論文 / 著書情報
Article / Book Information

論題(和文)	音声・動画像の因子分析を用いる話者ダイアライゼーション
Title(English)	
著者(和文)	西 史人, 井上 中順, 篠田 浩一
Authors(English)	Fumito Nishi, Nakamasa Inoue, Koichi Shinoda
出典(和文)	日本音響学会2015年秋季研究発表会講演論文集, , , pp. 175-176
Citation(English)	Proc. of Acoustical Society of Japan September 2015, , , pp. 175-176
発行日 / Pub. date	2015, 9

音声・動画像の因子分析を用いる話者ダイアライゼーション*

◎西 史人，井上 中順，篠田 浩一 (東工大)

1 はじめに

話者ダイアライゼーションとは「誰が，いつ」発話しているかを音声や画像の情報を用いて事前情報なしに推定するタスクである．トークショーや映画における話者ダイアライゼーションは電話や会議における話者ダイアライゼーションと比べ，BGM や環境音などの影響が大きい．そのため，音声と映像を用いたマルチモーダル話者ダイアライゼーションが効果的である．Felicien ら [1] はトークショーを対象にした実験で，音声情報と話者の服の色を特徴量として用いているが，本研究の対象である映画のように明暗の切り替わりが激しい映像で用いることは難しい．

そこで本研究では音声・動画像の因子分析を用いる話者ダイアライゼーションを提案する．

2 提案手法

2.1 概要

UBM の学習を認識対象に依存しない大量のデータを用いて行い，UBM とデータから抽出された特徴量から i-vector の抽出に必要なパラメータを推定する．認識では，VAD で音声と動画像を発話ごとに区切った後に発話セグメントから音声，画像の i-vector を抽出し，最後に各セグメントをまとめるクラスタリングを行う．以下，詳細を説明する．

2.2 学習

本研究では各発話から i-vector を抽出する．i-vector は GMM スーパーベクトルを話者と回線が形成する部分空間に射影して得られる特徴量である [2].

$$M = m + Tw \tag{1}$$

ここで，M は話者依存の GMM スーパーベクトル，m は話者に非依存の GMM スーパーベクトル，T は全変動空間を張る基底ベクトルからなる行列，w が i-vector である．発話 u における i-vector w_u は次の式で求める．

$$w_u = (I + T^t \Sigma N(u) T)^{-1} T^t \Sigma^{-1} F(u) \tag{2}$$

ここで $N(u)$，$F(u)$ は UBM パラメータと発話特徴量系列から求められる 0 次，1 次統計量である．

学習時には最尤法で学習した UBM のパラメータとあらかじめ用意した複数話者の発声を用いて，i-vector の抽出に必要な T，及び T で表現できない残留成分を表す対角共分散行列 Σ を推定する．画像についても同様の処理を行う．

2.3 認識

はじめに，音声を VAD によって区切り，MFCC を抽出する．

次に，各セグメントにおける統計量 $N(u)$ と $F(u)$ を UBM パラメータと MFCC から推定し，2 章で得られた T と Σ を用いて音声 i-vector を算出する．画像は VAD によって区切られた音声区間内に存在する顔画像を用いる．顔画像から HOG 特徴量を抽出し，画像 i-vector を作成する．

最後にすべてのセグメントから得られた i-vector に対し，k-means クラスタリングを行う．クラスタ数は既知とし，クラスタの重心を各セグメントからランダムに選ぶことで初期化を行う．

各セグメントと重心の距離 F_{ij} を以下の式で求める．

$$F_{ij} = aA_{ij} + (1-a)V_{ij} \ (0 \leq a \leq 1) \tag{3}$$

ここで，a は重み付けパラメータ，A_{ij} はセグメント i に対応する音声の i-vector w_{Ai} とクラスタの重心 C_{Aj} との距離である．

$$A_{ij} = 1 - \frac{w_{Ai} \cdot c_{Aj}}{||w_{Ai}|| \, ||c_{Aj}||} \tag{4}$$

画像による i-vector w_{Vi} とクラスタ重心 c_{Vj} 間の距離 V_{ij} も同様にして求める．

3 実験

3.1 実験条件

評価実験では映画「ハンナとその姉妹」のデータセットを用いた [3]．データセットには映画「ハンナとその姉妹」の各フレームにおける話者の顔座標，BGM の区間，発話者区間と発話者の情報が記されている．映画は全 106 分である．主要 5 人を対象としたダイアライゼーションを行う．主要な登場人物 5 人のみが現れるよう再編集した映画全 37 分を用いて評価を行った．

また，顔の位置は既知とし，顔画像が 1 つのセグメント内に複数存在する場合は，面積の大きい顔画像を 2 枚まで各セグメントに用い，画像における i-vector を抽出後，2 つのベクトルの平均を取ることで 1 つの

*Speaker diarization using factor analysis to audio and movie. by Fumito Nishi, Nakamasa Inoue, Koichi Shinoda (Tokyo Institute of Technology)

	VAD	Grand-truth
音声のみ	68.4	56.3
画像のみ	67.2	75.3
統合距離	65.9	55.4

Table 1　Voice activity detection (VAD) を用いた場合と Grand-truth でセグメンテーションを行った場合の Diarization Error Rate.　(%)

	左	正面	右
クラスタ 1	0	35.0	65.0
クラスタ 2	47.2	32.4	20.3
クラスタ 3	27.3	20.3	52.4
クラスタ 4	23.6	22.8	53.6
クラスタ 5	50.4	24.3	25.3
全クラスタ	37.1	25.3	37.6

Table 2　各クラスタの顔方向の割合 (%).

ベクトルにまとめる．評価基準には Diarization Error Rate (DER) を用いる．

特徴量について，音声は MFCC15 次元，パワー 1 次元とそれぞれの Δ，Δ Δ を結合した 48 次元，画像は 1 ブロックあたり 2×2 セルに対して 8 方向の 32 次元に x 座標 y 座標を加えた 34 次元の HOG 特徴量を用いる．UBM の混合数は 32 とする．音声，画像 UBM の学習にはそれぞれ，NIST 2004 SRE，Labeled Faces in the Wild (LFW) を用いた．

3.2　実験結果

Fig. 1 にセグメントが Grand-truth の場合における音声重みによる DER の推移を示す．DER が重み a に大きく依存することがわかる．Table 1 は音声のみ，画像のみ，音声と画像を統合した場合の DER を示す．また，セグメンテーションを VAD を用いて行った場合と Grand-truth を用いて行った場合の比較実験も行った．それぞれの場合において統合距離を用いた場合が最も良い結果となっている．これは映画内の雑音の影響を画像における i-vector が軽減しているためだと考えられる．Fig. 2 は各々のクラスタ重心に最も近い画像を示し，各話者と 1 対 1 で対応していることがわかる．Table 2 に各クラスタ内における顔方向の割合を示す．1 つの顔方向が 50%を超えるクラスタが 4 つ，残る 1 つも 47.2%と高い数値となっており，特徴量が顔方向に依存する傾向が見られる．

4　まとめ

本研究では，顔画像から抽出された HOG を用いて作成された画像における i-vector を音声における

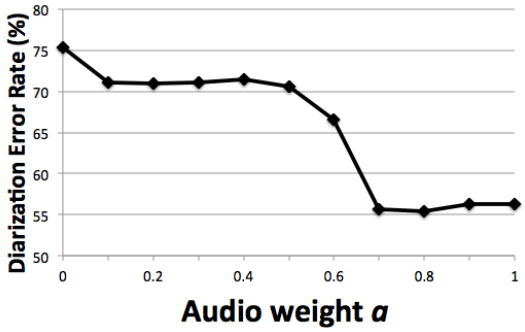

Fig. 1　音声重みによる DER の推移.

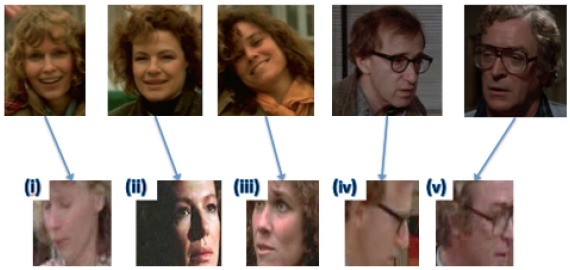

Fig. 2　画像 i-vector 各クラスターにおいて，重心に一番近い画像.

i-vector と組み合わせ，話者の発話区間を推定する話者ダイアライゼーションシステムを提案し，DER が 68.4%から 65.9%に改善することを示した．今後は最適な重み係数を事前に求める手法や顔特徴量を抽出する際の顔方向における正規化処理を行う．

参考文献

[1] Félicien Vallet *et al.* "A multimodal approach to speaker diarization on TV talk-shows", *IEEE Transactions on Multimedia*, Vol. 15, No. 3, pp. 509-520, 2013.

[2] Najim Dehak *et al.* "Front-end factor analysis for speaker verification", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788-798, 2011.

[3] Alexey Ozerov *et al.* "On evaluating face tracks in movies", *IEEE International Conference on Image Processing (ICIP 2013)*, 2013.