お母さんスイッチ

～態度とヒエラルキーを理解する音声対話システム～

2025年11月25日 大森唯詩・高橋蒼生・田中悠飛

目次

概要説明

我々の暮らしを便利に、豊かにするスマートスピーカー

・スマートスピーカーの普及により、それを使用することが次世代の子供たちにとって当たり前に
・彼らにとって「話せばなんでも応えてくれる存在」が日常になりつつある
・便利なだけでなく、問題を引き起こす可能性も

・従順なだけのスマートスピーカーではなく、「まるで母親のような許可/拒否/叱責」を生成する
・話者の権限(親/子)、話し方(丁寧/乱暴)を識別し、それに基づいて応答を変える音声対話AIを実現

システム構成

スイッチ用文法の作成/認識

- Juliusが受理するネットワーク文法を作成
- スマートスピーカーの典型的な文章
- 丁寧/乱暴な文章

話者識別

- 実験前半で作成したような話者識別モデルを利用し、話者をお母さんと子供に区別

話者に応じた応答分岐

- 受理した文章の丁寧度、話者のラベルに応じて返答を分岐

スイッチ用文法の作成/認識

- スマートスピーカーを利用する際によく使われるであろう文章を登録
- 後の態度判定を行うために、丁寧な文章と乱暴な文章を共に作成
- 認識した指令(TV_ON/VOLUME_UP)や態度(polite/rude)をjson形式で保存
- 登録した文法/コマンド：
  テレビ,電気,音楽のON/OFF
  アラームの設定
  カーテンの開閉
  おやつを与える
  暴言(うるさい/黙れ)
  感謝(ありがとう)

```
% NOUN_SWITCH
テレビ  てれび
電気　でんき
% NOUN_GIVE
おやつ  おやつ
% NOUN_ALARM
アラーム　　あらーむ
% NOUN_MUSIC
音楽　おんがく
% NOUN_VOLUME
音量　おんりょー
% NOUN_CURTAIN
カーテン　　かーてん
% VERB_SWITCH
つけて  つけて
消して  けして
つけろ  つけろ
けせ　けせ
% VERB_GIVE
ちょうだい　ちょーだい
ください　　ください
くれ　くれ
% VERB_ALARM
かけて  かけて
消して  けして
かけろ  かけろ
```

```
wav_path:    "logs/temp_ad10639ea9f84bada5c1401fd1560d76.wav"
raw_words:
  0:         "<s>"
  1:         "電気"
  2:         "つけて"
  3:         "<s>"
command:     "LIGHT_ON"
attitude:    "polite"
```

話者識別

- 授業での方法
  - ・GMM：音声特徴の一般的な分布を表現
  - ・各話者の音声データで成分を微調整→話者固有のモデル
  - ・supervectorを作成し、話者推定

- 今回

・最新の深層学習モデルECAPA-TDNNを使用

・→話者認識に特化

・深層学習に用いられた音声のサンプル数、環境の種類が膨大

・→高精度・雑音や反響に強い

◎

```
判定： child
確信度： {'parent': 0.3615270895, 'child': 0.6384729105}
```

```python
# ECAPA-TDNNモデルロード
classifier = EncoderClassifier.from_hparams(
    source="speechbrain/spkrec-ecapa-voxceleb",
    savedir="pretrained_models/ecapa"
)

def get_embedding(path, sr=16000):
    signal, sr = librosa.load(path, sr=sr)
    signal = signal.astype(np.float32)


    # numpy -> torch.Tensor に変換
    signal = torch.from_numpy(signal).unsqueeze(0)

    embedding = classifier.encode_batch(signal)
    return embedding.squeeze().cpu().numpy()


def enroll_speaker(wav_files):
    embs = [get_embedding(f) for f in wav_files]
    return np.mean(embs, axis=0)  # 話者の平均ベクトル

import os
import pickle
import pathlib

base_dir = "data"
speakers = ["parent", "child"]

models = {}
for spk in speakers:
    wav_files = [
        os.path.join(base_dir, spk, f)
        for f in os.listdir(os.path.join(base_dir, spk))
        if f.endswith(".wav")

    models[spk] = enroll_speaker(wav_files)
    print(f"{spk} 登録完了: {len(wav_files)} files")
```

話者識別

- 学習する音声サンプルの編集
  - ・ノイズ除去→バンドパスフィルタ　◯
  - ・音声区間を抽出　(主に無音区間の削除)　△
  - ・振幅正規化をして音量を一定に　◎

- 今回

・最新の深層学習モデルECAPA-TDNNを使用
・→話者認識に特化
・深層学習に用いられた音声のサンプル数、環境の種類が膨大
・→高精度・雑音や反響に強い
◎

```
判定： child
確信度： {'parent': 0.3615270895, 'child': 0.6384729105}
```

```python
# ECAPA-TDNNモデルロード
classifier = EncoderClassifier.from_hparams(
    source="speechbrain/spkrec-ecapa-voxceleb",
    savedir="pretrained_models/ecapa"
)

def get_embedding(path, sr=16000):
    signal, sr = librosa.load(path, sr=sr)
    signal = signal.astype(np.float32)


    # numpy -> torch.Tensor に変換
    signal = torch.from_numpy(signal).unsqueeze(0)

    embedding = classifier.encode_batch(signal)
    return embedding.squeeze().cpu().numpy()


def enroll_speaker(wav_files):
    embs = [get_embedding(f) for f in wav_files]
    return np.mean(embs, axis=0)  # 話者の平均ベクトル

import os
import pickle
import pathlib

base_dir = "data"
speakers = ["parent", "child"]

models = {}
for spk in speakers:
    wav_files = [
        os.path.join(base_dir, spk, f)
        for f in os.listdir(os.path.join(base_dir, spk))
        if f.endswith(".wav")

    models[spk] = enroll_speaker(wav_files)
    print(f"{spk} 登録完了: {len(wav_files)} files")
```

システムの統合
/UI設計

フロントエンド

- HTML/CSS/JavaScript　シングルページアプリケーションを構成
- Web Speech API　Googleに搭載されたリアルタイム音声認識用API

バックエンド

- Flask　pythonのwebアプリケーションフレームワーク
- identify.py　話者識別モジュール（既に学習済みのモデルを利用）
- attitude_analyzer.py　コマンド分類・態度判定モジュール

統合における技術的課題と解決策

ブラウザ録音はWebM形式である一方、
モデルが判定する音声は16kHzのWAVファイルを期待していた。

▼

pythonのlibrosaライブラリとsoundfileライブラリにより
WebM形式の録音を16kHzにしてさらに形式も変更

UI設計

テーマ：
サイバーパンク&ハッカー風インターフェイス

1. シンクロ率表示

2. 判別話者表示

3. 会話ログ

4. レスポンシブ設計

お母さんスイッチ

VOICE COMMAND AUTHORIZATION SYSTEM v2.0

SYSTEM STATUS

CURRENT STATE

LISTENING

UPTIME

00:06:02

SPEAKER IDENTIFICATION

2

CHILD

User Level Access

CONTROL PANEL

STOP LISTENING

RESET SYSTEM

母親シンクロ率 / AUTHORITY LEVEL

1

SYNCHRONIZATION

48%

0% - USER LEVEL　50% - ELEVATED　100% - ADMIN

STATUS MESSAGE

Moderate authority - Elevated privileges

VOICE INPUT

LISTENING...

CONVERSATION LOG

CLEAR

3

CHILD　13:55:37

INPUT: 電気つけて 音声データ送信済 電気ON 丁寧

OUTPUT: はい、電気をつけます。

CHILD　13:55:20

INPUT: 電気つけて 音声データ送信済 電気ON 丁寧

4

SPEAKER IDENTIFICATION

CHILD

User Level Access

CONTROL PANEL

STOP LISTENING

RESET SYSTEM

母親シンクロ率 / AUTHORITY LEVEL

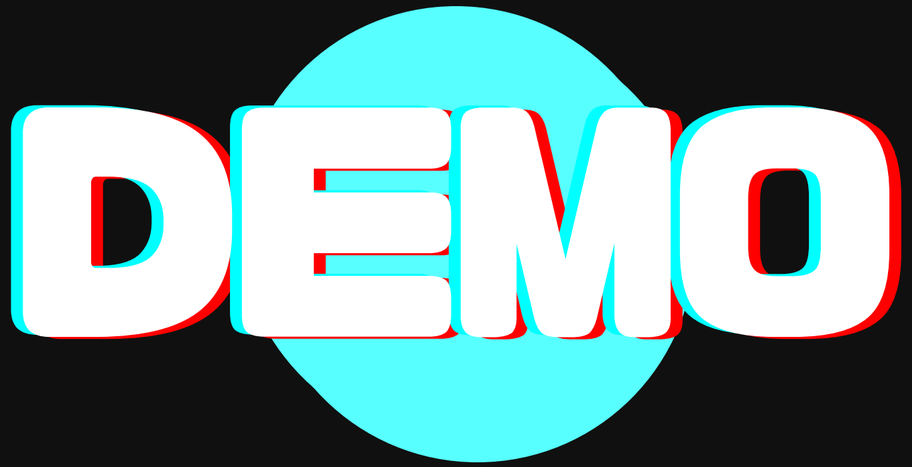
SYNCHRONIZATION

48%

0% - USER LEVEL　50% - ELEVATED　100% - ADMIN

STATUS MESSAGE

Moderate authority - Elevated privileges


DEMO

お母さんスイッチ

VOICE COMMAND AUTHORIZATION SYSTEM v2.0

SYSTEM STATUS

CURRENT STATE

PROCESSING

UPTIME

00:05:06

SPEAKER IDENTIFICATION

UNKNOWN

Waiting for input...

CONTROL PANEL

STOP LISTENING

RESET SYSTEM

母親シンクロ率 / AUTHORITY LEVEL

SYNCHRONIZATION 0%

0% - USER LEVEL　50% - ELEVATED　100% - ADMIN

STATUS MESSAGE

System initialized. Awaiting voice input.

VOICE INPUT

電気つけて

CONVERSATION LOG

CLEAR

No conversation history yet.

改善点/今後の展望

1. 話者識別の精度が不十分
2. 作成した文法にない指令、会話には対応できず柔軟性に欠ける
3. ウェイクワードの設定（アレクサ、Siriなど）
4. 新しいメンバーの声も自動的に登録する機能
5. お母さん以外のモードの実装

THANK YOU FOR LISTENING!