# 声質ヴィジュアライザの提案

松浦　泰仁[†]　角谷　亮祐[†]　秋庭　祐貴[†]
菅沼　佑太[†]　菊川　裕也[†]　馬場　哲晃[†]
串山　久美子[†]

声の強弱、高低、速度などの要素はコミュニケーションにおいて感情や雰囲気を伝達させるために不可欠なものである。本研究では、発話におけるこのような特性をフォントの特性に基づいて視覚化できる音声認識・文章作成アプリケーションのシステムと効果的なグラフィックユーザーインターフェイスの在り方について提案と議論を行う。

# A proposal of software that visualize vocal feature

YASUHITO MATSUURA[†]　RYOSUKE KAKUTANI[†]　YUKI AKIBA[†]
YUTA SUGANUMA[†]　YUYA KIKUKAWA[†]　TETSUAKI BABA[†]
KUMIKO KUSHIYAMA[†]

Vocal strength, pitch and tempo is necessary in communication to derive emotion and situation. This paper proposes word-processing and voice recognition application that can visualize these characteristics according to font features and effective GUI for the application.

## 1. 特徴

本稿では、文書作成アプリケーションにおいてパラ言語の伝達、すなわちイントネーションやリズムなどの言語の周辺的側面を表現する手法として、人の声質を文字に反映させるアプリケーションを提案する。音声認識技術は単に音声を文字に起こす事のみに留まらず、近年では Google 社の「Google 音声検索」や Apple 社の iPhone4S の音声ガイダンス機能「Siri」、東芝のスマートフォン向け音声翻訳アプリケーション「ポケット通訳」のような応用方法が盛んに研究開発されている。このような事例の殆どが、身近なマイクとして存在するスマートフォンの普及に起因したものであり、これに伴って音声認識アプリケーションもますます注目を浴びるであろうと推測する。しかしながら、従来の音声認識アプリケーションは読み込んだ音声の文字列を「情報」として扱うものであるか、またはいかに正確かつ即座に文字列を表示させるかという精度の向上に重点を置いたものが多い。

本研究では、音声そのものの特徴と表示される文字との相互関係に着目している。具体的には、発した声の大小・高低・速度といった「音響的特徴」によって、入力される文字のフォントの種類・ウェイト・級数といった「視覚的特徴」を操作することができる。これはフォントがただ無作為に変化するものではなく、声とフォントの特徴が共通認識的に結び付けられるようなものでなければならない。

声とフォントとが巧妙に一致する例として、漫画や一部の小説の「台詞」がある。登場人物が声を張り上げる場面では太く大きなゴシック体が使われ、貧弱な声では細く小さなフォントが使われる。(図 1)あるいは、幽霊の声を演出したい場合は震えが掛かったようなフォントが選択され、華麗な人物の声には相応のフォントが選択されることが望ましい。本研究はこのような情景的な文章を感覚的に作成することができる

**図1** 台詞におけるフォント使い[1)]

† 首都大学東京システムデザイン学部
Faculty of System Design, Tokyo Metropolitan University

## 2. システム実装

### 2.1 手法

音声の様々な要素を抽出するため、本研究の第一段階として語勢に着目する。語勢とは話す時と書く時の言葉の勢いであり、汎用性の高いデジタルテキスト入力においては表現が困難で、鉤括弧や引用符を用いた強調をわざわざ付加的に入力しなければならない。著者らは、語勢によって文字の大小が変化することでこの煩わしさを解消できるだろうと考える。

以下はシステム実装の手法とそのシステム概要図である。

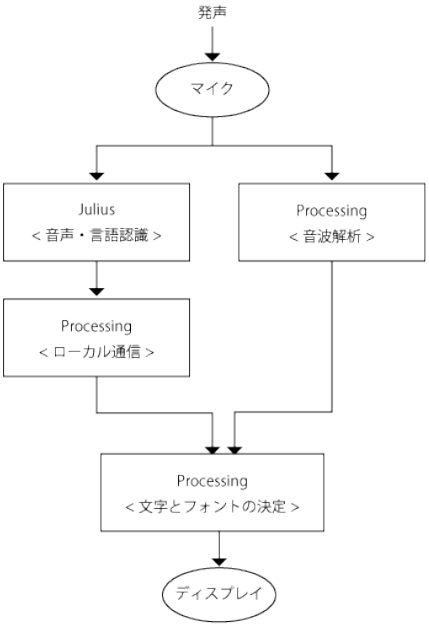


**図2** システム概略

使用環境は Apple 社の MacBookAir 及び内蔵マイク、OS は Mac OS X 10.6.8、言語開発環境は Xcode3.2.6、日本語音声認識エンジンには Julius 4.1.5[2)]、音響検出及び文字の画面出力には Processing4.1.2 を使用している。

### 2.2 シーケンスの解説

① Julius をモジュールモードで起動させ、音声認識サーバとして使用。表示情報を Processing 上に移行させる。

② Processing でマイクから検知した音波を FFT にかけ、文字のサイズを決定する。

③ ①と②を合わせた認識結果を単語毎に Processing 画面上に表示させる。

ここではシステムの検証のため、行程②において音波の振幅のみに限定しているが、完全版では同行程で波長から声の高さを判定し、フォントの種類を変化させるつもりである。

### 2.3 音声認識精度の考察

Julius をモジュールモードで起動し TCP/IP 経由でクライアント(Processing)と接続する場合と通常モード（ターミナル上で起動させる）の場合とで発声から認識結果表示までのタイムラグを感覚的に比較する。前者は、発声者が発話を終了して約 1 秒後に認識結果が表示されたが、後者では発話終了とほぼ同時に表示されているため、モジュールモード起動は表示タイムラグを増大させる要因になっていると判断した。これについては今後改良する必要がある。

日本語認識の精度に関しては Julius 自体に準拠している。Julius の最大の特徴は汎用性であり、デフォルトでも約 6 万語の言語モデルをカバーするが、文法・単語辞書のカスタマイズによって様々な用途に特化することも可能である。すなわち、ある文字誘導に基づいて発話される環境下（音声ガイダンスなど）や、専門用語、方言などを想定して対応することができる。また、高齢者や子供向けなどに応じて音響モデルを変更することで精度を向上させることも有効である。

発声環境については、デフォルトの設定の場合、ある程度静かな環境、マイクから 20~30cm 程度の距離、機械を意識した明瞭な喋り方であれば、登録された言語モデルに則った文章を作成することができる。ここでの「明瞭な喋り方」とは、例えば「こ、ん、に、ち、は」の様に文字間の区切りを意識するのではなく、アナウンサーのような自然な発話様式のことを指す。しかし本研究においてアナウンサーのような発声をユーザに強要してしまうと、自由な喋り方によるアクセントの現れが無くなると推察し、音響モデルに関して調整しなければならない。

### 2.4 マイク音量と文字サイズの対応

マイク音量と文字サイズの対応については、以下の図 3 の通りに行う。。最小レベルの文字サイズ a は可視性を考慮して 4.0 ポイント以上を最低条件とする。

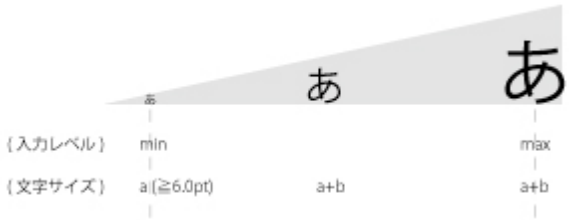


**図3** 入力レベルと文字サイズの対応

### 2.5　声質とフォントの対応

本章では、声の高さをフォントに影響させる手法について考察する。「声の質感」とは本来、同じ声種（ソプラノ、アルト、テノールなど）における個人差について言及する言葉であるが、ここではより広義な意味として声種の違いも「声質」と呼ぶことにする。本研究の目標としては同年代同士での絶妙な声の違いまでをも識別したいと考えるが、そのための分析手法はまだ研究途中である。まずは「高い声を出しているか、それとも低い声を出しているか」を判定することに留める。

フォントの持つ心理作用について、井上 [4]らの先行研究では、文字形態の印象に対する 心理要因および物理要因が示されている。

**表1** 文字形態とその印象

| 代表的な印象語（形容詞対） | 心理要因 | 物理要因 |
|---|---|---|
| 素直な，安定した，落ち着いた，くせのない | 素直さ | 線の太さ，陰影，バランス |
| 魅力的な，粋な，楽しい，リズミカルな | 魅力感 | 線の凹凸，白抜き |
| 重い，迫力のある，男性的な，うっとうしい，ごてごてした | 力量感 | 線の太さ，隙間 |
| 新しい，明るい，開放的な，派手な，温かい | 新鮮さ | 白抜き，大きさ |
| 丸い，柔らかい | 丸 | 丸さ |

これを元に、具体的な和文フォントを主観選抜する。声の印象は合唱において代表的な声種を 4 種設定する。声種間の差を表現するため、フォントファミリーはそれぞれ異なるものを採用した。

(A)　ソプラノ
「こんにちは」: DFP 痩金体 W3

(B)　アルト
「こんにちは」: DFP 教科書体 W3

(C)　テノール
「こんにちは」: 小塚ゴシック Pro M

(D)　バス
「こんにちは」: ヒラギノ角ゴ Std N

フォントは Mac OS X 10.6.8 標準フォントに加えて Adobe Creative Suite 4.0 と Microsoft Office 2008 for Mac をインストールした状態のフォント環境から選抜し、男声は力量感の出やすいゴシック系フォント、女声は繊細さの印象から明朝系フォントを選んだ。その中の音域の違いは、不安定と感じるフォントを高音域に割り当て、その逆を低音域に割り当てることで表現している。

さらに、今後の改良により各声種内でも声質（軽い、重いなど）の区別化が可能になったと仮定した時のフォントは以下のように変化させる。

(X)　軽い、明瞭、リズミカル
「こんにちは」: DFP 太丸ゴシック体 中抜き

(Y)　落ち着き、重い、安定
「こんにちは」: ヒラギノ角ゴ Std N

(Z)　恐怖、くせのある、不安定
「こんにちは」: DFP 祥南行書体 W5

このような変化を(A)~(D)のフォントを反映しながら起こすことで、より多くの変化をユーザが楽しむことができる。

## 3. インターフェイスの議論と提案

本研究の最大の目的は、文字を音声によって変動させること自体にあるのではなく、それを用いたエンターテインメント・アプリケーションの創出である。ユーザが声質を視覚化できることを自然に知覚でき、何度もアプリケーションを使用してみたいという意欲を掻き立てるようなエンターテインメント性が必須である。さらに留意しなければならないのは、本研究の特徴から逸れたものであってはならないということである。以上の観点を念頭に置き、最も適切なアプリケーションとそのインターフェイスを考察する。

初期案として、キャラクターの台詞の吹き出しに見立てたテキストボックスに文字が出現するタイプ「mojica」を考案した。キャラクターの存在、および吹き出しに見立てた部分には台詞が表示されるという行為のアフォーダンスがあるため、ユーザが何らかの台詞を発声する事を誘発する効果はある。しかし、エンターテインメント性を考慮すると、1 体のキャラクターが微動だにせず、淡々と台詞が入力されたとしても面白みに欠け、白紙に入力される場合と比べて効果は大差がないのでは、という意見が上がった。

**図4** mojica インターフェイス

これらの指摘を踏まえ、改善案として新たに「VV Typewriter」を考案した。これは、表情が伝わりにくかった mojica のキャラクターを喜怒哀楽の変化が豊かなキャラクターに変更しており、吹き出しと文字が変化する。台詞に従ってキャラクターが変化する。

**図5** VV Typewriter インターフェイス

VV Typewriter においてはキャラクターの挙動についてエンターテインメント性があるが、喋る台詞の内容自体のエンターテインメント性を追求する手法として、漫画の台詞に着目したことに立ち返った「My-serif」を考案した。「自分の台詞（マイセリフ）を作る」という、小学生の落書き遊びのような感覚で漫画の中の台詞を自由に作成できることがコンセプトである。インターフェイスの構成としては、台詞や吹き出しが半透明となっている漫画のコマが画面に出現し、漫画を読む方向に従って入力音声が台詞になる。ある程度の文字数を超えると噴出しの大きさが拡張され、入力したい吹き出しを移行させる際はキーボード入力で行う。また、フォントの変化と同様のシステムを用い、吹き出しの形状も変化させる。入力が１ページ完了するとスクリーンショットとして保存され、比較して楽しむことができる。

単にキーボード入力で吹き出しに文字を打ち込む事と比較すると、ユーザは台詞を喋るキャラクターになりきる必要があり、コミュニケーションの種となると考えられる。また、画面から離れてユーザ同士の演技を楽しむこともできるため、遊びの拡張性もユーザによって開拓されることが期待できる。

**図6** My-serif インターフェイス

## 4. 改善と展望

本研究はまだ完全な実装段階に至っておらず、各システムの連携が不明確である。以下に問題点を提示する。

### 4.1 声種の設定

音波解析について、ソプラノ～バスの各閾値をどのように設定するかが不明瞭である。これはおそらく音楽的分類よりも統計的に算出する方が望ましいと考えられる。なぜなら、ユーザは歌唱するわけではなくほ

ぼ自然な発話をするためであり、老若男女の様々な台詞別の音波を解析した上で適切なボーダーラインを設定する方が有効なためである。

### 4.2 文字組の維持

文字の大きさと「アキ」が変化する場合、段内の文字数や行間などのレイアウトが著しく崩れる場合が存在する。現行システムでは、市販の文書作成ソフトのような自動調整機能を実現させることは困難である。吹き出しというテキストボックスの存在により、文字が重なったり不自然に空いたりしないような文字組を実現させるためのアルゴリズム設定はまだ研究する必要がある。

### 4.3 音声認識の精度

周辺環境をできるだけ無音状態に保たなければならないことはエンターテインメント性を阻害することになる。だからといって、マイクの集音範囲を狭めると認識精度を落とすジレンマが発生する。前述の通り Julius は音響モデルを調節可能で、様々な音量や喋り方に設定上対応可能であるといえるが、それを周辺環境に応じてフレキシブルに変化させることは困難である。この解決策としては、音声認識ソフトを変更するか、あるいは設定に適切な発話をユーザに促す誘導を設ける必要がある。例えば My-serif においては、「キャラクターになりきれ」といった誘導文を提示することが考えられる。

### 4.4 フォントの候補不足

本稿でのフォント環境は非常に限定的であり、それぞれの声種・声質を表現するために最も効果的であるとは言い難い。より多くのフォントを候補に含める必要がある。また、フォントを設定する際の判断基準が主観による所が大きく、効果を実証するためには統計が必要である。

### 4.5 文字変化の意義

そもそも、言葉のイントネーションを文字のサイズに置き換えることが有効かどうかについて議論の余地がある。言葉がリズミカルに見えるかもしれないが、聴覚で感じたイントネーションの印象と視覚のそれとがどれほど一致するかが定かではない。本研究の目的はパラ言語情報そのものを正確に伝えることではなく、それを手段としてコミュニケーションを促進させることであるため、重大な意図のないアクセント変化まで視覚化する必要はないと思われる。したがって、重大なアクセントがある場合の抽出手法について研究の余地がある。

単語の大きさが変化する方が通じ易いと言える。なぜなら元々パラ言語情報を除いた文章における強調である鉤括弧や引用符は、単語や文節に掛けられることが多いためである。例えば「昨日は晴れだった。」という文章において、今日は違う天気であることを強調したい場合は「昨日は」に、雨や雪ではないことを強調したい場合は「晴れ」に強調符が付く。しかし、発話において強調の方法は単語や文節のみに限定されるとは言い切れない。また強調とは逆に平坦を表現したいこともあるだろう。そのような場合、やはり文字のイントネーションの有無は必要不可欠となる。

今後の応用について、文章作成という点から離れ、よりエンターテインメント性のあるアプリケーション例として以下のようなボイスタイピングゲームを提案する。

**図7** ボイスタイピングゲーム

このゲームは上段の文章のサンプル音声と文字の大きさの変化を参考にしながら音声入力し、下段に認識結果が表示され、最終的に上段との適合度を測定して点数表示するものである。サンプルの文章は楽曲の歌詞などにするとエンターテインメント性が確保できると考えられる。しかし、その場合はカラオケでの採点機内蔵のゲームと大差が無い恐れがある。

エンターテインメント性ではなく社会的意義のあるものとして、聴覚障害者の発話訓練ツールとしての応用を提案する。そのためには、以下の要件が付加される。

- 誤った発声を明確に提示し、正しい発音のための誘導を、口腔図などを用いて行う。
- 単語をより強く対応付けするために、正しい発音ができるとイラストが表示されるシステムを設ける。

しかし聴覚障害者の発話訓練は、幼児期から長期間に渡り専門家によって行われたとしても確実な効果が得られるものではないほど困難を極めるものであり、

容易には着手し難い事項である。

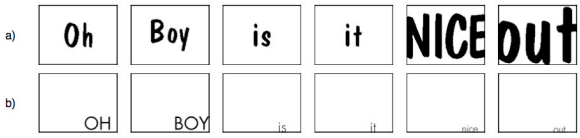


**図8** 同じ文章で異なる感情を表現[7]

## 5. 関連研究

本研究以前にも、音声認識とその視覚化についての研究、あるいは文字の特性についての研究は数多く行われている。

神原 [5]らは、ペン入力デバイスのブラシをオノマトペの音声入力によって変更するインタラクション手法を提案している。オノマトペに着目することで、従来のペイントソフトよりも感覚的に対応付けができること、ストロークの情報とブラシの切り替え情報を同時に送信できること、エンターテインメント性があることが特徴となっている。この研究においても音声認識エンジンとして Julius(Julian)が採用されており、ブラシストロークというリアルタイム性が重視される状況においても問題無く作動している。応用例において、オノマトペと関連した事象の入力であれば様々な用途に適用できることが考えられており、本研究においてもオノマトペに着目した使用シナリオの設定によって新たな活路が生まれるのではないかと期待する。

水口 [6]らは文字アニメーションの自動合成エンジン作成のため、文字アニメーションが持つ特性や表現のパターンに言及している。その中で Lee[7]らの同じ文章で異なる感情を表現する例（図 7）を紹介し、その例に沿った形でフォントのもつパラ言語的印象、文字の拡大・拡張、ある単語の実世界動作の模倣（snow という単語では、パラパラと上から下に降る粉雪や横に速く吹きつけるような吹雪などの違い）、漫画的な誇張表現、オノマトペに基づいた表現などを分析している。本研究と比べると、音声認識の研究ではない点とアニメーションという点において異なっているものの、言語印象の拡張についてきめ細やかな分析が為されており、本研究を今後一段と進めるものとなるだろう。

野原 [8]らは、自動会話システムにおける感情表現の手法について、フォントとキャラクターの効果を採用している。本研究の mojica と類似した効果を狙っているが、本研究はユーザの発声におけるパラ言語情報の表現に主眼がある。しかし、本稿では文章作成からのフィードバックが弱く、一方で会話という状況設定はインタラクティブであり、パラ言語情報を表現する目的としても有効である。

## 参 考 文 献

1) ハチワンダイバー: 柴田ヨクサル, ヤングジャンプコミックス
2) Julius: 京都大学 河原研究室, 名古屋工業大学 Julius 開発チーム, 情報処理振興事業協会(IPA), 奈良先端科学技術大学院大学 鹿野研究室.
3) http://julius.sourceforge.jp/
4) 井上,鎧沢,“文字形態から受ける印象と品質評価要因の検討 ”,電子情報通信学会論文誌,Vol.J67-B,No.3, pp.328-335,1984.
5) 神原 啓介, 塚田 浩二. オノマトペン. 2010
6) 水口 充, 田中 克己.文字アニメーションの自動合成の試み. 2005
7) Lee, J., Forlizzi, J., and Hudson, S. The kinetic typography engine: an extensible system for animationg expressive text. 2002
8) 野原 雄紀, 徳永 幸生, 杉山 精, 米村 俊一, 大谷 淳. 自動会話システムにおけるフォントによる感情表現の効果. 2009