INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Analysis of Face Mask Effect on Speaker Recognition

Rahim Saeidi, Ilkka Huhtakallio and Paavo Alku

Department of Signal Processing and Acoustics
School of Electrical Engineering, Aalto University, Finland
{rahim.saeidi, ilkka.huhtakallio, paavo.alku}@aalto.fi

Abstract

Wearing a face mask affects the speech production. On top of that, the frequency response and radiation characteristics of the face mask - depending on the material and shape of the mask - adds to the complexity of analyzing speech under face mask. Our target is to separate the effect of muscle constriction and increased vocal effort in speech produced under face mask from sound transmission and radiation properties of face mask. In this paper, we measure up the far-field effects of wearing four different face masks; motorcycle helmet, rubber mask, surgical mask and scarf inside anechoic chamber. The measurement setup follows the recording configuration of a speech corpus used for speaker recognition experiments. In matching speech under face mask with speech under no mask, the frequency response of the respective face mask is accounted for and compensated for before acoustic feature extraction. The speaker recognition performance is reported using the state-of-the-art i-vector method for mismatched and compensated conditions in order to demonstrate the significance of knowing the type of mask and accounting for its sound transmission properties.

Index Terms: face mask, speaker recognition,

1. Introduction

Speech is the most natural way of communication for humans. Nowadays, considering the recent advancements in *natural language understanding*, machines are also expected to be able to comprehend human speech or even talk to each other [1,2]. Humans are able to produce speech signal in different styles when communicating to a distant person, speaking in noisy environment or expressing emotions vocally. Not only the prosody of speech is affected by the manner of speaking, but also many acoustical voice properties change when speaking style is varied from whispering to shouting. This variability is undesirable in view of robust speech analysis and causes difficulties in automatic information extraction from speech signal.

In automatic information extraction such as recognition of speech content, speaker, language or gender, it is customary to build an acoustic model using hundreds of hours of speech collected in normal conditions, that is, the speaker is not under emotional or cognitive load and the ambient is noise free. The state-of-the-art techniques in different fields of speech technology are able to produce an accurate recognition of the underling information in normal conditions. However, when the modern recognition systems are being tested in unseen conditions, the outcome is less accurate and sometimes unreliable. A common condition for speech recordings collected from a crime scene is to have the signal captured by a far-field microphone and the speaker is covering his/her face with a mask. Forensic voice comparison becomes very challenging when speech under face mask is compared to interview quality normal speech. As another example, the doctors wearing an occupational surgeon mask would also expect the automatic speech recognition to work seamlessly.

Wearing a face mask introduces both *active* (change in speaking style and effort) ans *passive* (filtering due to mask material and shape) on the recorded speech. It is difficult to ascertain how much of the changes in speech spectrum is attributed to active effects of wearing a face mask. However, the passive effects of wearing a face mask can be compensated in part by measuring up the acoustic properties of the material of a face mask. One of the contributions of the current study is to report the measurements of the transfer functions for four forensically-relevant worn face masks. This study is a sequel to our earlier work in [3] by utilizing the same speech corpus and speaker recognition system configuration. The current study focuses on the importance of mask compensation in magnitude spectral domain in acoustic feature extraction. We present a comparative study demonstrating the usefulness of a simple *direct inversion* of the measured magnitude of a face mask's transfer function. The inversion compensating for the effect of a face mask is then used in speaker identification experiments under mismatched conditions.

2. Speaking Under Face Mask

There is little research on how wearing different face masks affects the acoustical properties of speech. By wearing a face mask, speech production mechanism adapts itself in order to compensate for the new situation. The amount of change in speech production depends on the mask type, the amount of its contact with speech production organs and how much wearing the face mask affects perception of own voice [4]. In [5], the effect of wearing a face mask is considered in terms of a *transmission loss* by playing speech through a loudspeaker and recording it again by a microphone that is separated from the loudspeaker by a face mask. Similarly, in [6], the transmission loss of 44 different woven fabrics were measured in an audible frequency range.

2.1. Speech Corpus

Our earlier study in [3] presented a collection of speech under face mask with the focus on forensic automatic speaker recognition. The recordings were made in a studio in the Faculty of Behavioral Science at University of Helsinki. The speech data is originally recorded in 44.1kHz and later downsampled to 8kHz to make it suitable for front-end processing configuration in speaker recognition experiments. The data has been recorded with 3 microphones simultaneously; a headset placed near the speaker's mouth, a microphone attached on the wall on the right side of the speaker and a microphone placed behind

http://dx.doi.org/10.21437/Interspeech.2016-518

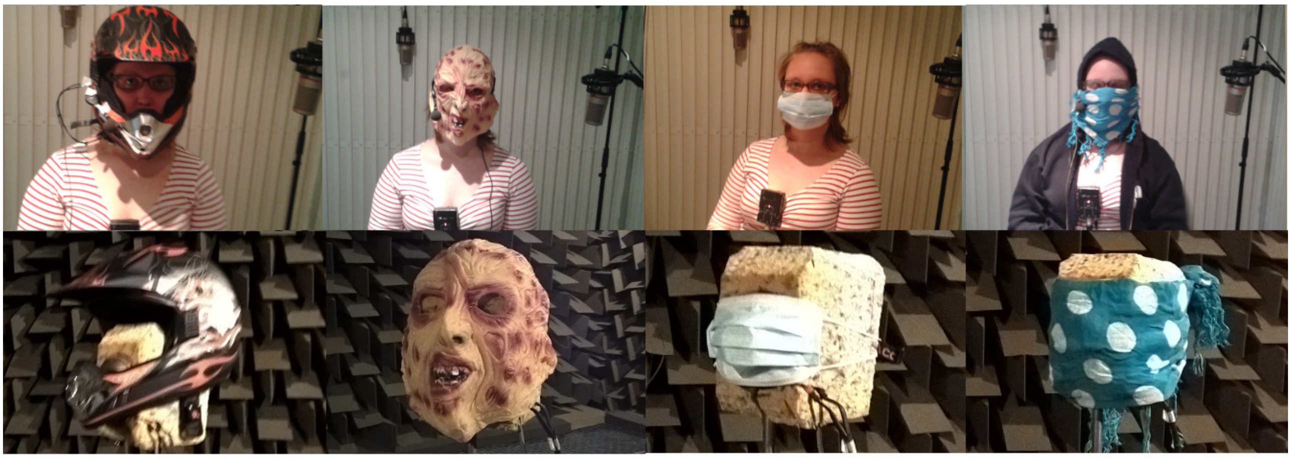

Figure 1: In the top row, a volunteer wears 4 different face masks: motorcycle helmet, rubber mask, surgical mask and scarf. The speech material under face mask is collected with support from Finnish *National Bureau of Investigation* and University of Helsinki. In the bottom row, we simulate the recordings with face mask inside anechoic chamber by using an artificial voice generator devised in a pretend dummy head.

the speaker. Four types of face masks as it is shown in top row of Figure 1 is considered: a motorcycle **helmet**, a latex **rubber mask** covering the whole face, a thin **surgeon mask** and a **scarf** which limits the jaw movement to some extent.

The control recording with **no mask** was recorded in normal condition. The corpus includes speech from 4 males and 4 females in both reading and spontaneous format. We have collected about 1.5 hours of speech data for every speaker including speech under four different face masks and no mask conditions each repeated in two sessions and recorded with three microphones. The interested reader can find further details of the speech data collection in [3].

2.2. Measuring Acoustic Properties of Four Face Masks

In order to measure up the passive effects of wearing the face masks targeted in this study, we used an artificial voice generator and carved a block of high density foam to hold the voice generator and to resemble a dummy head. The pretend dummy head is shown in action in bottom row of Figure 1, where we set up a recording condition similar to the speech data collection under face mask. Measurements were conducted in a large anechoic chamber of Department of Signal Processing and Acoustics in Aalto University, Espoo, Finland. The chamber has 7 m cone to cone spacing in all three dimensions. The measurement location was 2.5 m apart from the radiation aperture of the source.

Effect of masks on spectral balance of radiation was studied by measuring a far field response of B&K Type 4219 artificial voice mounted on the dummy head. A close range shot of the dummy head and recording devices is shown in Figure 2. G.R.A.S. Type 46AF free-field microphone was used together with signal conditioner G.R.A.S Type 12AG to record the sound pressure signal, while Yamaha MX-70 power amplifier fed the cone driver inside the B&K Type 4219 artificial voice. This setup allows us to obtain a parallel measurement of the voice immediately after it is being generated in the dummy head as well as from the microphone in the far-field. In this way, it is feasible to perform a reasonably accurate analysis of the passive effects of a face mask .

A schematic block diagram of the recording configuration is shown in Figure 3. Fuzzmeasure 3.0 software running on a Macbook Pro was used for measurements while RME FF400 sound card utilized the D/A conversion of the playback signal and A/D conversions of the response signal at 48 kHz sampling rate. A 4 seconds long logarithmic sine sweep from 1 Hz to 24 kHz was used as the excitation signal. The level was adjusted to give a desirable signal to noise ratio on distant microphone while keeping the distortion of the artificial voice at a reasonable level.

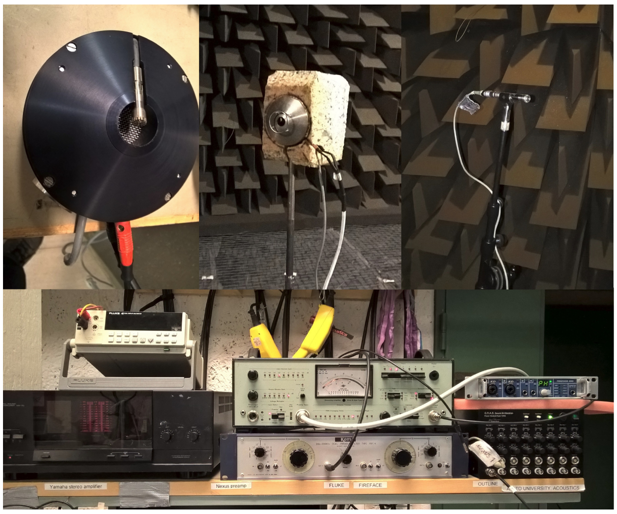

Figure 2: A snapshot of the recording devices. The top row shows B&K Type 4219 artificial voice along with G.R.A.S. Type 46AF free-field microphone. In the bottom row the amplifiers, A/D and D/A signal converters and filters are shown.

2.3. Analysis of the Measurements

We consider the recorded signal immediately after the voice generator as the reference and calculate the magnitude of the transfer function for each mask by dividing the magnitude spectrum of the signal measured by the far-field microphone by the

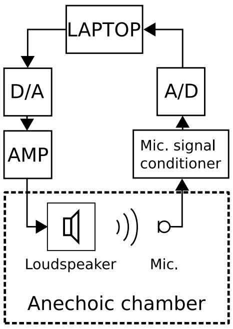

Figure 3: Block diagram of the recording setup for face mask tansfer function measurements.

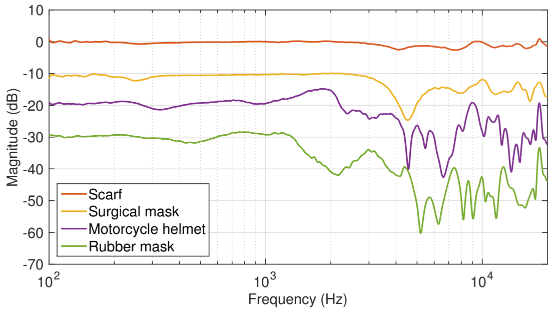

Figure 4: The measured magnitude of the mask transfer functions. The magnitudes of the transfer functions, $|H(f)|$, have been shifted by 10 dB for visual clarity.

spectrum of the reference signal. The magnitude of frequency response, $|H(f)|$, curves shown in Figure 4 are smoothed 4.2% per octave ($1/24$ octave). This smoothing is intended to increase the readability.

In line with the observation reported in [6], the measured magnitude frequency response of four masks in this study as shown in Figure 4 indicate that the most of the spectral distortion caused by wearing a face mask occurs in high frequency ranges rather than low frequencies. The red curve shows the effect of scarf on radiation. The attenuation is negligible below 3 kHz and has it maximum value of 2.5 dB at 4.1 kHz and 7.6 kHz. Otherwise the attenuation is less than 2 dB above 3 kHz.

The yellow curve shows the same for surgery mask. Again there is practically no effect below 3 kHz. At 4.5 kHz, a 14 dB dip is observed after which the attenuation stays around 5 dB. As it is also reported in [5], the high frequency loss can affect on the speech intelligibility. Magenta line representing the magnitude of transfer function for a motorcycle helmet shows an interesting phenomenon between 1 kHz and 2.2 kHz, where an increase in radiation takes place, peaking maximum of 5.2 dB at 1.8 kHz. A narrow dip with magnitude of 20 dB at 4.6 kHz is visible and a highly varying attenuation around 10 dB follows for the rest of the band, including a maximum attenuation of 22.6 dB at 6.6 kHz.

The magnitude of transfer function for both motorcycle helmet and rubber mask rises at some frequency bands that can be attributed to increased radiation load or increased directivity due to the mask shape. The data for rubber mask (in green) shows a noticeable attenuation (of ca. 6 dB) between 1.5 kHz and 4.5 kHz that follows a larger attenuation of about 20 dB at higher frequencies with the most prominent dip of 30 dB at 5.2 kHz. The varying attenuation characteristics at higher frequencies result from the changed diffraction pattern of the source due to mask with complex geometry.

3. Speaker Recognition

Speaker recognition techniques experienced a fast paced evolution in the recent years [7]. After the introduction of i-vectors [8], as a compact representation of a speech utterance, it became the state-of-the-art technique in speaker recognition. The most recent developments in the field report successful integration of deep neural networks to speaker recognition framework [9]. For the sake of simplicity in our speaker recognition experiments, we stick to the well-known i-vector extraction with universal background model (UBM) [10] acoustic models and probabilistic linear discriminant analysis (PLDA) [11] for modeling speakers' space.

As it is usually the case in a real-world forensic speaker recognition application, we take a state-of-the-art recognition system [12, 13] *off-the-shelf* and perform experiments in two conditions. In the first experiment, we look at closed-set speaker identification performance when there is a mismatch in terms of face mask between enrolment and probe utterances. In the second experiment, we employ the same setup as in the first experiment, but we utilize the frequency response of the respective mask for an utterance and perform a simple channel equalization before acoustic feature extraction to equalize the effects of face mask on magnitude spectrum of speech frames.

This study assumes that a face mask acts as a linear time-invariant filter. This assumption implies that the feedback effect that a face mask has on speech production is ignored and left for future studies. Throughout our recording configuration reported in Section 2.2, a measurement of the magnitude of frequency response, $|H(f)|$, is obtained for each of the four face masks (see Section 2.3). Next, we perform channel equalization in the form of *direct inversion* [14] or so-called *zero-forcing equalizer* by applying the inverse of $|H(f)|$ filter for a recorded speech under that particular face mask. The inverse filter in the form of $1/|H(f)|$ is utilized after speech spectrum estimation in the chain of acoustic feature extraction. This is illustrated in Figure 5.

3.1. Experimental Setup

We use a *linear prediction* model of order $p = 20$ in order to arrive at short time spectrum of speech. The speech signal is segmented to frames of 30 msec length with a hop rate of 15 msec. Next, 19 Mel-frequency cepstral coefficients (MFCCs) are extracted and appended by frame energies. After RASTA filtering [15], Δ and $\Delta\Delta$ features are calculated to form 60-dimensional feature vectors. We apply feature warping [16] after removing the unvoiced parts of speech.

A gender-dependent UBM with 2048 components is trained using a subset of NIST SRE 2004–2006, Switchboard cellular phase 1 and 2, and the Fisher English corpora. The total variability space [8] of 400 dimensionality is trained with the same data as for UBM. After extracting i-vectors, we used linear discriminant analysis (LDA) projection to enhance separability of speakers and reduce the i-vectors dimension to 200. Next, we remove the global mean of i-vectors, perform whitening using within-class covariance normalization (WCCN) [17] and apply spherical normalization of the resulting i-vectors [18]. At the end, a Gaussian PLDA [19] modelling is utilized.

We use the second session of the reading speech material

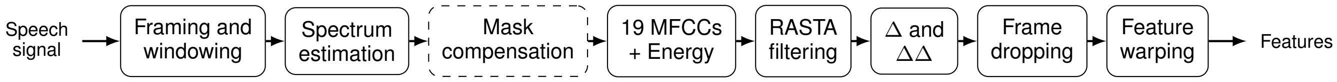

Figure 5: Front-end processing for acoustic feature extraction. Mask compensation is only applied for processing speech under a face mask by employing the inverse filtering in the form of $1/|H(f)|$ using the respective magnitude of transfer function $H(f)$ measured for that face mask.

Table 1: Closed-set correct speaker identification rate reported in percentage (%) 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 mask	Helmet	Rubber mask	Surgeon mask	Scarf
No mask	**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
Scarf	90.3	85.7	82.5	94.9	**97.0**
Number of tests	793	574	543	626	568

of each speaker as for enrolment. We make template i-vectors for speech under different face masks as well as no face mask in order to test the recognition system behaviour in comparing i-vectors originating from speech under different face masks. The i-vectors corresponding to the utterances recorded using each of the three microphones are extracted separately. In enrolment phase, these three i-vectors are averaged out in order to reduce the sensitivity of the recognition system to the mismatch caused by utterances collected from a different microphone in training and test phases. The training side, on average, includes around 25 seconds of active speech.

We designed the test protocol in such a way that recordings of spontaneous speech for all of the speakers are segmented into chunks of feature vector streams each correspond to acoustic features extracted from 2.5 sec of active speech. The segments are non-overlapping. In such a way, we acquire several hundreds of i-vectors respectively for utterances under different masks where all three microphones and both of the recording sessions are utilized. The total number of available i-vectors for each scenario of face mask is provided in the last row of Table 1. The difference in the number of test i-vectors for each of the face masks is a result of the application of a simple utterance-level energy-based voice activity detector. In addition, durations of spontaneous speech recordings under face mask are noticeably shorter than the corresponding recordings with no face mask present. The closed-set speaker identification experiments are designed to represent no cross-gender trial which means that each test segment is evaluated against 4 speaker templates. The top scoring speaker is identified as the underlying speaker and compared to the ground truth.

3.2. Experimental Results

The closed-set speaker identification results are presented in Tables 1 and 2. In general, when the enrollment and probe utterances are from the speech under the same mask, a high correct identification rate is observed. A degradation in recognition performance occurs when the template and test segment are from different masks. The results in Table 2 suggest that by introduction of direct inversion in the feature extraction chain for processing of speech under face cover, in most of the cases the deficit in recognition performance caused by mismatch of face mask is reduced.

Table 2: Closed-set correct speaker identification rate reported in percentage (%) with the same interpretations as in Table 1. In this experiment, the face mask effect is compensated by application of direct inversion with respect to the magnitude transfer function of a face mask. The type of face mask present in a speech utterance is considered to be known a priori.

Face cover	No mask	Helmet	Rubber mask	Surgeon mask	Scarf
No mask	**95.2**	95.3	91.2	93.8	93.5
Helmet	92.1	**98.1**	89.4	90.0	87.4
Rubber mask	91.4	95.1	**98.2**	94.4	92.5
Surgeon mask	95.4	95.7	93.3	**97.9**	95.8
Scarf	92.2	88.7	86.1	94.3	**97.5**

The recognition results presented in the first row of Table 2 are promising. When the enrolment utterances are from speech under no mask and direct inversion is applied in acoustic feature extraction of test utterances from different face masks, the recognition system presents marginal improved recognition rates except for surgeon mask case. Minor improvements in speaker identification can also be seen for the matched cases.

4. Conclusions

In line with our earlier study on collecting speech under face mask, we set up a new recording configuration to measure transfer functions of four different face masks. These measurements are targeted to separate the passive and active effects of wearing a face mask on recorded speech. Our measurements indicate severe high frequency distortion of speech caused by wearing a rubber mask and motorcycle helmet. Because band limited speech signals are used in the speaker recognition system, the high frequency distortion effects caused by wearing face masks, are not strongly pronounced in speaker recognition rates. Nevertheless, by introducing a simple channel equalization in the form of applying a zero-forcing equalizer in the magnitude spectrum domain, marginal improvements in correct speaker identification rate are observed. In future studies, a more accurate treatment of the face mask transfer function needs to be addressed. In doing the direct inversion, we envisioned that such compensation for feature extraction can be applied only on the magnitude spectrum hence ignoring the phase response of the measured transfer function.

5. Acknowledgements

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

6. References

[1] R. D. Mori, F. Bechet, D. Hakkani-Tur, M. McTear, G. Riccardi, and G. Tur. Spoken language understanding. *IEEE Signal Processing Magazine*, 25(3):50–58, 2008.

[2] M. T. Mills and N. G. Bourbakis. Graph-based methods for natural language processing and understanding; a survey and analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(1):59–71, 2014.

[3] R. Saeidi, T. Niemi, H. Karppelin, J. Pohjalainen, T. Kinnunen, and P. Alku. Speaker recognition for speech under face cover. In *Proc. Interspeech 2015*, pages 1012–1016, 2015.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] F. Richardson, D. Reynolds, and N. Dehak. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, 22(10):1671–1675, 2015.

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

[11] 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.

[12] 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.

[13] 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.

[14] J. Benesty, M. M. Sondhi, and Y. Huang, editors. *Springer Handbook of Speech Processing*, chapter Channel Inversion and Equalization. Springer, 2008.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson. Multi-condition 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.