Robust Speaker Recognition with Combined Use of Acoustic and Throat Microphone Speech

Md Sahidullah[1]*, Rosa Gonzalez Hautamäki*[1]*, Dennis Alexander Lehmann Thomsen*[2]*,*
Tomi Kinnunen[1]*, Zheng-Hua Tan*[2]*, Ville Hautamäki*[1]*, Robert Parts*[3]*, Martti Pitkänen*[3]

[1]Speech and Image Processing Unit, School of Computing, University of Eastern Finland, Finland
[2]Signal and Information Processing, Department of Electronic Systems, Aalborg University, Denmark
[3]Aplcomp Oy, Helsinki, Finland

sahid@cs.uef.fi, rgonza@cs.uef.fi, dalth@es.aau.dk, tkinnu@cs.joensuu.fi
zt@es.aau.dk, villeh@cs.uef.fi, parts@neti.ee, martti@aplcomp.fi

Abstract

Accuracy of automatic speaker recognition (ASV) systems degrades severely in the presence of background noise. In this paper, we study the use of additional side information provided by a body-conducted sensor, throat microphone. Throat microphone signal is much less affected by background noise in comparison to acoustic microphone signal. This makes throat microphones potentially useful for feature extraction or speech activity detection. This paper, firstly, proposes a new prototype system for simultaneous data-acquisition of acoustic and throat microphone signals. Secondly, we study the use of this additional information for both speech activity detection, feature extraction and fusion of the acoustic and throat microphone signals. We collect a pilot database consisting of 38 subjects including both clean and noisy sessions. We carry out speaker verification experiments using Gaussian mixture model with universal background model (GMM-UBM) and i-vector based system. We have achieved considerable improvement in recognition accuracy even in highly degraded conditions.

Index Terms: Speaker recognition, Noisy condition, Throat microphone, Fusion.

1. Introduction

Speech-based authentication systems with automatic speaker verification (ASV) technology provide a low-cost and flexible biometric solution to access control [1]. It yields high recognition accuracy when speech data from acoustically matched conditions are used in both enrollment and verification. But the performance of ASV systems degrades dramatically in the presence of channel or environmental mismatch. The channel effects are well-compensated with the help of advanced channel effect reduction techniques on i-vector representation [2]. Mismatch due to additive environmental noise, however, remains challenging [3]. Much research has been devoted on improving the accuracy of ASV systems in the presence of additive noise. Speech enhancement techniques are used to reduce the effect of noise in speech [4]. Robust features are also proposed which are invariant to certain variations in speech signal [5, 6]. Fusion of several sub-systems with different features and different speech activity detection (SAD) methods have been applied to noise-robust speaker identification [7]. The investigated features range from cepstral, cortical to prosodic features and the explored SAD methods include both supervised and unsupervised ones. Further, model domain techniques such as parallel model combination and multi-condition training [8] are also studied. Nevertheless, all these techniques are most effective for specific or known types of noises, none being a universal solution for environments with unpredictable noises.

Differing from the existing single-channel solutions to noise-robustness, we study the use of multiple microphones to record speech. As opposed to existing multi-channel speech acquisition methods where several identical microphones are placed in different physical locations from the signal source, we collect signals using two different kinds of microphones. The first one is a conventional *acoustic microphone* (AM) that uses acoustic transducer to convert sound energy into electrical signals. The second one is a skin-attached non-acoustic sensor [9], *throat microphone* (TM) or *laryngophone*, that picks up vocal fold vibration into signals. As the TM absorbs vibrations directly from the throat, the signal is immensely robust even in severely degraded environmental conditions.

Known applications of throat microphones includes speech communication in military, aviation, law enforcement, sports or other similar scenarios where the subjects wear helmets, masks or full-face breathing apparatuses. This study is a part of an ongoing OCTAVE project[1] that transfers ASV technology to novel logical and physical access control applications including demanding acoustic environments. Even if the use of throat microphones has been widely explored in speech processing (reviewed in Section 2), their use in ASV appears surprisingly small; most of the known work on throat microphone based speaker recognition are done by one research group [10, 11, 12]. Other than this, it is used to address the limitation of ASV system's ability to handle whisper speech [13]. Shahina *et al.* have also studied this for language identification task [14].

This study presents a work-in-progress report on setting up a practical data collection platform for the OCTAVE project and reports findings on a small pilot corpus (38 speakers) to study feasibility of throat mics in ASV. We aim at independent validation of some of the earlier findings concerning the potential of TM signals. We expand the prior work in TM-based speaker verification (reviewed in Section 2.2) with the inclusion of more modern ASV systems (GMM-UBM [15] and i-vectors [2]) and fusion of throat and acoustic microphone features. One of the challenges that we face is unavailability of throat-mic data to train universal background models (UBMs) and other system components. To this end, we propose to use the classic maximum a posteriori (MAP) recipe to adapt the UBMs.

[1] https://www.octave-project.eu/

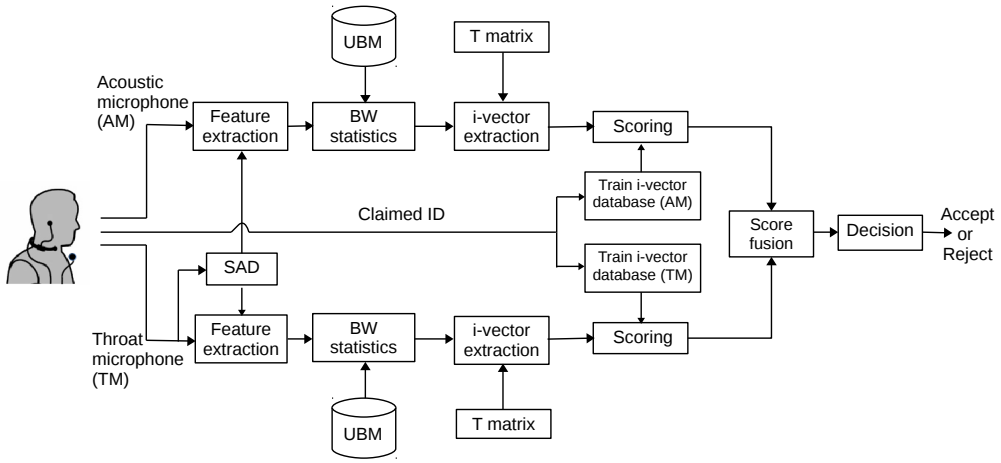

Figure 1: Block diagram of speaker recognition system with dual microphone setup.

2. Related work

2.1. General Use of Throat Mics in Speech Processing

Techniques for robust speech recognition and SAD in highly noisy, non-stationary environments using several heterogeneous sensors have been explored in [16]. The hardware prototypes integrate AM with bone microphones and TM among others into headsets. Another wearable recording system used in [17] integrates a close-talking, a monophonic far-field, and a TM in addition to a 4-channel far-field microphone array to create a multi-channel database for speech recognition. The database contains spontaneous, conversational, and partly noisy speech for seven synchronized audio channels. In [18], a technique for estimating clean acoustic speech features by combining TM and AM recordings using a probabilistic optimum filter (POF) mapping is proposed for speech recognition. Since TM speech is relatively more robust to environmental variations, it can be used to detect speech regions. In [19, 20], this idea is used and improved recognition performance is obtained when TM speech is used for SAD. In [21], various adaptation methods such as maximum likelihood linear regression and sigmoid low-pass filtering are studied in the context of whispered speech recognition with the help of TM signal. In [22] and [23], TM is used for voice quality assessment.

Throat microphone signals are also used in speech enhancement. They are also used as clean reference signal for the objective performance measure for speech enhancement algorithms [24]. Though TM speech is more robust in presence of ambient noise but its intelligibility is lower than AM speech. For this reason, sometimes the quality of TM speech also needs to be improved. In [25], a phone-dependent Gaussian mixture model-based statistical mapping have been explored for this purpose to construct probabilistic mappings between TM and AM speech signals. Various spectral mapping techniques are compared in [26] for the enhancement of TM speech.

2.2. Use in Automatic Speaker Verification

In spite of its high robustness against environmental noise, surprisingly throat microphone is not much studied for speaker recognition. This is possibly because in recent past, most of the advancements in speaker recognition research are made with NIST SRE where text-independent speaker recognition problem with telephone channel conversational speech of longer duration (approximately 5 min) is the main concern. The major applications of this task are in forensic and surveillance. However, text-dependent speaker recognition is more suitable for access control, both for physical (e.g., entrance to a protected area) and logical (e.g., tele-banking, secure service over internet, etc.) access. In this work, the main motivation for exploring throat microphone in this context is due to its inherent robustness in realistic noisy conditions. Our work is a part of the ongoing OCTAVE project for developing real-time voice biometrics for physical and logical access in a highly degraded environment (e.g., airport ground). Our goal is to explore the use of throat microphone in such a degraded condition for speaker verification task.

The previous studies in throat microphone based speaker recognition used auto-associative neural network (AANN) for modeling target speakers [10, 11, 12]. Performance was evaluated for closed-set speaker identification task. In this work, we evaluate the performance of acoustic and throat microphone based speaker verification system for GMM-UBM and i-vector based speaker recognition. Moreover, since the signals are synchronized, we use SAD labels computed for TM for more accurate detection of speech segments, specially in noisy condition. Further, we use score fusion to combine the recognition scores of AM and TM system. The block diagram of overall speaker recognition system is depicted in Fig. 1.

3. Collection of Speech Corpus

Dual microphone speech data for text-dependent speaker recognition are collected using a web-based user interface from three different sites. All the recordings are made using a similar model of AM and TM. We use Scarlett 2i2 USB 2.0 audio interface manufactured by Focusrite for recording two channels simultaneously into a stereo file[2]. The TM signals are recorded in left channel and the AM in the other as shown in Fig. 2. Voice samples are recorded using a web interface with the Microsoft

[2] http://us.focusrite.com/usb-audio-interfaces/scarlett-2i2

Edge web browser, and data are stored in remote server. The sampling frequency for original recordings is set at 44.1 kHz. These phrases (as listed in Appendix A) are same as the common phrases used in the Part I sub-condition of the on-going RedDots project on text-dependent speaker recognition [27]. Voice samples from five different sessions are collected for each subject, out of which one is noisy and the rests are from relatively silent condition such as common office environment. In both cases, we use clean speech for training. Therefore, the test case with clean speech is called as *matched condition* whereas test with noisy session is referred here as *mismatched condition*.

We collect data from 38 subjects. Speaker recognition experiments are conducted on a database of 30 speakers (23 male and 7 female) and speech-data from the remaining eight speakers are used in domain adaptation. We use speech signals from three different clean sessions for training text-dependent speaker model. For each speaker, 10 different models are created for each common phrases separately. Total 300 target models are trained for 30 speakers. Voice samples from the remaining two sessions are used in test. In order to evaluate the performance in clean and noisy condition, we use two sessions as two separate test conditions. Trials are designed such that the texts or *spoken-content* of target model and test segment are identical. In each condition, there are 9000 trials and out of them 300 are *genuine* or *target* while rests 8700 are *impostor* or *nontarget*.

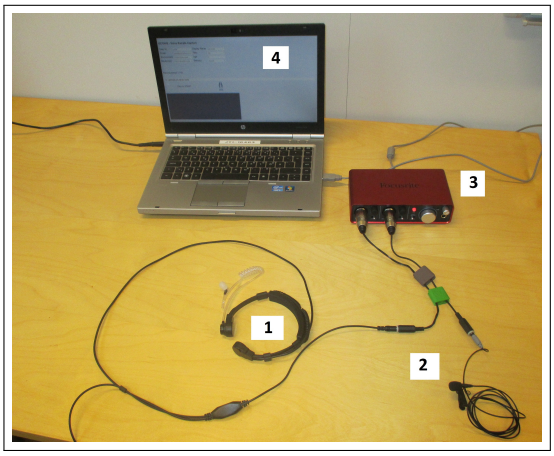

Figure 2: Data collection setup with (1) Throat microphone, (2) Close-talk or acoustic microphone, (3) USB audio interface, (4) Web-based user interface.

4. Experimental Setup

4.1. Description of Features and Classifiers

The recorded utterances have a sampling rate of 44.1 kHz. In order to use it with suitable UBM data, we down-sample them at 16 kHz. Then we compute mel-frequency cepstral coefficients (MFCCs) as spectral features for representing both AM and TM signals. The MFCCs are extracted from speech frames of duration 20 ms with 50% overlap. We use 20 filters in mel scale to compute 20 coefficients including the energy. Then we perform RelAtive SpecTrAl (RASTA) processing to reduce the linear channel effects [28]. We obtain 60-dimensional features after augmenting delta and double-delta coefficients computed with a window of three frames. Finally, we drop the non-speech frames using an energy-based SAD [1].

Experiments are performed both with GMM-UBM [15] and i-Vector system [2]. In both cases, the UBMs are trained in a gender-independent manner from all 6300 speech files of the TIMIT corpus. We choose TIMIT as it has microphone speech of good quality (16 kHz) in English language similar to the AM speech to be used in the evaluation. UBMs are trained with 512 mixtures using 10 iterations of the expectation-maximization (EM) algorithm. Target models are created using a *maximum-a-posteriori* (MAP) algorithm with relevance factor of 14. The i-vector extractor (i.e., T-matrix) is trained for 400 total factors with five iterations of EM. We compute recognition score as log-likelihood ratio and cosine similarity for GMM-UBM and i-vector system, respectively.

4.2. Performance Evaluation

We use equal error rate (EER) and minimum detection cost function (minDCF) to assess speaker recognition accuracy. EER is calculated in the detection threshold when the false alarm (P_{fa}) and the false rejection rate (P_{miss}) are equal, whereas minDCF is the minimum of

$$C(\Theta) = w_{\text{miss}} \times P_{\text{miss}}(\Theta) + w_{\text{fa}} \times P_{\text{fa}}(\Theta),$$

where Θ is the detection threshold. Here, w_{miss} and w_{fa} are weights for the miss and false alarm rate. We set $w_{\text{miss}} = 0.01$ and $w_{\text{fa}} = 0.99$ following the NIST evaluation plan.

5. Results

5.1. Speaker Recognition with Individual Microphones

In the first experiment, we compute speaker verification accuracy using GMM-UBM and i-vector system. The results are shown in Table 1 for signals from individual microphones in matched and mismatched conditions. For both the systems, we observe that AM-based system performs better than TM-based in matched condition. However, in mismatched condition speaker recognition systems using TM speech outperforms AM-based systems. This agrees with the previous studies in this field [11]. Further, we observe the similar trend with i-vector based system. We also note that state-of-the i-vector system gives relatively poor performance as compared to classical GMM-UBM system. This is most likely due to the lack of suitable development data for i-vector extractor as the performance of i-vector system is highly data-dependent [2]. Besides, the speech segments are very short in duration for which i-vector system is not very efficient [29].

Table 1: *Text-dependent speaker recognition results in terms of EER (in %) and minDCF (× 100) with single microphone for matched and mismatched conditions using GMM-UBM system. Results are shown for two separate microphones (AM and TM).*

Classifier	Condition	AM		TM	
		EER	minDCF	EER	minDCF
GMM-UBM	Matched	**0.06**	**0.03**	2.72	1.48
	Mismatched	10.33	4.79	**6.67**	**2.67**
i-vector	Matched	**1.33**	**0.39**	4.31	1.70
	Mismatched	12.67	4.67	**9.00**	**3.23**

5.2. Domain Adaptation with Limited In-domain Data

The results in Table 1 were obtained by using a UBM and T-matrix both trained on TIMIT. Though TIMIT is a good choice as compared to any NIST corpus, it is not the most appropriate as the speech files of the current database are phonetically constrained. On the other hand, we do not have any publicly avail-

able throat microphone corpus to use with TM speech effectively in speaker modeling. This dataset handicap can be partly solved by applying *domain adaptation* [30]. For this purpose, we use a limited amount of speech-data that are collected using our setup. Speech features from eight speakers (6 male and 2 female) are used to adapt the pre-trained UBM using relevance MAP. We separately create two different adapted UBMs: one for AM speech and another for TM with corresponding speech features. Only speech files of the four 'clean' sessions are considered. Subsequently for T-matrix estimation, we use those 320 ($8 \times 4 \times 10$) sentences as in-domain data in addition to the existing 6300 segments of TIMIT. The results are shown in Table 2. We have found that utilization of in-domain data considerably reduces recognition error rates in most cases. Comparing the results of Table 1 and Table 2, we further observe that the relative improvements for TM-based systems are considerably higher than the improvements for AM-based system.

Table 2: *Same as Table 1 but for domain adaption with limited in-domain data.*

Classifier	Condition	AM		TM	
		EER	minDCF	EER	minDCF
GMM-UBM	Matched	**0.33**	**0.21**	1.67	0.93
	Mismatched	8.67	4.05	**4.40**	**1.88**
i-vector	Matched	**0.67**	**0.33**	2.00	1.06
	Mismatched	9.72	3.80	**5.00**	**1.94**

5.3. SAD Using Throat Microphone Speech

In previous cases, the speech frames are detected by applying an energy-based SAD on the respective signals. Now we perform experiments with AM-based system where SAD labels are generated using TM speech as they are more robust to noise. These results are shown in Table 3. In comparison to the results described in Table 2, the improvement is considerable for mismatched condition. For example, in GMM-UBM system, EER and minDCF drop from 8.67% and 0.0405 to 4.67% and 0.0199, respectively.

Table 3: *Text-dependent speaker recognition results in terms of EER (in %) and minDCF ($\times$ 100) with AM speech whereas the energy SAD labels are obtained from AM and TM signal, respectively.*

Classifier	Condition	AM-based SAD		TM-based SAD	
		EER	minDCF	EER	minDCF
GMM-UBM	Matched	0.33	0.21	**0.03**	**0.07**
	Mismatched	8.67	4.05	**4.67**	**1.99**
i-vector	Matched	0.67	0.33	**0.33**	**0.13**
	Mismatched	9.72	3.80	**7.67**	**2.48**

5.4. Score Fusion of AM and TM Systems

Finally, fusion is performed to combine two different microphone based systems. We apply equal weight based score fusion. The other possibility was to use feature-level fusion or so-called input fusion, we can not apply it here due to lack of suitable parallel data for joint training of AM and TM features. The results of combined system are shown in Table 4. In comparison with the results of obtained with single microphone based system as shown in Table 2 (TM) and Table 3, we observe significant improvement in all cases except for GMM-UBM system in matched condition. Notable improvement is observed for mismatched condition for both GMM-UBM and i-Vector system.

Table 4: *Text-dependent speaker recognition results in terms of EER (in %) and minDCF ($\times$ 100) for equal weighted score level fusion.*

Classifier	Condition	Score Fusion	
		EER	minDCF
GMM-UBM	Matched	0.33	0.07
	Mismatched	1.67	0.67
i-vector	Matched	0.33	0.07
	Mismatched	1.71	0.81

6. Conclusions

We developed a dual channel, acoustic and throat microphone, speech collection system. The developed system was used to collect a preliminary corpus of 38 speakers, where one session was recorded in noisy conditions and the other four sessions in an office environment. In the experiments, we demonstrated that speech recorded via throat microphones is more robust against additive noise mismatch than recordings made using conventional acoustical microphones, as was expected. We also showed that using the TM recording for SAD computation clearly improves the speaker recognition performance. Score fusion of systems based on AM and TM recordings are able to improve the EER from 6.67% to 1.67% for the mismatched case. As a future work, we plan to extend the number of speakers for our corpus and investigate further how to utilize TM recordings in conjunction with AM recordings in speaker recognition.

7. Acknowledgements

We are thankful to Mr. Amir Hossein Poorjam and Mr. Ivan Kukanov for their help in data collection at UEF. The paper reflects some results from the OCTAVE Project (#647850), funded by the Research European Agency (REA) of the European Commission, in its framework programme Horizon 2020. The views expressed in this paper are those of the authors and do not engage any official position of the European Commission.

Appendix A

The following ten common sentences/phrases are used as text-material for all the subjects:

1. A watched pot never boils.
2. Actions speak louder than words.
3. Artificial intelligence is for real.
4. Birthday parties have cupcakes and ice cream.
5. Jealousy has twentytwenty vision.
6. My voice is my password.
7. Necessity is the mother of invention.
8. OK Google.
9. Only lawyers love millionaires.
10. There's no such thing as a free lunch.

8. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12 – 40, 2010.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[3] M.I. Mandasari, M. McLaren, and D.A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *ICASSP*, March 2012, pp. 4249–4252.

[4] P.C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[5] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *In 2001: A Speaker Odyssey - The Speaker Recognition Workshop*, 2001.

[6] S.O. Sadjadi and J.H.L. Hansen, "Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification," *Speech Communication*, vol. 72, pp. 138–148, 2015.

[7] O. Plchot, S. Matsoukas, P. Matejka, N. Dehak, J. Ma, S. Cumani, O. Glembek, H. Hermansky, S.H. Mallidi, N. Mesgarani, R. Schwartz, M. Soufifar, Z.-H. Tan, S. Thomas, B. Zhang, and X. Zho, "Developing a speaker identification system for the DARPA RATS project.," in *ICASSP*, 2013, pp. 6768–6772.

[8] J. Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds, "Robust speaker recognition in noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1711–1723, 2007.

[9] S.A. Patil and J.H.L. Hansen, "The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification," *Speech Communication*, vol. 52, no. 4, pp. 327 – 340, 2010.

[10] M.A. Marx, G. Vinoth, A. Shahina, and A.N. Khan, "Throat microphone speech corpus for speaker recognition," *MES Journal of Technology and Management*, pp. 16–20, 2009.

[11] A. Shahina, B. Yegnanarayana, and M.R. Kesheorey, "Throat microphone signal for speaker recognition," in *Proceedings of ICSLP*, 2004.

[12] N. Mubeen, A. Shahina, A.N. Khan, and G. Vinoth, "Combining spectral features of standard and throat microphones for speaker identification," in *Proceedings of International Conference on Recent Trends In Information Technology*. IEEE, 2012, pp. 119–122.

[13] W. Jin, S.S. Jou, and T. Schultz, "Whispering speaker identification," in *Multimedia and Expo, 2007 IEEE International Conference on*, 2007, pp. 1027–1030.

[14] A. Shahina and B. Yegnanarayana, "Language identification in noisy environments using throat microphone signals," in *Proceedings of International Conference on Intelligent Sensing and Information Processing*, 2005, pp. 400–403.

[15] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[16] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, "Multi-sensory microphones for robust speech detection, enhancement and recognition," in *Proceedings of ICASSP*, 2004, vol. 3, pp. iii–781–4 vol.3.

[17] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, "The design and collection of COSINE, a multimicrophone in situ speech corpus recorded in noisy environments," *Computer Speech & Language*, vol. 26, no. 1, pp. 52 – 66, 2012.

[18] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *Signal Processing Letters, IEEE*, vol. 10, no. 3, pp. 72–74, 2003.

[19] T. Dekens, Y. Patsis, W. Verhelst, F. Beaugendre, and F. Capman, "A multi-sensor speech database with applications towards robust speech processing in hostile environments.," in *LREC*, 2008.

[20] T. Dekens, W. Verhelst, F. Capman, and F. Beaugendre, "Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection," in *Proceedings of EUSIPCO*, 2010, pp. 23–27.

[21] S. Jou, T. Schultz, and A. Waibel, "Adaptation for soft whisper recognition using a throat microphone.," in *INTERSPEECH*, 2004.

[22] V. Uloza, E. Padervinskis, I. Uloziene, V. Saferis, and A. Verikas, "Combined use of standard and throat microphones for measurement of acoustic voice parameters and voice categorization," *Journal of Voice*, vol. 29, no. 5, pp. 552 – 559, 2015.

[23] F. Bozzoli and F. Angelo, "Measurement of active speech level inside cars using throat-activated microphone," in *Audio Engineering Society Convention 116*. Audio Engineering Society, 2004.

[24] S. Ntalampiras, T. Ganchev, I. Potamitis, and N. Fakotakis, "Objective comparison of speech enhancement algorithms under real world conditions," in *Proceedings of the 1st international conference on PErvasive Technologies Related to Assistive Environments*. ACM, 2008, p. 34.

[25] M.A.T. Turan and E. Erzin, "Source and filter estimation for throat-microphone speech enhancement," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 24, no. 2, pp. 265–275, 2016.

[26] K. Vijayan and K.S.R. Murty, "Comparative study of spectral mapping techniques for enhancement of throat microphone speech," in *Communications (NCC), 2014 Twentieth National Conference on*, 2014, pp. 1–5.

[27] K.A. Lee, A. Larcher, W. Wang, P. Kenny, N. Brummer, D.A. van Leeuwen, H. Aronowitz, M. Kockmann, C. Vaquero, B. Ma, H. Li, T. Stafylakis, J. Alam, A. Swart, and J. Perez, "The RedDots data collection for speaker recognition," in *Proceedings of Interspeech*, 2015.

[28] H. Hermansky and N. Morgan, "RASTA processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[29] A. Poddar, M. Sahidullah, and G. Saha, "Performance comparison of speaker recognition systems in presence of duration variability," in *Proceedings of IEEE INDICON*, 2015.

[30] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for I-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 4047–4051.