Speaker Recognition For Speech Under Face Cover

Rahim Saeidi[*], *Tuija Niemi*[†], *Hanna Karppelin*[‡],
Jouni Pohjalainen[*], *Tomi Kinnunen*[⋆], *Paavo Alku*[*]

[*]Department of Signal Processing and Acoustics, Aalto University, Finland
[†] Forensic Laboratory, National Bureau of Investigation, Finland
[‡] Faculty of Behavioural Science, University of Helsinki, Finland
[⋆]Speech and Image Processing Unit, School of Computing, University of Eastern Finland
rahim.saeidi@aalto.fi, tuija.niemi@poliisi.fi, hanna.karppelin@helsinki.fi,
jouni.pohjalainen@aalto.fi, tomi.kinnunen@uef.fi, paavo.alku@aalto.fi

Abstract

Speech under face cover constitute a case that is increasingly met by forensic speech experts. Wearing face cover mostly happens when an individual strives to conceal his or her identity. Based on the material of face cover and the level of contact with speech production organs, speech production becomes affected by face mask and a part of speech energy gets absorbed in the mask. There has been little research on how speech acoustics is affected by different face masks and how face covers might affect performance of automatic speaker recognition systems. In the present paper, we have collected speech under face mask with the aim of studying the effects of wearing different masks on state-of-the-art text-independent automatic speaker recognition system. The preliminary speaker recognition rates along with mask identification experiments are presented in this paper.

Index Terms: Speaker Recognition, Face Cover

1. Introduction

Speech is the most natural way of communication between humans. Apart from the spoken words, speech signal conveys information about identity of the speaker, emotional state, acoustic environment, language and accent. Speaker recognition is the task of identification or detection of the underlying speaker in a recorded speech. Forensic speaker recognition entails detection of individuals from any possible scenario of speech recording in a crime scene. From this perspective, recognition systems encounter difficulties in dealing with modified, forged or naturally altered material in their evaluation stage [1,2]. According to the guidelines of *European network of forensic science institute* (ENFSI), these challenges could lead to a decision; not to proceed with further comparison analysis [3]. The state-of-the-art techniques in speaker recognition cannot produce reliable speaker comparison results under challenging forensic conditions which limits the admissibility of recorded speech evidence to a court of law.

In forensic speech analysis, intentional voice modifications play a significant role in misleading automatic recognition system or even a human expert listener. Imitation, synthesized speech and speaking under face cover can be mentioned as examples. Studying speech under face cover has gained more attention after James Foley's case in Iraq [4] where forensic speech scientists aim at finding a militant who speaks under face cover. In this specific case, a technique called *language analysis for the determination of origin* helped investigators to limit the pool of suspects to specific geographical area.

There are various intrinsic and extrinsic factors that cause undesirable variabilities in the speech signal in view of an automatic recognition system. *Intrinsic variability* refers to human factors in speech production such as vocal effort, speaking rate and emotion. *Extrinsic variability*, on the other hand, refers to how the acoustic speech signal reaches the recognition system or a human listener. This involves transmission media which introduces both environmental noise (surrounding signals from the street or from other devices) as well as channel distortions due to recording device or transmission channel such as telephone line or IP network. Extrinsic factors affecting a recognition system correspond to issues that cause a change in natural speech signal after it is being generated. Intrinsic factors, on the other hand, are a collection of the effects that result in variation in realization of an acoustic event in the generation phase.

Covering face, which is an event that frequently happens in crime cases, involves *both the intrinsic variability* (i.e. face cover affects production of speech) and *the extrinsic variability* (i.e. signal absorption in the face cover). The material of the face cover, degree of lip/nose contact, restricted jaw elevation and skin stretching are the most important factors related to speech under face cover.

In this paper, we report text-independent speaker recognition where speakers wear 4 different forensically relevant face masks. We introduce a new speech corpus which is collected to support the research in this study. By employing a state-of-the-art i-vector based recognition system, we train speaker-specific models with speech recorded under different face masks. The normal speech referred to as "no mask" serves as a natural choice for training utterance. In test phase, we report recognition rates under both matched and mismatched conditions with respect to the use of face cover. We further look into mask classification scenario where the type of face mask in a short utterance is being identified from a closed-set of face masks.

2. Speaking Under Face Cover

One of the frequent situation in the caseworks referred to forensic speech scientists is when the talker wears a face mask. Nevertheless, there has been limited research addressing the effects of different face covers on speech acoustics and consequently on speaker recognition systems performance. Wearing a face cover affects the recorded speech in both active and passive

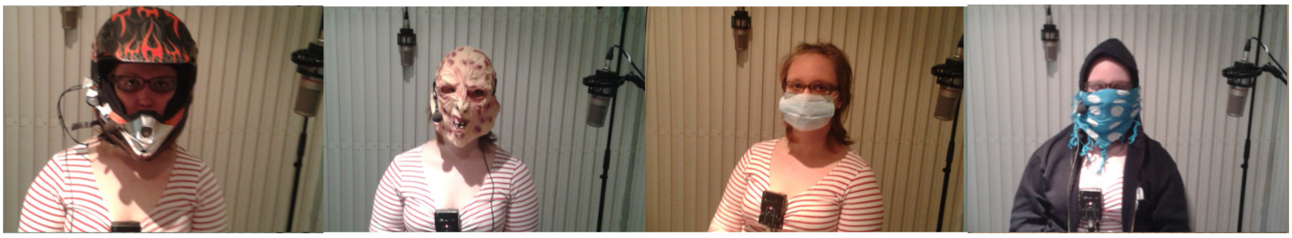


Figure 1: The speech material under face cover is collected with support from Finnish *National Bureau of Investigation* and University of Helsinki. A Finnish female volunteer wears 4 different face covers: motor cycle helmet, rubber mask, surgical mask and hood + scarf. The speech is recorded with 3 microphones. Both spontaneous and reading speech are considered.

manner. Apart from the acoustic absorption properties of the mask, by wearing a face mask, some of the speech articulation mechanism are also affected. Depending on the mask type and amount of its contact and pressure on face, the lips and jaw movements mostly become restricted. These muscle constrictions would in turn change the normal articulation of consonants like /p/ and /m/. The limited movement of the jaw caused, for example, by wearing a motorcycle helmet may result in a reduction of the range of the first formant of open vowels [5]. On the other hand, the talker might increase the vocal effort in order to compensate for the effect of face cover. Such effects are not extensively studied in the literature and the resulting effect on speaker recognition is consequently not clear.

In [6], along with other voice disguises, the effect of wearing a surgical mask is investigated for automatic speaker recognition in a speaker-specific way. The authors looked into the identification scores for each member of a group of target speakers separately and found that wearing a surgical mask affects the recognition system performance quite adversely. In [7], the intelligibility of speech produced with three face covers; *niqab* (a cloth mask worn by muslim women), *balaclava* (a ski mask that exposes part of face only) and surgical mask is investigated. The authors found that listeners can reliably identify the target words independent of the type of the mask.

In an attempt to measure frequency response of the masks in [7], *transmission loss* is measured by playing speech through a loudspeaker and recording it again by a microphone that is separated from the loudspeaker by face mask. In this setup of measurements, the authors found minor differences between the transmission loss across different mask fabrics. Earlier in [8], the acoustic transmission loss of 44 woven fabrics were measured in different audible frequency bands. According to the measurements in [8], the transmission loss depends to a large extent on the weight, thickness, and density/porosity of the fabric. It was observed that sound energy absorption in different fabrics results in more energy loss in high frequency ranges than low frequencies. In a recent study, looking into wearable microphones [9], no difference was observed between the transmission characteristics of different shirt types or between shirts and the bare-microphone condition.

Audio-Visual Face Cover (AVFC) corpus [5, 10] is a speech database consisting of carefully controlled, high-quality audio and video recordings of talkers whose faces were covered by a comparatively large variety of forensically-relevant face and head coverings. It consists of phonetically controlled /C_1ɑ: C_2/ syllables by using two each of 18 consonants in two syllable positions with the nucleus selected to be open back vowel /ɑ:/. The database entails recordings from 10 native British English speakers (5 males and 5 females). Despite a major limitation of the corpus; the highly-controlled speech material in the form of (mainly) nonsense syllables, the neat design of the corpus allowed detailed acoustic analysis of the effect of face masks on fricatives and plosives [11, 12].

3. Corpus Description

This study presents an ongoing data collection for speech under face cover with the focus of forensic automatic speaker recognition. Four types of face masks that are typically worn for commission of crimes or in situations of public disorder are considered. These face masks are shown in Figure 1.

- **Helmet**: The subject wears a motor cycle helmet.
- **Rubber mask**: A latex mask covering the whole face is utilized which has holes for eyes and mouth.
- **Surgeon mask**: The subject wears a thin mask typically being used for anti-pollution purpose or in surgical operations.
- **Hood + scarf**: The subject is wearing a hood which limits the jaw movement. On top of the hood, a light scarf covers speaker's mouth and nose.

Recordings were made in the Faculty of Behavioral science's (located in Kruununhaka, Helsinki) studio at University of Helsinki which is a soundproof, about 5 square meters

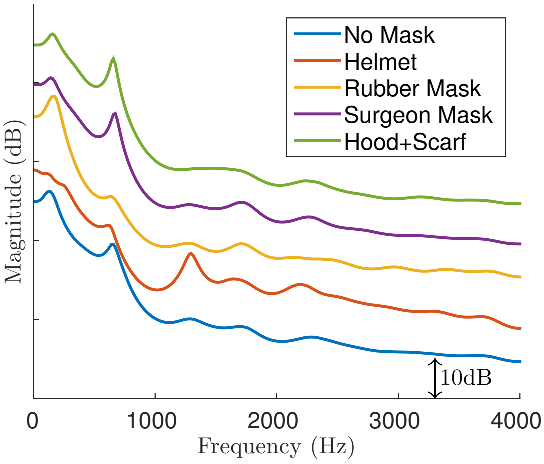
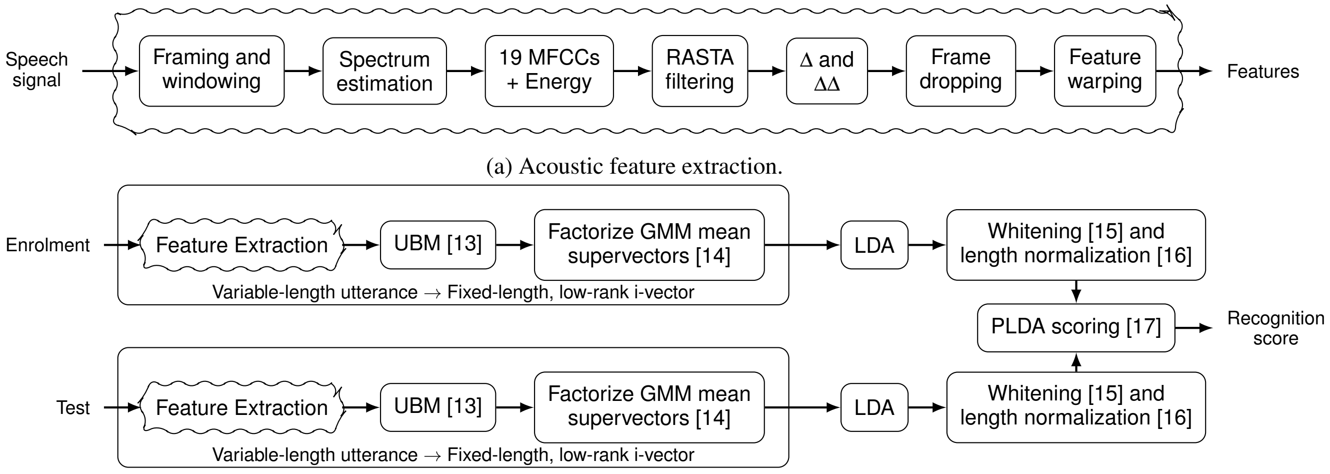


Figure 2: Long-term average spectrum for a male speaker calculated using linear prediction (order of $p = 20$) from voice parts of an utterance phonated in no mask and through 4 different masks. The audio is captured with close talking microphone. The speaker reads fixed text in the utterances. The spectra are shifted 10dB for better visual comparison.


Speech signal
Framing and windowing
Spectrum estimation
19 MFCCs + Energy
RASTA filtering
Δ and ΔΔ
Frame dropping
Feature warping
Features
Enrolment
Feature Extraction
UBM [13]
Factorize GMM mean supervectors [14]
Variable-length utterance → Fixed-length, low-rank i-vector
LDA
Whitening [15] and length normalization [16]
PLDA scoring [17]
Recognition score
Test
Feature Extraction
UBM [13]
Factorize GMM mean supervectors [14]
Variable-length utterance → Fixed-length, low-rank i-vector
LDA
Whitening [15] and length normalization [16]

(a) Acoustic feature extraction.

(b) Extracting, post-processing and comparing i-vectors.

Figure 3: Block diagram of i-vector based speaker recognition system in our experiments.

large room and it has two windows and double doors. The data is originally recorded in 44.1kHz, 16 bit Mono format but the sampling frequency is reduced to 8kHz for the sake of speaker recognition experiments in this paper. The data has been recorded with 3 microphones simultaneously; a headset placed near the speaker's mouth (AKG C444 model), a microphone attached on the wall on the right side of the speaker and a microphone placed behind the speaker (both AKG C4000B model).

The volunteers were asked to read a set of sentences for one recording and for the next recording they chose a picture from a set of comics and paintings to speak spontaneously. The first recording is designed to encompass a phonetically rich fixed text read by all speakers. The list of sentences is provided in Section 7. The second recording is deemed to simulate spontaneous speech, different from reading speech and vary across speakers and sessions. Each speaker's recordings include fixed text and spontaneous speech under control condition (no mask) and the recording were repeated by wearing 4 different types of face masks. Each recording scenario repeated in two sessions on the same day. The Speakers aged between 21 and 28 years old. All were native Finnish speakers and university students. Prior to taking part, the participants were informed about the procedure both in written and verbal form so that they could grant their informed consent to participate.

Each recording lasts between 60 to 90 seconds. The control recording dubbed as *no mask* was recorded under normal vocal effort and no face cover. The speech collection includes 4 males and 4 females. Considering data collection using 3 microphones, under 4 masks plus no mask condition, read speech and spontaneous types and 2 sessions, we have 60 audio files per speaker amounting to 1.5 hours of speech data for every speaker. In Figure 2, the *long-term average spectrum* of fixed text utterance for a male speaker is shown across 4 different masks as well as "no mask" condition. This analysis suggests that surgeon mask and hood+scarf mostly affect the spectral properties above 1 kHz where in case of helmet and rubber mask, the deviation from "no mask" is observed in low frequency range as well. We leave detailed acoustic analysis of different face masks in this corpus for future research and focus on automatic speaker recognition in the next section.

4. Speaker and Mask Recognition

Text-independent speaker verification has gained considerable attention in the last two decades [18]. The so-called *i-vector* approach [14] is the state-of-the-art approach in text independent speaker recognition. The structure of our i-vector based recognition system is shown in Figure 3b. For i-vector recognition system, as it typically happens in real forensic applications, we take a state-of-the-art recognition system [19, 20] *off-the-shelf* where recognition system parameters cannot be adapted to the test condition because of data scarcity.

4.1. Experimental Setup

The block diagram of feature extraction stage is depicted in Figure 3a. A short time spectrum is estimated for speech frames of 30 msec with a frame shift of 15 msec. We used *linear prediction* method for spectrum estimation with a prediction order of $p = 20$. Next, 19 Mel-frequency cepstral coefficients are extracted and appended by frame energies. After RASTA filtering [21], Δ and $\Delta\Delta$ features are calculated to form 60-dimensional feature vectors. At last, active speech is retained based on frame-level energy and feature warping [22] is applied.

A gender-dependent *universal background model* (UBM) [13] with 2048 components is trained using a subset of NIST SRE 2004–2006, Switchboard cellular phase 1 and 2, and the Fisher English corpora. To factorize the GMM mean supervectors, the total variability space [14] is trained with the same data as for UBM with 400-dimensions. In post-processing of utterance-level i-vectors, we used linear discriminant analysis (LDA) projection to enhance separability of classes (speakers) and reduce the i-vectors dimension to 200. Prior to Gaussian probabilistic linear discriminant analysis (PLDA) [17, 23] modelling, we remove the mean, perform whitening using within-class covariance normalization (WCCN) [15] and normalize the length of i-vectors [16].

We chose one session of fixed text speech for each speaker as for making the speaker template. Speaker templates are made separately for each speaker under different masks as well as with no mask condition. In the template, the i-vectors extracted from three different microphones are averaged in order to reduce the recognition results sensitivity to channel mis-

Table 1: Closed-set correct speaker identification rate reported in percent (%) when speaker models are trained and tested with different masks. The rows correspond to face mask in the template and columns represent the face mask in test.

Face cover	No cover	Helmet	Rubber mask	Surgeon mask	Hood + scarf
No cover	**95.2**	94.9	88.6	94.2	93.3
Helmet	88.5	**97.7**	86.0	88.8	88.4
Rubber mask	90.3	96.5	**97.1**	94.1	91.4
Surgeon mask	95.1	96.7	90.1	**97.9**	95.6
Hood + scarf	90.3	85.7	82.5	94.9	**97.0**
Number of tests	793	574	543	626	568

match. The training side, on average, includes around 25 seconds of active speech while for test side only non-overlapping segments of 2.5 seconds active speech are considered in these experiments. The test segments are extracted from spontaneous speech uttered under different masks where all three microphones and two sessions are utilized. In the identification experiments there is no cross-gender trial and each test segment is evaluated against 4 gender-matched speaker templates. The top scoring speaker is identified as the underlying speaker.

4.2. Experimental Results

The closed-set speaker identification results are presented in Table 1. In matched condition, the speaker identification rate for the "no cover" case is slightly inferior compared to other cases. At this point, the reason behind is not clear and our interpretation is hindered by short duration of tests segments and limited number of trials available for each condition. When the template and test segment are from the speech under the same mask, a high correct identification rate is observed. The comparison of the highlighted diagonal elements of the table suggests that although under the matched mask condition the recognition system performs well, a degradation in recognition performance occurs when the template and test segment are from different masks. The amount of degradation depends on the mask type. When the speaker template is derived from speech with no face mask, the highest decline in performance happens for speech under rubber mask. Interestingly, testing with other face masks does not degrade the recognition performance dramatically. Arguably, compared to other face masks in this study, wearing a rubber mask entails the highest contact with facial organs and results in more active compensation from the talkers during speaking.

The results in Table 1 show that the test data in specific mask condition match the template trained by speech under the same mask best. This observation motivates us to develop an automatic face mask classification system. The 60-dimensional acoustic features of fixed-text segments for all speaker (irrespective of their gender) are pooled together and a Gaussian mixture model (GMM) [24] with 64 components and diagonal covariance structure is trained with maximum likelihood criterion for each mask. The same test segments that we used for testing the speaker identification system are employed in mask classification experiment. The results are shown in Table 2. In light of our experiments, speech with no mask can be correctly classified in 75% of trials. As it is highlighted in Table 2, surgeon mask and hood + scarf are less confused with rubber mask and helmet.

Table 2: Confusion matrix for closed-set mask identification reported in percent (%). The test segmets are the same ones used for speaker identification as in Table 1.

Face cover	No cover	Helmet	Rubber mask	Surgeon mask	Hood + scarf
No cover	73.9	28.7	16.9	40.7	24.5
Helmet	6.3	44.4	10.9	**9.9**	**4.6**
Rubber mask	4.5	12.5	56.2	**7.2**	**5.8**
Surgeon mask	4.7	**6.3**	**2.8**	17.1	10.4
Hood + scarf	10.6	**8.0**	**13.3**	25.1	54.8

5. Conclusions

We presented a first study on the effect of wearing different face masks on the state-of-the-art automatic text-independent speaker recognition system. A relatively small speech corpus is collected in support of this study consisting of 8 speakers and 4 different forensically relevant face masks. This paper presents preliminary studies on matching spontaneous speech under face cover with normal reading speech in context of speaker identification and mask classification tasks. The i-vector based speaker recognition system experiences performance deterioration when used in mismatched face mask conditions. However, the small relative degradation indicates the capability of the state-of-the-art recognition systems in partially mitigating the face mask mismatch. As the future research we need to look into the acoustical changes attributed to different parts of speech individually in order to gain more knowledge on the effect of wearing a face cover on speech signal. The gained knowledge can be employed in building more robust speaker recognition system for use across a wide variety of forensic situations. An in-depth study is in place for efficient classification of face mask.

6. Acknowledgement

This work was supported by Academy of Finland (project numbers 256961, 284671, 253120 and 283256). We acknowledge the computational resources provided by the Aalto Science-IT project.

7. Fixed-text sentences

Gerberat ja jaguaarit eivät ole demagogien alaa.
Agraariseen kotiseutuun liittyy nostalgiaa.
Bodaajankin näkökulma on huomioitava.
Fissio- ja fuusioenergian käsitteet ovat problemaattisia.
Barbaarimainen käytös vaikutti lähinnä tökeröltä.
Estradi vapautui ballerinan kerättyä flegmaattiset aplodit.
Pengerrykset sabotoivat vehreän maiseman.
Kofeiini on efektiivistä ainetta.
Täällä Ninni on purrut hammasta.
Guldenista tunnetaan myös nimitys floriini.
Abstrakti ajattelu hyödyttää ergonomiassa.
Lahjaröykkiö jökötti kadulla kuin kivivuori.
Anglikaaniseen eli episkopaaliseen kirkkoon kuuluu konfirmaatio.
Kaftaanikankaiden hankinta on lisännyt produktiviteettia.
Bakteerisolujen kahdentuminen tapahtui dramaattisella vauhdilla.
Snobien ja päihdeongelmaisten hankaluuksien lähtökohtana lienee identiteettikriisi.
Pääjohtaja falsifioi gangsteriliigan alibin.

8. References

[1] N. K. Ratha, J. H. Connell, and R. M. Bolle. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal*, 40(3):614 –634, 2001.

[2] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li. Spoofing and countermeasures for speaker verification: a survey. *Speech Communication*, 66:130–153, February 2015.

[3] Terms of reference for forensic speaker analysis (FSAAWG-TOR-FSI-001), 2008. `http://www.enfsi.eu/`.

[4] Voice, words may provide key clues about James Foley's killer, 2014. `http://www.cnn.com/2014/08/22/world/europe/british-jihadi-hunt/index.html`.

[5] N. Fecher. *Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants*. PhD thesis, Language and Linguistic Science, The University of York, UK, 2014.

[6] C. Zhang and T. Tan. Voice disguise and automatic speaker recognition. *Forensic Science International*, 175(2–3):118–122, 2008.

[7] C. Llamas, P. Harrison, D. Donnelly, and D. Watt. Effects of different types of face coverings on speech acoustics and intelligibility. *York Papers on Linguistics*, 9(2):80–104, 2009.

[8] M. E. Nute and K. Slater. The effect of fabric parameters on sound-transmission loss. *The Journal of The Textile Institute*, 64(11):652–658, 1973.

[9] M. VanDam. Acoustic characteristics of the clothes used for a wearable recording device. *Journal of the Acoustic Society of America*, 136(4):263–267, 2014.

[10] N. Fecher. The audio-visual face cover corpus: Investigations into audio-visual speech and speaker recognition when the speaker's face is occluded by facewear. In *Proc. Interspeech 2012*, 2012.

[11] N. Fecher and D. Watt. Speaking under cover: The effect of face-concealing garments on spectral properties of fricatives. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)*, pages 663–666, August 2011.

[12] N. Fecher and D. Watt. Effects of forensically-realistic facial concealment on auditory-visual consonant recognition in quiet and noise conditions. In *International Conference on Auditory-Visual Speech Processing (AVSP 2013)*, 2013.

[13] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, 2000.

[14] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 19(4):788–798, 2011.

[15] A. O. Hatch, S. Kajarekar, and A. Stolcke. Within-class covariance normalization for SVM-based speaker recognition. In *Proc. Interspeech 2006 (ICSLP)*, pages 1471–1474, Pittsburgh, Pennsylvania, USA, September 2006.

[16] D. Garcia-Romero and C. Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Proc. Interspeech 2011*, pages 249–252, 2011.

[17] S. J. D. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *11th International Conference on Computer Vision*, pages 1–8, 2007.

[18] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 52(1):12–40, 2010.

[19] R. Saeidi and D. A. van Leeuwen. The Radboud University Nijmegen submission to NIST SRE-2012. In *Proc. NIST SRE 2012 workshop*, Orlando, US, December 2012.

[20] R. Saeidi, et al. I4U submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification. In *Proc. Interspeech 2013*, pages 1986–1990, 2013.

[21] D. Hardt and K. Fellbaum. Spectral subtraction and RASTA-filtering in text-dependent HMM-based speaker verification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1997)*, pages 867–870, Munich, Germany, April 1997.

[22] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)*, pages 213–218, Crete, Greece, June 2001.

[23] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson. Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, 2012.

[24] D. Reynolds and R. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3:72–83, January 1995.