

感情音声合成を利用した コミュニケーション支援システムの提案

儀間 裕太^{†1} 伊藤 守^{†1}

概要: 人はコミュニケーションを行なうとき、相手の発する声から感情を判断することができる。ところが、声を出せない人は、何らかの意思伝達装置を利用して意思を伝えるため、感情を伝えることが難しい。本研究では、声を出せない人を対象に、意思と感情を伝達することができるコミュニケーション支援システムを提案する。本システムは、PC、Web カメラ、アイトラッカー、およびアプリケーションから構成され、本アプリケーションは、視線検知、表情検知、感情音声合成、語句予測変換、および視線入力用 GUI (Graphical User Interface) の機能を有している。

Proposal of Communication Support System using Emotional Speech Synthesis

YUTA GIMA^{†1} MAMORU ITO^{†1}

Abstract: When a person performs communication, the person can determine the feeling of the partner from his/her voice. However, a person who cannot utter a voice has difficulty in conveying feelings due to the use of an intention transmission device. This research proposes the communication support system which can transmit both intention and feelings. This system consists of a pc, a web camera, an eye tracker, and application software. The software has gaze detection, facial expression detection, emotional speech synthesis, predictive conversion of words and phrases, and GUI (Graphical User Interface) for gaze input.

1. はじめに

人はコミュニケーションを行なう際、主に自分の意思を声に変えて伝える。それに加えて、身振り手振りの仕草や顔の表情の変化が聴者に伝わることになる。聴者は、受け取ったそれらから、言葉の持つ意味や、話者の感情を理解する。そして今度は、聴者が話者となり、自分の意思を声に変え、伝える。これが繰り返し行われ、対面コミュニケーションは成立する。しかし、病気の症状など、何らかの理由で声が出せない人は、自分の意思を自分の声で伝えることができない。

こうした課題に対応するために使用されるものが、意思伝達装置である。意思伝達装置には、視線、体の動きなどを検知することで、利用者の意思を文章に変える機能が備わっており、その文章を音声合成や事前に登録した利用者の声へ変換して伝えることができる。しかし、一方で、利用者の感情を音声のみで伝える手段はあまり提案されていない。

スマートデバイスでは当然となっている画面のタッチ操作や、音声アシスタントの音声操作などの NUI (Natural User Interface) が、日常生活における入力手段として増えていることが示されている [1]。その他にも、意思伝達装置で実現されている視線検知や、Web カメラの画像から人

の顔を抽出して表情を検知する技術など、人間が自然に行なう動作を検出する技術開発も多く行われている [2] [3]。

そこで本研究では、視線検知技術、表情検知技術、感情音声合成技術を組み合わせて、音声のみで感情を伝える手段を提供するコミュニケーション支援システムを提案し、評価用プロトタイプを開発したので報告する。本システムでは PC、表情検知を行なう Web カメラ、視線検知を行なうアイトラッカー、感情音声合成サービスへ通信するためのインターネット接続を使用する。利用者が PC の画面に表示される文字ボタンを見ると、見ているところにマウスカーソルが移動、その箇所を見ながら左目を瞬きすると、文字ボタンがクリックされる。これにより、利用者は文字を選択でき、繰り返すことで文章を作成することが可能となる。その操作と同時に顔の表情を検知し、感情の推定とその強度を計測し続ける。作成した文章と計測した感情情報を付加し、感情音声合成サービスと通信を行なう。そして、感情を表現する音声再生されるため、利用者は自分の感情を音声のみで伝えることができ、聴者は利用者の意思と感情を理解すると期待できる。

また、本システムの有効性の指標について、視線操作の使用性、表情から推定された感情と実際の感情との合致度、および感情音声合成を聞いたときの聴者の感情の理解度の 3 点を設定し、それぞれ検証を行なった。

^{†1} 神戸情報大学院大学 情報技術研究科
Kobe Institute of Computing; Graduate School of Information Technology

2. 関連研究

音声合成に感情を含める方法として，飯田ら [4] [5] [6] は感情音声データベースを用いた感情表現可能な合成音声（感情合成音声）を提案している．その後さらに，男女の感情音声データベースを作成し，男女毎の喜び，怒り，悲しみの感情を含んだ合成音声を作成できるシステムを試作している．

また、安村 [7]のコミュニケーションとインタラクションによると、人と人とがコミュニケーションしているとき、気持ちが入ってくるとなんらかの感情がそこに含まれているのが一般的であり、人同士のコミュニケーションではノンバーバルな情報のやり取りの方が主体となっていると論じている。

3. 提案システム

本研究で提案するシステムは図 1 のように、PC、Web カメラ、アイトラッカー、およびアプリケーションから構成される。なお、それに加えてインターネット接続環境が必要となる。また、本アプリケーションは、視線検知、表情検知、感情音声合成、語句予測変換、および視線入力用 GUI の機能を有しており、以下にそれぞれの詳細を記す。

3.1 視線検知

視線検知は、アイトラッカーを使用する。本研究では画面上に並べた文字を視線操作で選択することから、より高精度な検知性能が必要であると考えた。そこで、他製品に比べ視線検知の精度が高い EyeTribe 社製 The Eye Tribe Tracker を採用した。また、視線操作でボタンクリックを行なうため、マウスでの操作感に近付け、左目の瞬きを対応付けて実装することとした。

3.2 表情検知

表情検知は、Intel 社が公開している Intel Perceptual Computing SDK[a]を利用することで可能となる。この SDK は、3D カメラを使用することで多彩な機能を実現できる。本研究に使用する表情検知は 3D カメラが必須ではないため、Web カメラを使用した。また、検知した表情から感情を推定し、その情報は視線入力用 GUI アプリケーションへ送信される。

3.3 感情音声合成

本研究で使用する喜び、怒り、悲しみの感情音声合成に対応しており、Web API ライブラリが公開されているサービス[b]ということで、HOYA サービス社の VoiceText を選定した。ベータ版であり、利用登録することで無償利用できる。

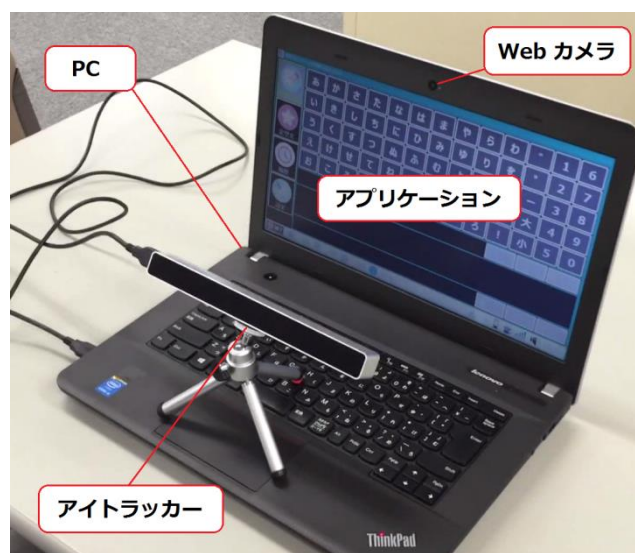


図 1 システム構成

Figure 1 System Configuration



図 2 Google サジェスト機能

Figure 2 Google Suggest Function

3.4 語句予測変換

語句予測変換とは、文字を入力した際に全ての文字を入力するより前に、単語候補を提示する機能である。図 2 のように Web ブラウザにおける語句予測変換はサジェスト機能[c]と呼称されることもある。表示された候補を選択して入力することが可能になるため、文章作成時間の短縮が期待できる。より多く検索に利用されているため [8], Google Suggest API を採用した。

3.5 視線入力用 GUI

本システムのメインとなる視線入力用 GUI は、誤操作を防止するため、全画面表示で起動され、タスクバーは表示されない。アプリケーション起動後は、視線で直感的に操作することが可能となる。次に、画面構成について図 3 に示す。また、画面下に配置したステータスバーの左部分には、取得した表情から無感情、喜び、怒り、悲しみの 4 つの推定された感情の割合を表示し、その時点で一番強い感情を青く表示するようにした。ここで、無感情は便宜的に呼称しており、喜び、怒り、悲しみ以外に推定された感情を示す。なお、ステータスバーの右部分には PC のステータスとして、Web カメラ有無、アイトラッカー接続有無、インターネット接続有無、およびミュート、音量などのス

a) <https://software.intel.com/forums/intel-perceptual-computing-sdk>
b) <https://cloud.voicetext.jp/webapi>
c) <https://www.google.co.jp/>

ピーカー設定を表示する。それから、画面左には、4つの機能を切り替えるためのタブが表示されている。図3にも示している文字入力から文章を作成して音声を再生する文字入力、事前に登録した文から再生する文章を選択する定型文、入力履歴から再生する文章を選択する履歴、および声質などを調整する設定の4つについて以下にそれぞれの詳細を記す。

まず、文字入力機能について説明する。視線操作が可能な状態で文字盤に目を向けると、図4のように見ている箇所のボタンが黄色に変化して強調される。その状態で、左目を瞬きすることで、その文字が入力され、その操作を繰り返して行なうことで、図5のように文章を作成することができる。作成した文章をそのまま確定することもできるが、入力された文章から Google Suggest API を利用した漢字変換と語句予測変換の候補が図6のように表示され、利用者はそれらの候補から選択することもできる。文章を作成して確定、または変換候補から選択した文章は、図7のように画面一番下へ表示される。表示された文章に対して、入力から完了までに取得した感情の割合から最も強い感情を付加し、感情音声合成サービスへ送信する。送信したデータが正常であれば、感情音声合成サービスから音声データを受信してスピーカーから再生される。また、よく使う文章として登録したい場合は、ここで定型文として登録することも可能である。

次に、定型文機能について説明する。これは頻繁に使用する文章やお気に入りの文章を、文字入力機能によって予め登録しておくことができる機能である。登録された文章は図8のように表示され、選択するだけで音声を再生できるため、文章を入力する手間と時間の短縮が期待できる。文字入力機能と同じく、再生したい文章に目を向け、左目を瞬きする。それと同時に表情から感情を推定して、感情音声合成サービスとの通信を行い、音声をスピーカーから再生する。その他に、登録されている定型文を削除することも可能である。

さらに、履歴機能について説明する。これはこれまで音声として再生した文章が図9のように表示される。定型文機能と同じく、選択するだけで音声を再生できるため、同様の短縮効果が期待できる。なお、操作方法などは定型文機能と同じであるが、履歴の削除機能は提供しない。

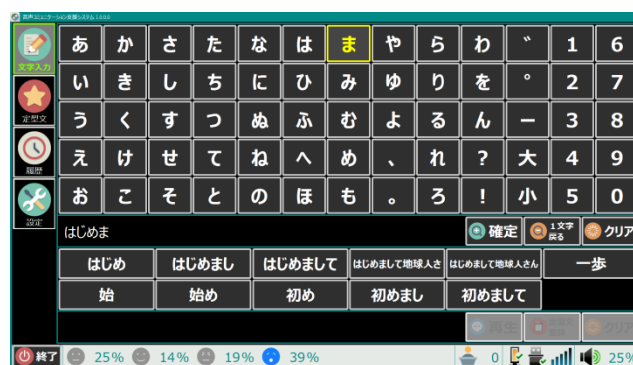


図3 画面構成と文字入力機能

Figure 3 Screen Structure and Character Input Function



図4 文字盤

Figure 4 Character Dial



図5 文字入力

Figure 5 Character Input

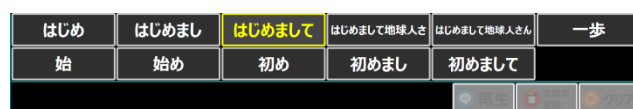


図6 変換候補

Figure 6 Conversion Candidates



図7 音声再生

Figure 7 Audio Playback

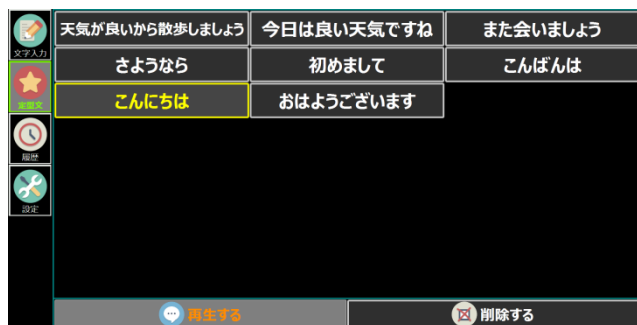


図 8 定型文機能

Figure 8 Fixed Phrase Function

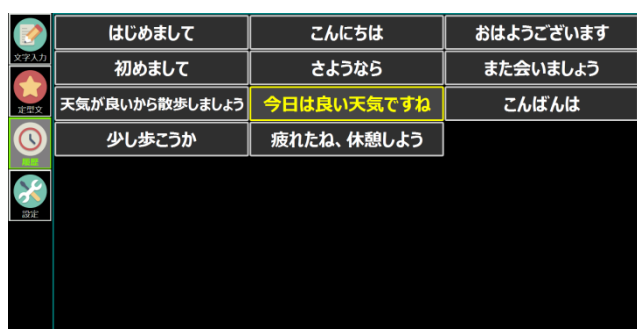


図 9 履歴機能

Figure 9 History Phrase Function



図 10 設定機能

Figure 10 Configuration Function

表 1 本システムの音声パラメータ

Table 1 Speech Parameters of the System

パラメータ	説明
API キー	無料利用登録した際に発行される、Web API を利用するためのキー
話者	声質のベースとなる話者であり、男声ではたける・しょう・サンタ・くまの 4 つ、女声ではひかり・はるかの 2 つから選択する
音量	音声再生される音量を 50 から 200 の間で選択でき、基本は 100 となる
速度	音声再生される速度を 50 から 400 の間で選択でき、基本は 100 となる
音程	再生される音声の高低を 50 から 200 の間で選択でき、基本は 100 となる



図 11 感情手動選択

Figure 11 Manually Select the Emotion



図 12 メッセージ手動入力

Figure 12 Manually Input the Message

最後に、設定機能について説明する。図 10 のような画面構成となっており、本システムで再生される音声の声質設定、および操作方法設定の 2 つに分けて説明する。初めに、本システムで再生される音声の声質設定について、表 1 に示す。これは感情音声合成サービスである VoiceText Web API のマニュアル[a]を参考にして、調整可能パラメータとして実装したものとなる。調整方法として、例えば 50 代女性の声質に近付けたい場合、ベースとなる話者にひかりを選択し、速度と音程を小さくすることで、期待する声質に近付けることができた。続いて、2 つの操作方法設定について説明する。感情手動選択は、病気などの症状で表情を顔に表現することができない、または困難な人があることを想定して実装した。この機能を有効にすることで、表情から感情を推定せずとも音声に付加する感情を選択でき、感情を音声で表現できる。メッセージ手動入力については、手を動かしてキーボード入力を行なうことで、文章を作成できる人のためのオプションとして実装した。

a) <https://cloud.voicetext.jp/webapi/docs/api>



図 13 システム利用の様子
Figure 13 System Overview

4. 有効性検証

本システムの有効性を検証するにあたり、まずは健常者に図 13 のように使用してもらい、視線操作の使用性、表情から推定された感情と実際の感情との合致度合、および感情音声合成を聞いたときの聴者の感情の理解度についてインタビューを行なった。

4.1 視線操作の使用性

まず、その人に合ったキャリブレーションのデータでないと、適切な操作感が得にくいことが分かった。当初、事前にキャリブレーションを行なったデータを使用してシステムを使用してもらったが、人によって頭部の位置や PC までの距離が異なり、精度が乱れたためである。次に、本システムで実装した視線操作は、左目のみを閉じることでボタンクリックを実現するが、その動作ができない人がいることが分かった。左目を閉じると同時に、右目も閉じてしまうため、ボタンをクリックできなかったのである。右目であれば閉じることができる、という意見もあり、ボタンクリックの実装については課題が残った。

4.2 表情から推定された感情と実際の感情との合致度合

概ね、表情から感情を推定できていたようである。それから、人によって推定されやすい感情があるということが分かった。顔の形や各部位の位置などによって、推定されやすい感情があるのではないかと考えられる。

4.3 感情音声合成を聞いたときの聴者の感情の理解度

本システムで実装した感情音声合成は、喜び、怒り、悲しみに加えて、感情情報を付加しない無感情の 4 つを選択する。その中で、最も理解度が高かったものは、悲しみとなった。また、喜びと怒りについては、再生する文章にもよるが、相互に間違って認識されることが多かった。これは、悲しみが明確に暗い音声であることに対し、喜びと怒りは共に強く勢いがある音声となるためであると考えられる。

5. まとめ

本研究では、視線検知技術、表情検知技術、感情音声合成技術を組み合わせて、音声のみで感情を伝える手段を提供するコミュニケーション支援システムを提案した。有効性の検証について、開発した評価用プロトタイプを用いて、まずは健常者を対象に行なった。今後は、本研究が対象とする人に実際に使用してもらい、有効であることを確認したい。

謝辞 本研究は、神戸情報大学院大学における研究室共同プロジェクトの一部として実施されたものであり、研究を推進するにあたり、貴重なご意見やご指摘を頂いたプロジェクトメンバー、ご指導を頂きました先生方に心より感謝申し上げます。

参考文献

- [1] 東京工芸大学, “ナチュラルユーザーインターフェースに関する調査,” 東京工芸大学, 2012.
- [2] 皆川明洋, 小田切淳一, 堀田悦伸, “複数のセンシング技術を活用した非接触ユーザーインターフェース,” 富士通, 2013.
- [3] 中野泰彦, 宮川あゆ, 佐野聡, “ドライバの覚醒度検知技術,” 富士通, 2008.
- [4] 飯田朱美, N.CAMPBELL, 安村通晃, “感情表現が可能な合成音声の作成と評価,” 情報処理学会, 1999.
- [5] 飯田朱美, S.DESIRAZU, 伊賀聡一郎, N.CAMPBELL, 安村通晃, “感情表現が可能な音声合成を用いたコミュニケーションエイドの試作,” 電気情報通信学会, 1999.
- [6] 飯田朱美, 伊賀聡一郎, 樋口文人, N.CAMPBELL, 安村通晃, “対話支援のための感情音声合成システムの試作と評価,” ヒューマンインタフェース学会, 2000.
- [7] 安村通晃, “コミュニケーションとインタラクション,” 慶應 KFC ジャーナル, 2003.
- [8] アウンコンサルティング株式会社, “2014_世界 40 カ国の地域を検索エンジンシェアと推移【2014 年 6 月版】,” アウンコンサルティング株式会社, 2014.