

Vocal Effortの変動に頑健な話者識別のためのフィルタバンク特徴量の設計

吉田 朋矢[†] 生嶋 竜実[†] 武田 龍[†] 駒谷 和範[†]
[†] 大阪大学 産業科学研究所

1. はじめに

話者識別とは、事前登録された話者の音声に基づき、入力された音声かどの登録話者のものかを識別する技術である。この技術はスマートスピーカ等で、パーソナライズされた応答を可能にする。例えば、「音楽を流して」という音声のみから話者を判別し、その話者の好みであった楽曲を提供できる。

登録音声と実際の入力音声では、話者とスピーカ間の距離が必ずしも一致しない(図1)。例えば、スピーカの遠くから話しかける場合、通常よりも声を大きく出す。この場合、残響などの空間的特性と話者の声の出し方の特性(VE: Vocal Effort)が変動するため、話者識別の精度が低下する。空間特性は残響除去等の信号処理で対応可能なため、本稿ではVEの変動の問題を扱う。

VEの変動を扱った従来研究では、VEの強弱を複数含む音声データを用いて、識別用のモデルを高精度化している[1]。VEの強弱に応じた識別器を学習し、入力音声のVEに応じて識別器を切り替えている。一方、識別に用いる音声特徴量は、一般的なメル周波数ケプストラム係数(MFCC: Mel Frequency Cepstral Coefficient)を用いている。

本研究ではVEは低周波帯域に強く影響を与えるという仮説の基、VEの変動に頑健な特徴量を得るため、線形フィルタバンク(LF: Linear Filterbank)を用いる。フィルタバンクの配置が低域に集中するメルフィルタバンク(MF: Mel Filterbank)とは異なり、LFではフィルタバンクの配置が一様となっている。このため、LFはMFよりもVEの影響を受ける特徴量の次元が少なくなることが期待される。本稿では、LFとMFを特徴量に使用した場合の識別精度を実験的に評価する。

2. VEが異なる音声間での話者識別

2.1 埋め込みを用いた話者識別

本研究で用いる話者埋め込みに基づく話者識別器の全体像を図2に示す。話者識別器は特徴量変換、埋め込み抽出、類似度計算、棄却判定から構成される。特徴量変換では、音声信号を特徴量の系列に変換する。埋め込み抽出では、その特徴量系列から話者埋め込みを抽出する。話者埋め込みとは話者性を表す固定長のベクトルである。類似度計算では、登録音声の話者埋め込みと識別音声の話者埋め込みの類似度を計算する。棄却判定では閾値により登録話者か否かを判定する。

この話者識別器について訓練段階、登録段階、識別段階に分けて説明する。訓練段階では、大規模なデータセットで話者埋め込み抽出部の訓練を行う。登録段階では、登録する音声は特徴量変換を経て話者埋め込み抽出部に入力され、得られた登録音声の話者埋め込みが保持される。識別段階では、識別する音声は特徴量変換を経て話者埋め込み抽出部に入力される。得られた識別音声の話

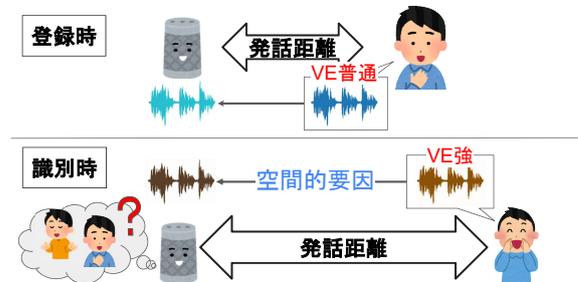


図1: 本研究の焦点: Vocal Effortによる識別精度低下

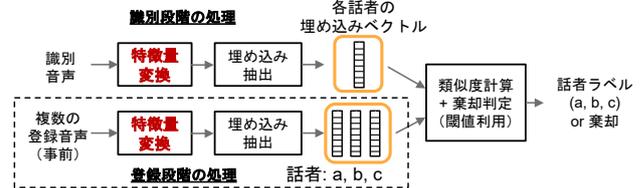


図2: 話者識別器の全体像: 特徴量変換が焦点

者埋め込みと、登録音声の話者埋め込みで類似度を算出する。このうち、類似度が最大となる登録音声か、もっとも識別音声に近いとみなされる。ある閾値を最大の類似度が超えれば、その話者と判定する。逆に、閾値を下回れば、未知話者と判定する。こうして、識別音声か登録した話者のうち誰のものか、あるいは登録した話者のものではないかを判定する。

2.2 従来の特徴量

話者埋め込み抽出への入力特徴量には、フィルタバンク特徴量が用いられる。時間周波数領域での入力音声 $S \in \mathbb{R}^{F \times T}$ からフィルタバンク特徴量 $X \in \mathbb{R}^{D \times T}$ への変換はフィルタバンク $W \in \mathbb{R}^{D \times F}$ を用いて次式によって行われる。

$$X = 20 \log_{10}(W \|S\|) \quad (1)$$

ただし F, T は時間周波数領域での周波数のインデックスの最大値および入力音声の時間フレームのインデックスの最大値を表す。 D はフィルタバンクの次元数を表す。 $\|\cdot\|$ は要素ごとに絶対値を取る操作を表す。

従来、フィルタバンク W にはメルフィルタバンク(MF)が用いられていた(図3上)。MFとはメル尺度に対して等間隔なフィルタバンクである。メル尺度は次式で定義される。

$$\text{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (2)$$

ここで、melはメル尺度、 f は周波数を表す。MFは低周波帯にフィルタバンクが集中している。そのため、低周波帯に変動を多く含む場合、特徴量に大きな影響が出る。

Design of filter bank features for robust speaker identification against vocal effort variation: Tomoya Yoshida, Tatsumi Ikushima, Ryu Takeda, and Kazunori Komatani (Osaka Univ.)

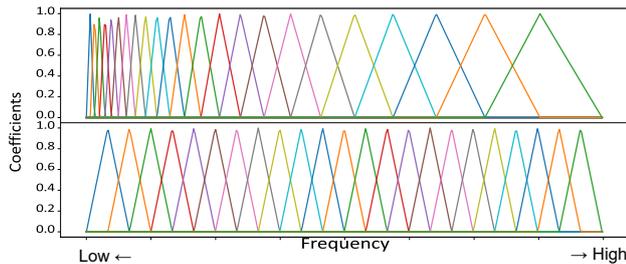


図3: フィルタバンクの比較. 上がMF, 下がLF.

3. VEの変動に頑健なフィルタバンク特徴量

本研究では線形フィルタバンク (LF) を用いる. 予備検討において VE の異なる音声のうち従来の話者識別器で精度が低下するもののスペクトログラムを比較した. これにより, 低周波帯のパワーに変化を確認した. これをもとに, VE の変動が低周波帯に集中しているという仮説を立てた.

LF とは周波数に対して等間隔なフィルタバンクである (図3下). LF では MF のように低周波帯にフィルタバンクが集中していない. そのため, 低周波帯の変動の影響を受ける次元が少ない. また, フィルタバンクを掛ける最低周波数を制限した. これにより, VE の変動に対する頑健性が期待できる.

VE の一種であるささやき声を扱った研究 [2] では, LF および低周波帯除去と同様の手法により頑健性を獲得している. これは, ささやき声と通常音声の変動が低周波帯に集中していることに起因する. 本研究で扱う VE においても同様の効果を期待する.

4. 実験

4.1 実験条件

入力特徴量を変更した場合の話者識別の精度を検証した. 特徴量変換には, LF, LF を適用する周波数帯から 1,000Hz 以下を除去したもの (LF-1000), ベースラインの MF を用いた. すべての特徴量変換で次元数は 23 とした. 話者埋め込み抽出には x-vector [3] を用いた. 類似度計算にはコサイン類似度を用い, その後閾値による未知話者判定を行った.

訓練データには英語読み上げ音声である miniLibriSpeech[‡] の一部を用いた. miniLibriSpeech の音声には VE の変動はほとんど含まれない. 話者数が 23 名, データ長が 1.9-17.3[s] で, 話者ごとに 37 発話を訓練データとした. 訓練データの総データ量は 4 時間強である.

テスト時の登録と識別には著者らが収集した音声を用いた. この音声には発話距離が中距離のもの (normal-VE) と遠距離のもの (high-VE) が含まれている. normal-VE, high-VE とともに話者の口から 10cm 程度の位置にあるマイクで収録した. normal-VE ではマイクに向けて, high-VE では 2.3m ほど離れた位置にある対象に向けて読み上げた音声を収録した.

登録に使用するデータ全体は話者数が 5 名, データ長が 4.1-8.3[s] で, 話者ごとに 5 発話を含む normal-VE の音声である. このうち, 未知話者 1 名と 1 発話を順に選択していく. これにより, 25 個の登録セット作成し

表1: 特徴量ごとの登録セット全体に対する平均 EER および平均 accuracy.

	EER	accuracy
MF	0.347	46.1%
LF	0.312	49.3%
LF-1000	0.430	29.3%

表2: 話者ごとの平均 accuracy.

	話者 A	話者 B	話者 C	話者 D	話者 E
MF	54.0%	27.0%	5.0%	39.0%	81.0%
LF	68.0%	66.0%	0.0%	46.0%	42.0%

た. 識別に使用するデータは話者数が 5 名, データ長が 2.8-13.7[s] で, 話者ごとに 40 発話を含む high-VE の音声である.

登録音声セットごとに 3 種の話者識別器でそれぞれテストを行い, 等価エラー率 (EER: Equal Error Rate) および accuracy を算出した. EER は未知話者判定の可否のみで算出しており, 正しい登録話者と判定されるか否かは関係していない. accuracy は話者識別全体の精度を評価するために算出しており, EER での閾値を用いている.

4.2 結果と考察

表1に平均 EER と accuracy を示す. 識別精度は高い順に LF-1000, MF, LF であった. LF を使用した場合の平均 accuracy は MF を使用した場合に比べて 3.2 ポイント高かった. これは, VE の変動を含む音声間での話者識別において, 精度低下につながる特徴が仮説通り低周波領域に集中していることを示している. 一方, LF-1000 を用いた場合には MF よりも精度が 20.0 ポイント低下している. これは, 低周波成分を除外したことで話者識別に有用な特徴が失われたためである. これらのことから, VE の変動を含む話者識別に用いる特徴量には LF が有効であるとわかる.

表2には MF, LF のそれぞれの特徴量を用いた 2 種の話者識別器の平均 accuracy を話者ごとに示している. 話者 C と E では, MF に比べ LF で精度が低下している. この一因として, これらの話者では VE による変動の影響が高周波帯にまで及んでいることが考えられる. これについては調査による原因特定と対策が必要である.

5. おわりに

本研究では, VE の変動に頑健な話者識別のための特徴量として LF を検討した. 実験の結果, テストデータ全体での平均精度から LF が有効であった.

今回は小規模なデータセットを用いたが, 今後はより大規模なデータを用いた評価実験が必要である.

参考文献

- [1] Mahesh Kumar Nandwana, et al. Analysis and mitigation of vocal effort variations in speaker recognition. In *proc. ICASSP*, pp. 6001–6005. IEEE, 2019.
- [2] Xing Fan, et al. Speaker identification with whispered speech based on modified LFCC parameters and feature mapping. In *proc. ICASSP*, pp. 4553–4556. IEEE, 2009.
- [3] David Snyder, et al. X-vectors: Robust DNN embeddings for speaker recognition. In *proc. ICASSP*, pp. 5329–5333. IEEE, 2018.

[‡]<https://www.openslr.org/31/>