新技術説明会
New Technology Presentation Meetings!

顔画像処理を用いた音声を利用しない音声認識技術（読唇技術）の改善

九州工業大学
大学院情報工学研究院
准教授　齊藤 剛史

新技術説明会（2019年12月10日）

サイレント音声認識

- 音声情報を用いずに発話内容を認識すること
- Silent Speech Recognition（SSR）

- アプローチ
  - 磁気センサ
  - 画像（ビデオカメラ、距離センサ、サーモグラフィ、超音波センサ）　←　読唇
  - 非可聴つぶやきマイク
  - 振動センサ
  - 表面筋電
  - 脳波（非侵襲型、侵襲型）
  - など

SSRにおけるアプローチの比較

アプローチ	無音声	騒音環境下	喉頭摘出	非侵襲	費用
音声	×	×	×	○	○
磁気センサ	○	○	○	△	△
画像	○	○	○	○	○
非可聴つぶやきマイク（NAM）	△	○	△	○	○
振動センサ	×	○	×	○	△
表面筋電（EMG）	○	○	○	○	△
脳波（EEG）	○	○	○	△	△
脳波（ECoG）	○	○	○	×	×

読唇技術とは？

- 音声情報を用いず、口唇の動き（視覚情報）を用いて発話内容を読み取る技術

- 応用
  - 聴覚・発話障害者のコミュニケーション支援
    - 口頭摘出者
    - 手術後に一時的に発声できない患者
  - 騒音環境下における音声認識
  - 公共環境における無音声の復元
  - 複数人同時発話の解析
  - 音声無しアーカイブスからの発話内容の復元
  - バイオメトリクス認証

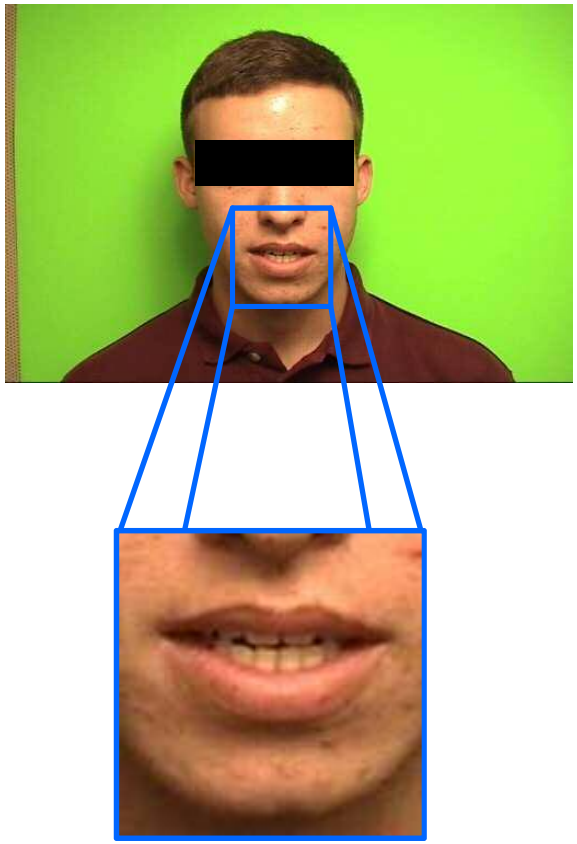
読唇技術とは？

- 画像処理技術を用いた工学的な研究は1980年代～
  - **Structural Analysis of Lip-contours for Isolated Spoken Vowels using Fourier Descriptors**
    - N. L. Hesselmann,
    - Speech Communication, 1983.
  - **Characteristics of the Mouth Shape in the Production of Japanese – Stroboscopic Observation**
    - Yumiko Fukuda（福田 友美）, Shizuo Hiki（比企 静雄）　← 東北大
    - Journal of the Acoustical Society of Japan, 1982.
  - **画像処理による読唇の試み**
    – 母音口形の識別およびそれに基づく単語認識 –
    - 松岡 清利, 古谷 忠義, 黒須 顕二　← 九工大
    - 計測自動制御学会論文集, 1986.
  - **画像解析による日本語母音の識別**
    - 内村 圭一, 道田 純治, 都甲 昌美, 相田 貞蔵　← 熊本大
    - 電子情報通信学会論文誌D, 1988.

従来技術とその問題点

- 音声情報を用いず、口唇の動き（視覚情報）を用いて発話内容を読み取る技術

- 従来読唇技術の特徴量の種類
  - 画像（Image-based）
  - モーション（Motion-based）
  - 幾何特徴（Geometric-feature-based）
  - モデル（Model-based）

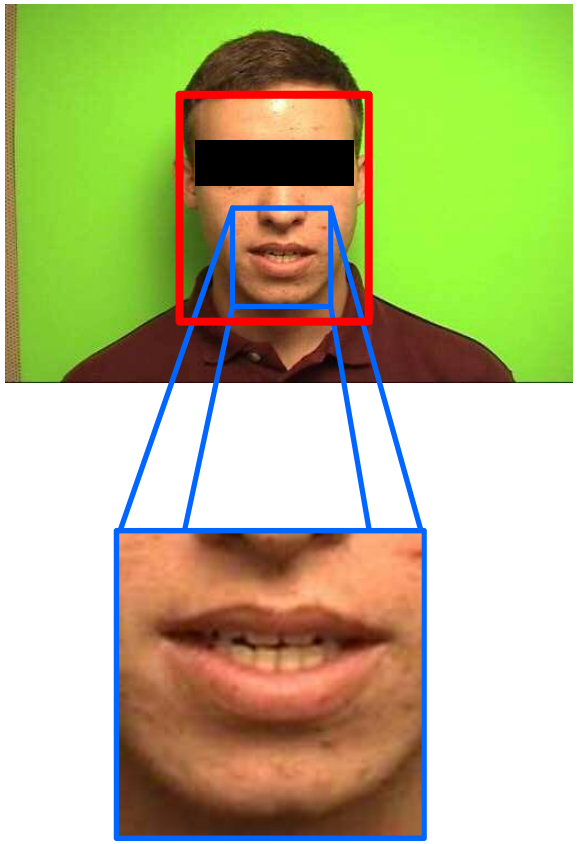


口唇周辺のみを考慮

新技術の概要

- 音声情報を用いず、口唇の動き（視覚情報）を用いて発話内容を読み取る技術

- 読唇技術の特徴量の種類
  - 画像（Image-based）
  - モーション（Motion-based）
  - 幾何特徴（Geometric-feature-based）
  - モデル（Model-based）



口唇以外（顔全体）から特徴を求める
⇒　属性情報と表情特徴を利用

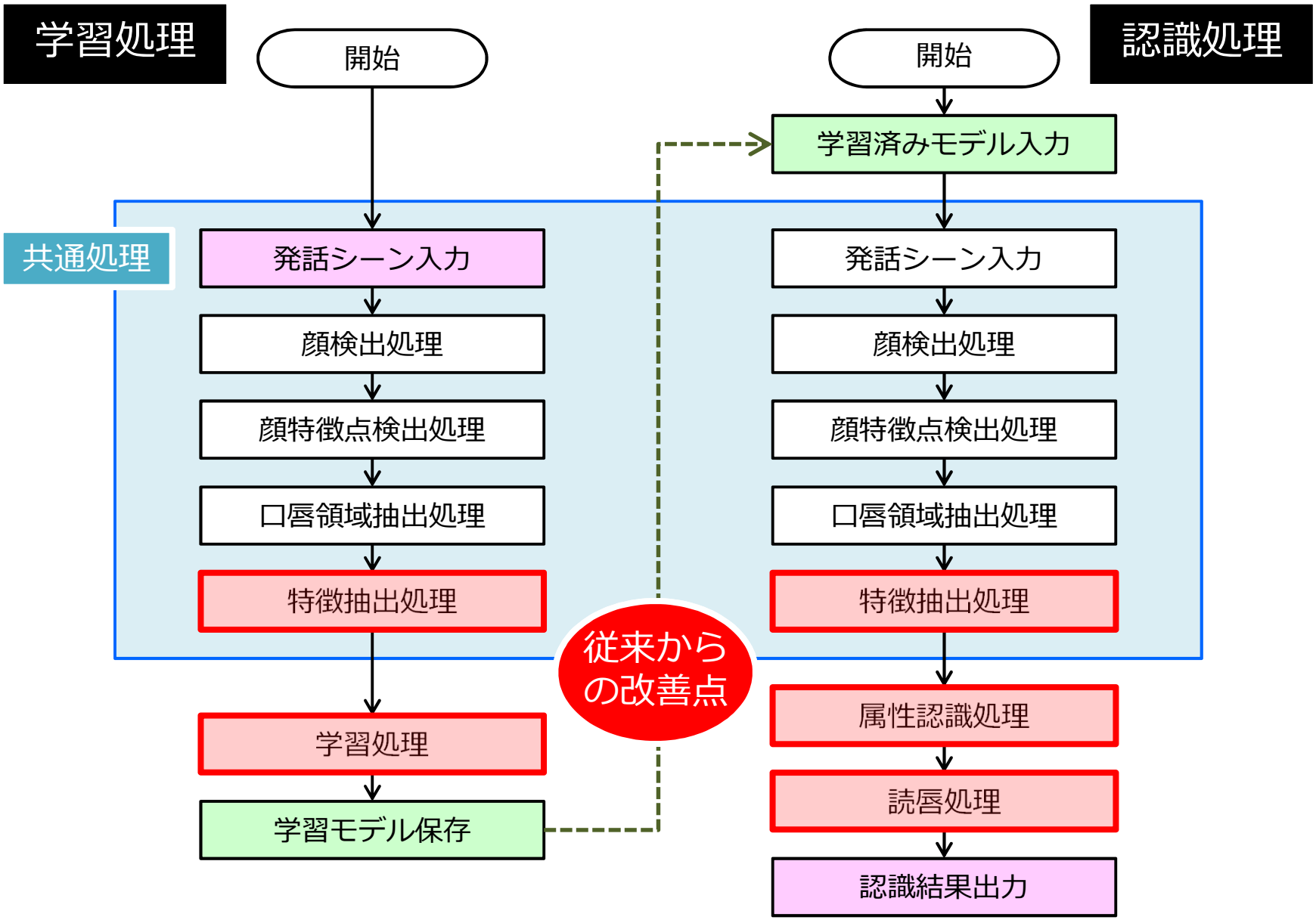
新技術の特徴

- 発話者の属性情報や表情特徴の導入
  - 属性情報：性別や年齢など
  - 表情特徴：「怒り」、「嫌悪」、「恐れ」、「幸福」、「悲しみ」、「驚き」など

- 認識精度の向上

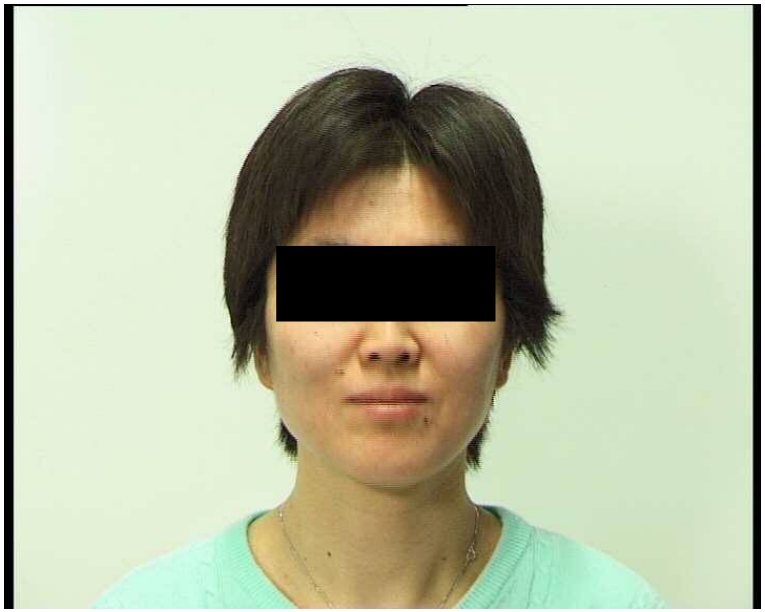
- 障害者のコミュニケーション支援の実現
- 音声情報不要の雑多な騒音環境や音声が収録できない環境における発話内容認識の実現

処理の流れ


学習処理
開始
共通処理
発話シーン入力
顔検出処理
顔特徴点検出処理
口唇領域抽出処理
特徴抽出処理
学習処理
学習モデル保存
従来からの改善点
認識処理
開始
学習済みモデル入力
発話シーン入力
顔検出処理
顔特徴点検出処理
口唇領域抽出処理
特徴抽出処理
属性認識処理
読唇処理
認識結果出力

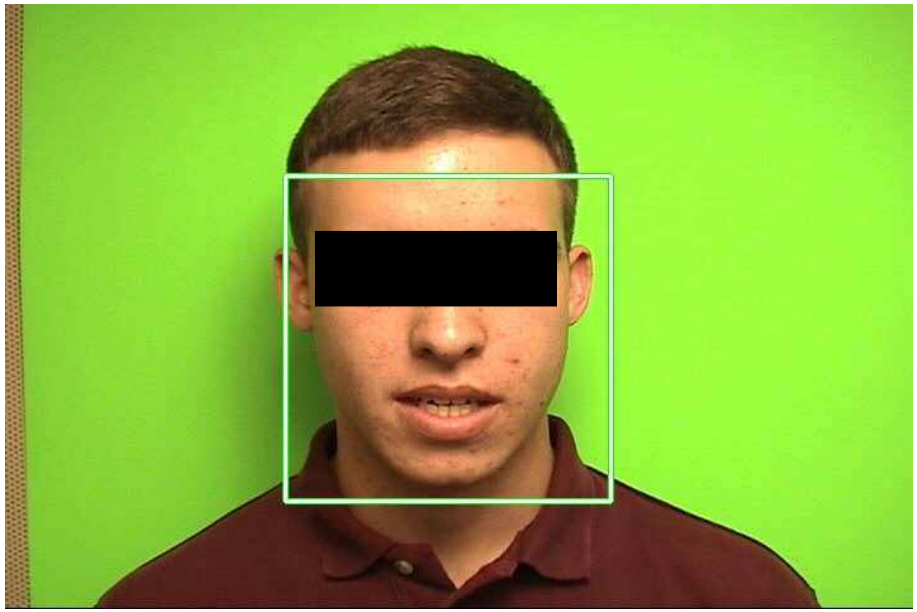
発話シーン入力

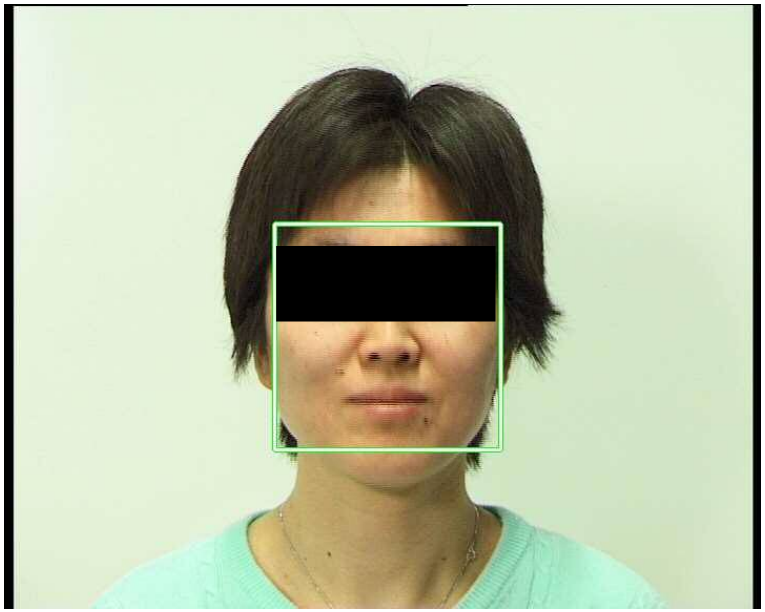
- 条件：顔画像（顔全体が写る）
  - 背景は一様でなくてもよい。



顔検出処理

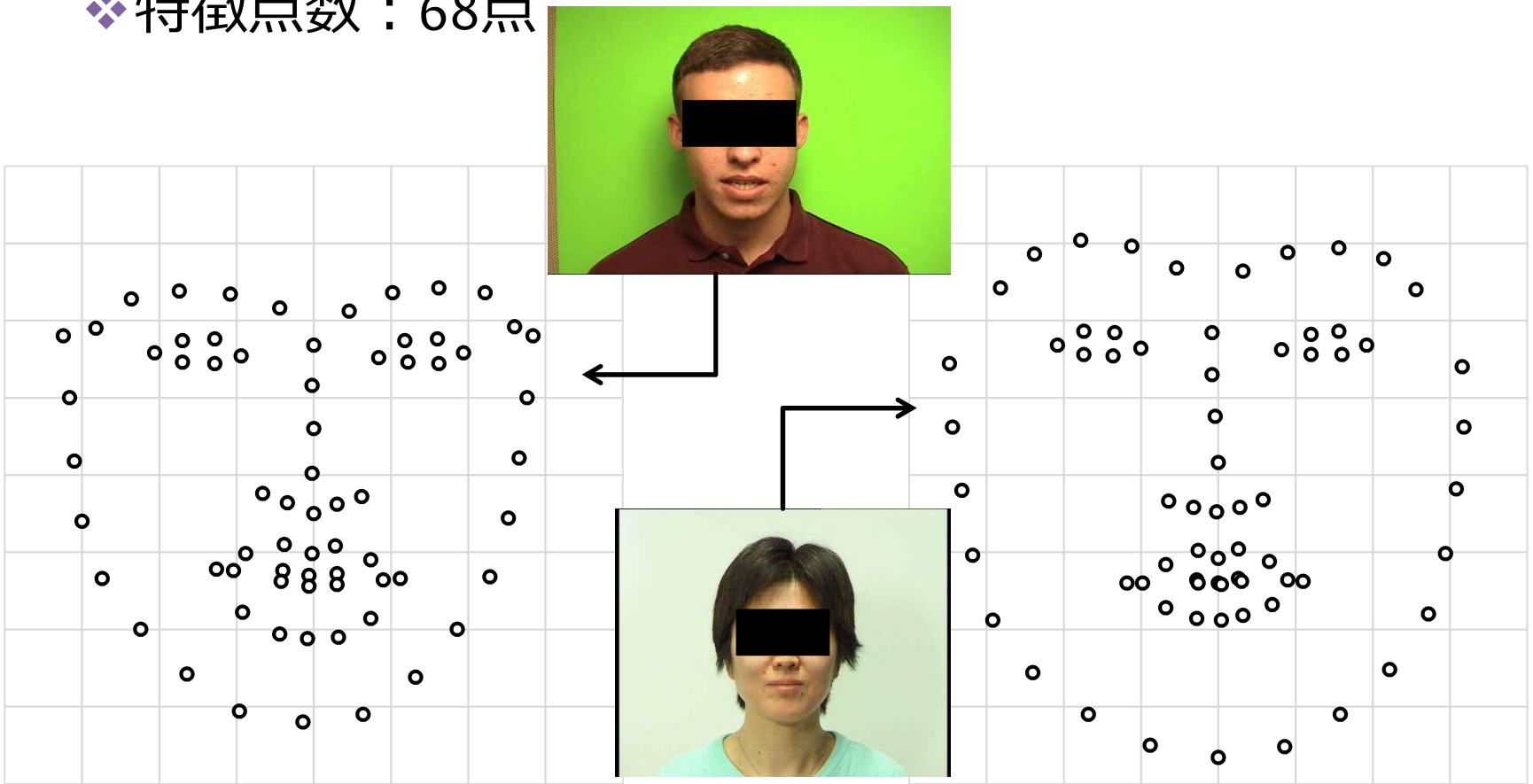
- 既存手法の利用
  - 例えば、OpenCV の detectMultiScale 関数





顔特徴点検出処理

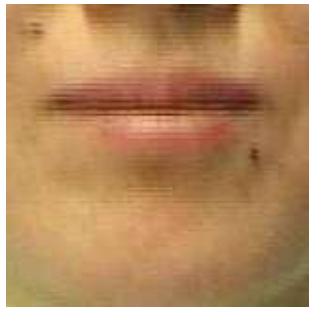
- 既存手法の利用
  - 例えば、dlib の face landmark detector
  - 特徴点数：68点



口唇領域抽出処理

- 従来の読唇手法
  - ❖口唇周辺の情報を利用するため口唇領域ROIを抽出



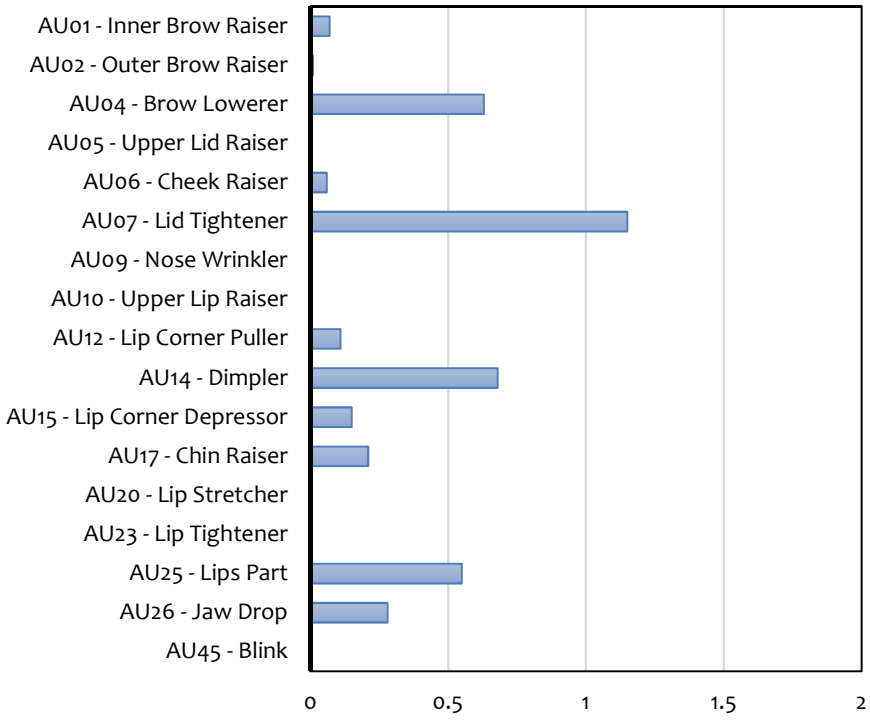



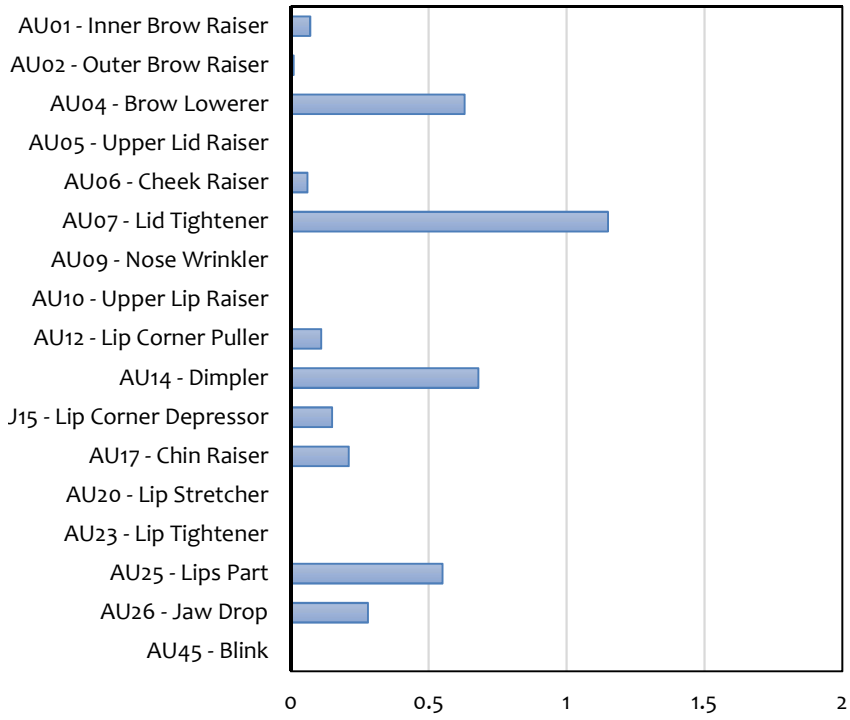
特徴抽出処理

- 従来の読唇技術
  - 口唇画像を利用した特徴量
  - 特徴点を利用した特徴量
- 新技術
  - 表情特徴

表情特徴例

- 表情特徴数：17個





実験に使用したデータベース

- 全て研究用として公開されている。

データベース名	言語	クラス数	発話内容	話者数（男女数）
CUAVE	英語	10	0～9の10数字	36名（19M+17F）
OuluVS	英語	10	10挨拶文	20名（17M+3F）
CENSREC-1-AV	日本語	10	0～9の10数字	学習用42名（22M+20F） 評価用51名（25M+26F）
SSSD	日本語	25	0～9の10数字 +15挨拶文	72名（38M+34F）

実験に使用したデータベース

- CUAVE（Clemson University Audio-Visual Experiments）
  - **Moving-Talker, Speaker-Independent Feature Study, and Baseline Results Using the CUAVE Multimodal Speech Corpus**
    - Eric K. Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N. Gowdy
    - クレムゾン大学（Clemson University）（USA）
    - EURASIP Journal on Applied Signal Processing, 2002
  - 現在は入手不可？



実験に使用したデータベース

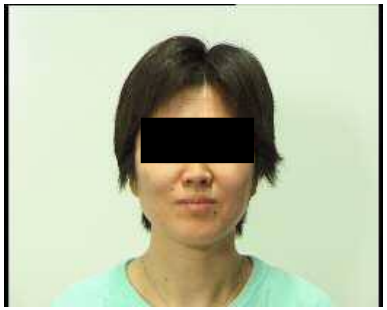
- CUAVE（Clemson University Audio-Visual Experiments）
  - 発話内容：10数字（En）
    - “zero”, “one”, “two”, … , “nine”
  - 画像サイズ：720x480[pixel]
  - フレームレート：29.97 fps
  - 話者数：36（19M+17F）



実験に使用したデータベース

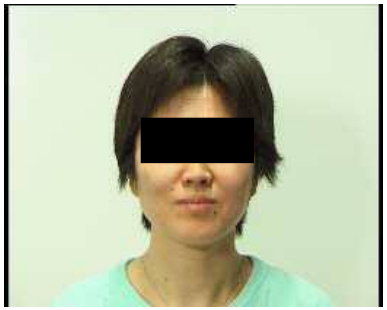
- OuluVS
  - **Lipreading with local spatiotemporal descriptors**
    - G. Zhao, M. Barnard, and M. Pietikainen
    - オウル大学（University of Oulu） (Finland)
    - IEEE Transactions on Multimedia, 2009.
  - メールで問い合わせ

    http://www.cse.oulu.fi/CMV/Downloads/OuluVS


実験に使用したデータベース

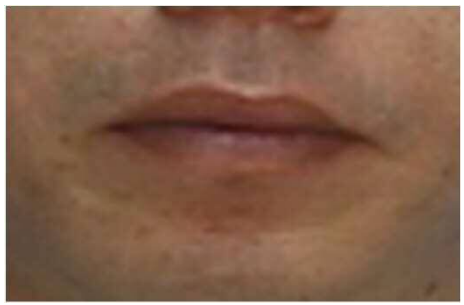
- OuluVS
  - 発話内容：10文（En）
    - “Hello”, “Excuse me”, “I am sorry”, … , “Have a good time”
  - 画像サイズ：720x576[pixel]
  - フレームレート：25 fps
  - 話者数：20（17M+3F）


実験に使用したデータベース

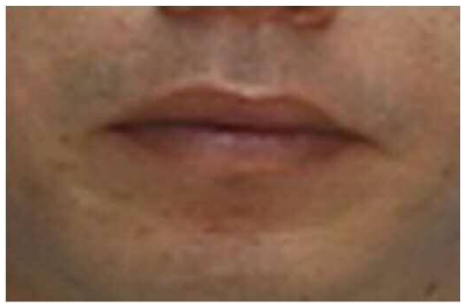
- CENSREC-1-AV（マルチモーダル音声認識評価環境データベース）
  - **音声・画像のモダリティ間の相互作用に着目した音声認識のモデル適応**
    - 大西正真, 田村哲嗣, 速水悟,
    - 電子情報通信学会 技術研究報告, 2011
  - 音声資源コンソーシアム

    http://research.nii.ac.jp/src/CENSREC-1-AV.html



実験に使用したデータベース

- CENSREC-1-AV（マルチモーダル音声認識評価環境データベース）
  - 発話内容：連続数字1～7桁（Ja）
  - 画像サイズ：81x55[pixel]
    - 口唇領域画像
  - フレームレート：29.97 fps
  - 話者数：93（47M+46F）



実験に使用したデータベース

- SSSD（Speech Scene database by Smart Device）
  - **SSSD：スマートデバイスを用いた読唇技術向け日本語データベース**
    - 齊藤 剛史, 窪川 美智子
    - 電子情報通信学会 技術研究報告, 2018.
  - 本研究室

    http://www.slab.ces.kyutech.ac.jp/SSSD/index_ja.html



実験に使用したデータベース

- SSSD（Speech Scene database by Smart Device）
  - 発話内容：25単語（Ja）
    - 10数字、15あいさつ文
  - 画像サイズ：300x300[pixel]
    - 顔下半分画像（LF-ROI）
  - フレームレート：29.97 fps
  - 話者数：72（38M+34F）



実験に使用したデータベース

- 実験に用いたデータ数

データベース名	実験手法	学習データ数	評価データ数
CUAVE	Leave-one-out cross-validation	1,750サンプル（35名×10単語×5サンプル）	50サンプル（1名×10単語×5サンプル）
OuluVS	Leave-one-out cross-validation	950サンプル（19名×10単語×5サンプル）	50サンプル（1名×10単語×5サンプル）
CENSREC-1-AV	Hold-out	840サンプル（42名×10単語×2サンプル）	510サンプル（51名×10単語×1サンプル）
SSSD	Hold-out	12,000サンプル（16名×25単語×30サンプル）	6,000サンプル（8名×25単語×40サンプル）

属性認識実験結果（性別）

- 条件
  - データベース：SSSDのみ
  - 今回は年齢は利用せず、性別のみを利用

学習データ	評価データ	認識率
16名（16M）	8名（8M）	**72.8%**
16名（16F）		68.9%
16名（8M+8F）		70.9%
16名（16M）	8名（8M）	72.6%
16名（16F）		**73.1%**
16名（8M+8F）		72.9%

男性データを認識する場合、男性データのみで学習したモデルの精度が最も高い。

女性データを認識する場合、女性データのみで学習したモデルの精度が最も高い。

読唇実験結果

- 条件
  - 男女を分けず男女混合データで学習している。
  - AU-only：表情特徴のみを利用
  - Lip-only：口唇特徴のみを利用（従来手法）
  - Lip+AU：口唇特徴と表情特徴を利用（提案手法）

Database	発話内容	AU-only	Lip-only	Lip+AU
CUAVE	10数字	71.2%	79.9%	**80.6%**
OuluVS	10挨拶文	69.8%	83.1%	**86.6%**
CENSREC-1-AV	10数字	59.6%	74.3%	**77.1%**

口唇特徴と表情特徴を利用する場合の精度が最も高い。

新技術の特徴・従来技術との比較

- 新技術は従来技術に比べて、
- 利点
  - ❖認識精度の向上
- 欠点
  - ❖処理量の増加

想定される用途

- 携帯電話・無線機等の音声通信分野
- 医療・福祉現場
- 玩具・ゲーム等のアミューズメント分野

- 障害者・高齢者のコミュニケーション支援
  - 生活の質（QOL）を高める支援システム
  - 就労促進

- 音声認識用インタフェースの代用
  - カーナビ、スマートフォンなど
  - 騒音環境下における音声認識の補助
  - 公共の場所における音声通話の実現

想定される業界

- 想定されるユーザ
  - 聴覚・発話障害者
  - 高齢者
  - 医療・福祉従事者
  - 全ての人（音声認識と同じユーザ）

- 想定される業界
  - 福祉分野
  - インタフェース（携帯端末、カーナビ）に関する分野

実用化に向けた課題

- 認識精度の向上
- 認識対象の改善
  - 孤立単語認識　⇒　連続文章認識
- データベースの充実

企業への期待

- 発話シーン収集の協力
- 障害者に対する有用性の評価
  - 医療・福祉分野に関連する企業の協力を期待

本技術に関する知的財産権

- 発明の名称　　：読唇装置及び読唇方法
- 出願番号　　：特願2019-213234
- 出願人　　：九州工業大学
- 発明者　　：齊藤剛史

お問い合わせ先

- 国立大学法人九州工業大学
- イノベーション推進機構　グローバル産学連携センター　知的財産部門　石田精
- TEL：093-884-3499
- FAX：093-884-3531
- E-mail：chizai@jimu.kyutech.ac.jp