)。海外だったらどこに行きたいです (か)。海外だとタイに行きたい (です)。なんかめずらしいです (ね)。行ったことある

んです (か)。1 回行ったことがあるんです (よ)。あそこは物価が安いから贅沢できるから (です)。それは 1 度行ってみたい (です)。どこにいつ行ってみたい (か)。僕は台湾行きたいです (ね)。台湾は日本人に人気高いですよ (ね)。一度行ったことがあって、あそこで食べる小籠包がめちゃくちゃおいしいんです (よ)。

相談

お疲れ様 (です)。おつかれさま (です)。今、論文の書き方を悩んでいて、相談してもいいです (か)。だいじょうぶ (です)。ありがとうございます (ます)。今、卒論で使う参考文献を探してるんですけど、なかなか見つからないんです (よ)。いつもどうやって探してます (か)。僕は Google scholar を使って検索して検索して (ます)。僕もそれを使って調べてるんですけど、調べてもなかなか出てこないですよ (ね)。英語のワードで調べると出やすいかもしれない (です)。なるほど、それで調べてみ (ます)。ついでになんですけど、参考文献の書き方ってどうすればいい (か)。文献の書き方は決まりがあって、先輩たちの論文を参考にすればいい (よ)。書き方をまとめたものがありますので送っときます (ね)。ありがとうございます (ます)。

ディベート

私はコロナでの自粛は反対 (です)。菅総理大臣が緊急事態宣言が延長するという発言がでるほど世間ではコロナに対する認識を改めるような意見が出て (ます)。しかし、たとえこの状況下でも国民全員が自粛を守ってしまうと経済が停滞し、日本全体が大きなダメージを受けると考えるから (です)。確かに、経済が停滞してしまうのは確か (しょう)。しかし、たとえ経済が停滞したとしても人命を最優先にしなければこれから先に発展する経済自体が無くなってしまう可能性があります (ます)。感染者数の分析を行ったところ、大型連休などのいわゆる気のゆるみが発生したタイミングと感染者数の相関関係は見られなかったよう (です)。これは外出自粛による効果はなかったという事実であるとも考えられ (ます)。それはあくまでも結果論であり、もし自粛に取り組んでいなければさらなる拡大が生じていた可能性も考えられ (ます)。未知の現象に対策を考えるのは当然のことだと考え (ます)。

プレゼン

それでは、リズムゲームのすすめという題目で神戸大学の〇〇が発表をさせていただきます (ます)。まず、わたくしがリズムゲームを進める理由といたしまして 3 つ挙げられ (ます)。一つは音楽に合わせてたたくのは楽しいということが挙げられ (ます)。リズムゲームには集中力が必要になり (ます)。何度も繰り返し練習していくうちに短時間での成長を実感することができ (ます)。次にリズムゲームのタイプについて述べさせて

いただき (ます). リズムゲームの種類は2種類ありまして, アーケード型と家庭用ゲーム機という風に分けられ (ます). これら2つの種類の比較を行い (ます). アーケード型について (です). アーケード型はゲームセンターにしかないため, 1プレイ 100円かかるため上達するのに時間がかかり (ます). それに対し, 家庭用ゲーム機はゲーム機やスマホでできるため1プレイごとにお金がかからないため上達するのが容易になり (ます). そのため, わたくしとしましては家庭用ゲーム機の方がお勧めになり (ます).