T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	マルチモーダルi-vectorを用いた話者ダイアライゼーション
Title(English)	Multimodal i-vectors for speaker diarization
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)	情報処理学会研究報告 SLP, vol. 107, no. 4, pp. 1-6
Citation(English)	, vol. 107, no. 4, pp. 1-6
発行日 / Pub. date	2015, 7
権利情報 / Copyright	ここに掲載した著作物の利用に関する注意: 本著作物の著作権は（社）情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。 The copyright of this material is retained by the Information Processing Society of Japan (IPSJ). This material is published on this web site with the agreement of the author (s) and the IPSJ. Please be complied with Copyright Law of Japan and the Code of Ethics of the IPSJ if any users wish to reproduce, make derivative work, distribute or make available to the public any part or whole thereof.

Powered by T2R2 (Tokyo Institute Research Repository)

マルチモーダル i-vector を用いた話者ダイアライゼーション

西　史人[†1]　井上　中順[†1]　篠田　浩一[†1]

概要：映画を対象とするマルチモーダル話者ダイアライゼーションにおいて，マルチモーダル i-vector を用いる手法を提案する． i-vector とは話者認識において使われている特徴量であり，発話者の情報を表した低次元ベクトルである．音声の i-vector に，動画中の話者の顔画像から抽出した i-vector を結合することで作られたマルチモーダル i-vector に対して教師無しクラスタリングを行う．評価実験は映画「ハンナとその姉妹」のデータセットで行い，Diarization Error Rate (DER) は音声のみを用いた場合比べ，68.3%から 65.5%に改善された．

キーワード：話者ダイアライゼーション，マルチモーダル，i-vector，話者識別

Multimodal i-vectors for Speaker Diarization

Fumito Nishi[†1]　Nakamasa inoue[†1]　Koichi Shinoda[†1]

Abstract: We propose multi-modal i-vectors, which extend the audio i-vector framework for speaker verification to a multi-modal speaker diarization in movies. In addition to the audio i-vector, which represents a speech utterance in an audio stream by a low-dimensional vector, we extract a visual i-vector from faces in a video segment. the audio and visual i-vectors are concatenated as a multi-modal i-vector clustered in an unsupervised way. We evaluate our method on the Hannah movie dataset. Our experiments show that diarization error rate is improved from 68.3% to 65.5% compared with audio stream only.

Keywords: speaker diarization, multimodal, i-vector, speaker verification

1. はじめに

近年，インターネット上の動画配信，テレビ放送などから，映像コンテンツが豊富に提供されている．大量の映像の中から目的とするものを検索する際に，映像内で誰が，いつ発話をしているかという情報は非常に有用である．しかし，そのような情報を人の手で付与することは非常に手間がかかり，全ての映像に対して人手で処理をする事は非現実的である．そこで，自動的に発話情報を，事前情報無しで付与することができれば多大な労力が必要な手動でのアノテーションを回避することができる．

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行うタスクである [1]．動画の自動アノテーションに用いる以外にも話者適応の前処理に用いることで，音声認識の精度を向上させることが期待できる．

電話や会議における話者ダイアライゼーションと比べ，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では BGM や環境音などの影響によって，音声情報のみでダイアライゼーションを行うと精度が低下してしまう．このような環境下では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ことが示されている [2]．たとえば Felicien ら [3] 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ることで精度を上げている．しかし，この手法は衣装の変更がないことや，話者が常に座っていることを前提としているので，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の特徴量としての i-vector に，顔画像から抽出された i-vector を加えたマルチモーダル i-vector を用いる手法を提案する．i-vector とは GMM スーパーベクトルを因子分析して得られる特徴量で，話者分類におけ

[†1] 現在，東京工業大学
Presently with Tokyo institute of technology

© 2015 Information Processing Society of Japan

る精度が高いことが知られている [4]．評価実験は「ハンナとその姉妹」のデータセットで行い，音声のみの i-vector を用いて評価を行った場合と比較して Diarization Error Rate (DER) が 4.3%改善した．

本論文の構成は以下の通りである．第 2 章では従来話者ダイアライゼーションに用いられてきた手法について述べる．第 3 章では提案手法であるマルチモーダル i-vector について，第 4 章では「ハンナとその姉妹」のデータセットを用いた実験について示し，第 5 章で結論を述べる．

2. 関連研究

先行研究 [2, 3, 5–7] において，話者ダイアライゼーションは，1) セグメンテーション，2) 特徴抽出，3) クラスタリングの 3 工程で行われる．本章では各工程における関連研究について述べる．

2.1 セグメンテーション

一続きの区間を均等に分割する手法 [5] や, Voice Activity Detection (VAD) を用いて得られた音声区間のみを用いる手法 [3] がある．

2.2 特徴抽出

先行研究 [3] では Mel-Frequency Cepstral Coefficients (MFCC) と Line Spectral Frequency (LSF) を特徴量として用いている． MFCC は音声認識によく用いられ，人の聴覚特性に基づいた特徴量である．LSF は口の形を表現したモデルに基づく特徴量で暗号符号化の手法として携帯電話などでも用いられている [8]．

話者認識の分野において，MFCC から話者情報を抽出する際に i-vector を用いる手法が効果的であることが示されている [4]． i-vector は Gaussian-Mixture-Model supervector (GMM スーパーベクトル) に因子分析の手法を用いることで得られる．GMM スーパーベクトルとは事前に学習しておいた Gaussian Mixture Model-Universal Background Model (GMM-UBM) を事前分布として用いて発話を GMM でモデル化し，GMM の平均ベクトルを結合することによって得られる特徴量である．平均ベクトルを結合することで音韻の影響を低減することができ，より話者の特徴を表すことができる．

画像特徴量としては服の色を用いた，フレーム毎の HSV ヒストグラムとショット毎の累積 HSV ヒストグラムを用いた研究がある [3]．本研究の対象である映画では明るさが頻繁に変化するため色による認識は難しく，また話者はシーンによって別の服装をしていることも多い．したがって明度の変化に頑健である特徴量を用いる必要がある．明度に対して頑健な特徴量の例として，物体検出で用いられることの多い Histograms of Oriented Gradients (HOG) 特徴量や回転，スケール変化に頑健な Scale-Invariant Feature Transform (SIFT) などがある [9]．O.Deniz は HOG 特徴量が顔認識においても高い性能が得られることを示している [7]．

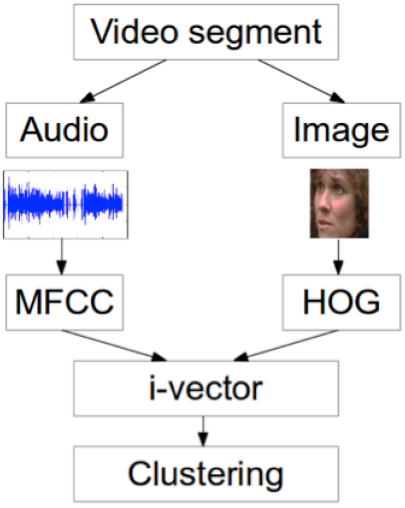

図 1 システム概要図
Fig. 1 System Overview

2.3 クラスタリング

特徴を各セグメントから抽出した後は，各セグメントにおける話者を同定するためにクラスタリングを行う．クラスタリングには大きくわけて教師ありと教師無しの 2 つがある．話者ダイアライゼーションは未知の話者に対して行われるため，教師なしクラスタリングが用いられる．具体的には階層的クラスタリングや，k-means クラスタリングが用いられることが多い．i-vector の評価を行う際，距離尺度としてコサイン距離を用いる場合が最も高い性能となることが示されている [10]．

3. 提案手法

図 1 に提案手法の概要を示す．まず，各セグメントから音声，画像の各特徴量を抽出する．次に i-vector を各特徴量から求め，マルチモーダル i-vector を作成した後に k-means クラスタリングをコサイン距離で行う．

3.1 セグメンテーション

特徴量抽出の前に，発話毎に音声を区切る必要がある．本研究では音声パワーと音声スペクトルの重心を利用した VAD を用いることでセグメンテーションを行った．

3.2 i-vector

まず，低次の特徴量として音声は MFCC，画像は HOG を抽出する．ここで, HOG は各顔画像から抽出を行う．また, i-vector の抽出に必要な UBM は, 学習データの MFCC, HOG 特徴量から，それぞれあらかじめ学習をしておく．

次に，各セグメントに対する i-vector を MFCC，HOG 特

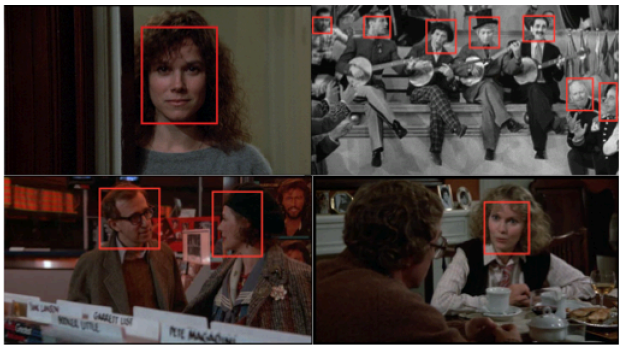

図 2 「ハンナとその姉妹」における映像の例
Fig. 2 Example video segments from the Hannah movie dataset. Bounding boxes for each face are provided.

徴量それぞれから抽出する．i-vector とは GMM スーパーベクトルを因子分析し，話者とチャネル情報空間からなる空間でモデル化するという考え方に基づく特徴量である [4]．M を対象の発話から推定された GMM の平均を連結した GMM スーパーベクトルとすると，i-vector w は次の式で示される．

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

ここで，m は話者，チャネルに非依存の 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構成される低ランクの矩形行列である．T は学習に用いる発話を全て別の話者から発せられるものとみなして固有声の抽出と同じ方法で求めることができる [11]．具体的には L フレーム $y_1, y_2, ..., y_L$ からなる発話 u の i-vector w_u は T と発話 u と UBM を用いた統計量に基づいて以下のように計算される．

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

ここで, $N(u)$ と $F(u)$ は, それぞれ 0 次, 1 次の Baum-Welch 統計量

$$N_c = \sum_{t=1}^{L} P(c|y_t, \Omega) \tag{3}$$

$$F_c = \sum_{t=1}^{L} P(c|y_t, \Omega)(y_t - m_c) \tag{4}$$

を要素とする行列であり，

$$N(u) = \mathrm{diag}([N_c \mathbb{1}]_{c=1}^{C}) \tag{5}$$

$$F(u) = [F_c]_{c=1}^{C} \tag{6}$$

で与えられる．ここで，Ω は混合数 C の UBM のパラメータ，$P(c|y_t, \Omega)$ は y_t が混合要素 $c(c = 1, 2, ..., C)$ から生成される事後確率，m_c は UBM の混合要素 c における平均ベクトル，$[\cdot]_{c=1}^{C}$ は括弧内のベクトルを連結したベクトルである．また，Σ は T で捉えることのできなかった残余を示しており，これは因子分析によって推定される [11]．

3.3 マルチモーダル特徴量融合

音声と映像の i-vector を融合する手法として，Feature fusion 法もしくは Decision fusion 法を用いる．本実験において，Feature fusion は k-means クラスタリングの前に融合，Decision fusion は k-means クラスタリング後に重み付けされたスコアを足し合わせたものである [12]．Decision fusion の場合の最終的なスコア F は以下の式であたえられる．

$$F = aA + (1 - a)V (0 \leq a \leq 1) \tag{7}$$

ここで，A は音声の i-vector から得られたスコア，V は映像の i-vector から得られたスコア，a が重み付けパラメータである．また，k-means クラスタリングでは以下のコサイン距離を用いてクラスタリングを行う．

$$\cos(w_1, w_2) = 1 - \frac{w_1 \cdot w_2}{||w_1||\,||w_2||} \tag{8}$$

ここで，w_1，w_2 は各セグメントから抽出された i-vector である．

4. 実験

4.1 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 [13]．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，主要 5 人を対象としたダイアライゼーションを行う．また，本研究では 5 人のみが登場するように再編集を行った映画を用いて評価を行った．

評価方法としては Diarization Error Rate(DER) を用いる．DER は

$$\begin{aligned} DER =& E_{speaker} + E_{false-alarm} \\ &+ E_{missed-speech} + E_{overlap} \end{aligned} \tag{9}$$

によって求められる．ここで各 E は誤ったラベル付けを行った時間の割合であり，それぞれ $E_{speaker}$ は誤った話者ラベルが付与された場合，$E_{false-alarm}$ は発話区間にラベルが付与されていない場合，$E_{missed-speech}$ は発話のない区間に話者ラベルが付与された場合，$E_{overlap}$ は複数話者が発話してる箇所でそのうちの誰のラベルも割り当てられていない場合である．

セグメンテーションには音声パワーと音声スペクトルの重心を用いて検出した無音以外の区間を用い，HTK [14] を用いて特徴量の抽出を MFCC15 次元，パワー 1 次元とそれぞれの Δ，Δ Δ の 48 次元で行う．画像特徴量としては HOG を用い，ラベル付けされた顔画像に対し 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 次元で抽出を行う．i-vector 抽出のための GMM の混合数は 32 であり，ALIZE [15] を用いて計算をする．

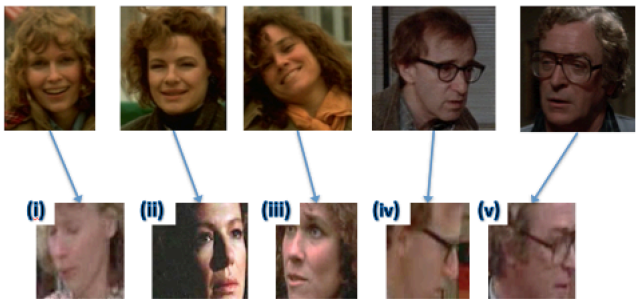

図 3 画像 i-vector 各クラスターにおいて，重心に一番近い画像
Fig. 3 Five speakers in the movie and centroids of resulting five clusters.

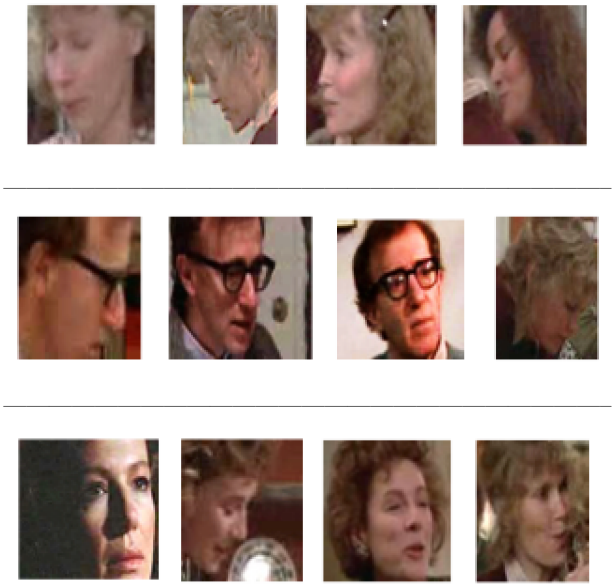

図 4 図 3 で示したクラスタにおける，各クラスタ内の顔画像の例．右端の画像が誤って検出されたもの．
Fig. 4 Examples of faces in profile in each cluster. The rightmost faces are wrongly assigned.

	VAD	Grand-truth
音声のみ	68.3	56.2
画像のみ	67.6	70.6
Feature Fusion 法	67.4	56.0
Decision Fusion 法	65.5	55.2

表 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 (%)． 音声重みを変化させていく Decision fusion 法が最もよい性能を示している
Table 1 Diarization error rate (%) using voice activity detection (VAD) and grand-truth (Manual) for segmentation. Decision fusion reports the best result obtained by using different audio weights.

表 1 はそれぞれ音声のみ，画像のみ，マルチモーダル i-vector を用いた場合の DER を示している．また，セグメンテーションを自動で行った場合と Grand-truth を用いた場合の比較実験も行った．それぞれの場合において，マルチモーダル i-vector が最も良い結果となっている．これは映画内の環境音や BGM などによる音声への悪影響を画像における i-vector が補完しているためであると考えられる．

図 3 はそれぞれのクラスターの中心に一番近い画像を示しており，それぞれ各話者と一対一で対応していることがわかる．しかし図 4 で示されるように，別の顔が間違えて分類される場合も見られた．間違いの多いクラスターでは顔の向きに依存している傾向が見られる．よって顔の向きを正規化することでより精度が高くなることが期待できる．

図 6 は重み a を最適化した際に，クラスタリングに関する変化が顕著に見られたものである．顔が大きく映っている場合はクラスタリングの精度が上がる一方，顔が小さく写ってしまう場合には逆に精度が落ちている場合が見られた．これは顔が小さい場合，顔の特徴となりうる部分が潰れてしまい，特徴を表現することが難しくなるためであると考えられる．

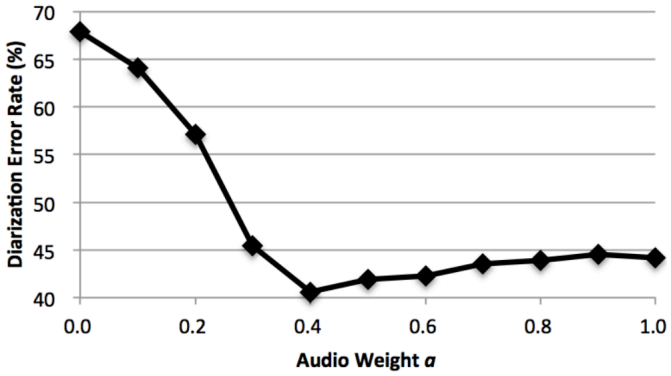

図 5 音声重みを a とした場合の Diarization Error Rate． 3 秒以上の音声ファイルを用い，BGM を用いていないデータでの結果
Fig. 5 Diarization error rate by audio weight a. Long utterances without BGM are used for testing.

実験環境	音声のみ	マルチモーダル
全てのテストデータ	56.2	56.0
BGM を除いたデータ [1]	49.5	48.2
3 秒以上のみを用いたデータ [2]	49.1	47.9
3 秒以下，BGM を除いたデータ	44.2	41.9

表 2 各環境における精度の比較．
Table 2 Diarization error rate (%) under different conditions. No BGM: utterances without background music are used for testing. Long utterances: utterances more than 3 sec are used for testing. Feature fusion is used for “ Multi-modal ”.

4.2 実験結果

4.3 分析

4.3.1 BGM の影響

表 2 に評価データから BGM のある箇所を取り除いた際の結果を示す．これによると，音声のみで評価した場合の DER が低くなっていることがわかる．この場合でもマルチモーダルによって性能は改善されている．これは映画における音声には BGM 以外にも雑音が多く含まれているため，画像によって補完ができているためだと考えられる．

4.3.2 短い発話の影響

３秒以上の音声を用いた場合の結果を表 2 に示す．長い音声のみを用いた場合，それに加えて BGM を除いた場合はいずれにおいても DER が低くなる．音声のみの場合，精度が低いのは発話の短さが原因の一つとなっていると考えられる．実際，[16, 17] は短い発話の場合，i-vector の精度が悪くなってしまうことを示している．

4.3.3 音声，映像の重み付け

図 5 では Decision fusion 法で音声の重みを変化させた場合のグラフである．DER は音声のみの場合の 44.2%から最大でマルチモーダル i-vector における音声の重みが 0.4 の場合の 40.6%まで改善した．

5. まとめ

本研究では映像の各区切りから得た MFCC，HOG を用いて抽出されたマルチモーダル i-vector を用いて各話者を推定するマルチモーダル i-vector を用いた話者ダイアライゼーションシステムを提案した．提案手法によって DER が 68.3%から 65.5%に改善することが示された．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ことが必要になる．

図 6 重みを最適化した際，正しくクラスタリングされるようになった例（上段）と誤ってクラスタリングされた例（下段）

Fig. 6 True positive (Upper) and false positive (Lower) shots at the most effective weight a.

参考文献

[1] Sue E Tranter and Douglas A Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp. 1557–1565, 2006.

[2] Elie El Khoury, Christine Senac, and Philippe Joly. Face-and-clothing based people clustering in video content. In *Proceedings of the international conference on Multimedia information retrieval*, pp. 295–304. ACM, 2010.

[3] Félicien Vallet, Slim Essid, and Jean Carrive. A multimodal approach to speaker diarization on tv talk-shows. *IEEE Transactions on Multimedia*, Vol. 15, No. 3, pp. 509–520, 2013.

[4] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 4, pp. 788–798, 2011.

[5] Johann Poignant, Laurent Besacier, and Georges Quénot. Unsupervised speaker identification in tv broadcast based on written names. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 1, pp. 57–68, 2015.

[6] Claude Barras, Xuan Zhu, Sylvain Meignier, and J Gauvain. Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 5, pp. 1505–1512, 2006.

[7] Oscar Déniz, Gloria Bueno, Jesús Salido, and Fernando De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, Vol. 32, No. 12, pp. 1598–1603, 2011.

[8] 板倉文忠. 音声分析合成の基礎技術とその音声符号化への応用. 電子情報通信学会研資, Vol. 6, pp. 4–5, 2006.

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1, pp. 886–893. IEEE, 2005.

[10] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Interspeech*, Vol. 9, pp. 1559–1562, 2009.

[11] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel. Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 3, pp. 345–354, 2005.

[12] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing*, Vol. 22, p. 23, 2004.

[13] Alexey Ozerov, Jean-Ronan Vigouroux, Louis Chevallier, and Patrick Pérez. On evaluating face tracks in movies. In *IEEE International Conference on Image Processing (ICIP 2013)*, 2013.

[14] Steve J Young and Sj Young. *The HTK hidden Markov model toolkit: Design and philosophy*. Citeseer, 1993.

[15] Anthony Larcher, Jean-François Bonastre, Benoit GB Fauve, Kong-Aik Lee, Christophe Lévy, Haizhou Li, John SD Mason, and Jean-Yves Parfait. Alize 3.0-open source toolkit for state-of-the-art speaker recognition. In *INTERSPEECH*, pp. 2768–2772, 2013.

[16] Ahilan Kanagasundaram, Robbie Vogt, David B Dean, Sridha Sridharan, and Michael W Mason. I-vector based speaker recognition on short utterances. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pp. 2341–2344. International Speech Communication Association (ISCA), 2011.

[17] Achintya Kumar Sarkar, Driss Matrouf, Pierre-Michel Bousquet, and Jean-François Bonastre. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In *INTERSPEECH*, 2012.