THE EFFECT OF NOISE ON MODERN AUTOMATIC SPEAKER RECOGNITION SYSTEMS

Miranti Indar Mandasari, Mitchell McLaren and David A. van Leeuwen

Centre for Language and Speech Technology
Radboud University Nijmegen, The Netherlands
{m.mandasari, m.mclaren, d.vanleeuwen}@let.ru.nl

ABSTRACT

Motivated by the application of speaker recognition in forensic area, this paper presents a study on noise robustness of several automatic speaker recognition system approaches, ranging from simple dot-scoring and a standard i-vector system with cosine distance scoring to a state-of-the-art i-vector Probabilistic Linear Discriminant Analysis (PLDA) system. Using the recent NIST 2010 Speaker Recognition Evaluation (SRE) data, the systems are analyzed in added noise conditions with a range of signal to noise ratios. Various experiments were conducted to study the influence of the noise on the speech activity detection and Wiener filtering in the front-end of the system.

Index Terms— speaker recognition, i-vector, PLDA, noise conditions, forensics

1. INTRODUCTION

Traditionally, automatic speaker recognition systems are developed and tested in a clean speech environment. However, in many applications of speaker recognition, the speech samples provided to the system may suffer from some background noise. In noisy conditions, the performance of speaker recognition system is expected to drop, especially in a low signal-to-noise ratio (SNR) situation [1, 2, 3].

In the last decade, some research has studied the behaviour of speaker recognition systems in noisy speech conditions [1], and a number of techniques has been proposed make speaker recognition systems more noise-robust [2, 3, 4]. However, to the best of our knowledge, there has been no research reported yet on the noise-robustness of the modern i-vector speaker recognition approach that has recently become mainstream in this field. Encouraged by many reports on good performance offered by the i-vector system in clean speech conditions, we are interested to see how the i-vector system behaves when the SNR ratio becomes less favorable.

As a continuation of previous work on the system evaluation in short duration conditions [5], and motivated by the application of automatic speaker recognition to forensics, this paper presents a study on i-vector based speaker recognition systems in noisy speech conditions. In forensic cases, a speaker recognition system can be utilized for preparing legal evidence to the court by processing a speech sample recorded from the crime scene (speech trace) and comparing this to speech material from the suspect. In a specific forensic case scenario, the speech trace can be corrupted by noise or other various forms of deterioration to the signal. Often, an incriminating recording is made using a telephone in a car (engine, wheel rolling and wind noise) or in a café or public place (voice babble and music background noise). The effect of environmental noise on the recording is at least twofold. On one hand, the noise is added to the speech signal at the transducer, leading to a lower SNR at the receiver's end. Additionally, this may reduce the coding efficiency in case a speech compression system is used, e.g., with GSM calls, which may lead to non-linear distortions in the speech signal. On the other hand, the Lombard reflex in human speakers will cause the speaker to change the vocal effort, in an unconscious effort to increase the SNR at the receiver's end (thus counteracting the first mentioned effect) but simultaneously changing their voice's spectral characteristics. An attempt to study the latter effect has been carried out in one of the NIST SRE-2010 evaluation conditions where the vocal effort of speaker's was manipulated without inducing additional noise in the recording [6]. In this study we concentrate on the effect of the lowering of the SNR through added noise, ignoring the effects of the Lombard reflex at this point.

Using the most recent NIST Speaker Recognition Evaluation data (SRE 2010) [6], a set of experiments was carried out in order to study the behavior of several state-of-the-art speaker recognition systems in noisy speech conditions. Perhaps one of the most undervalued components of any speaker recognition system is the Speech Activity Detection (SAD) algorithm. This process acts at the very front-end of the processing chain, and is likely to be influenced by added noise. In a first experiment we therefore look at the influence of noise on a full speaker recognition system compared to one where the SAD component is given an "oracle" clean speech version of the signal. In a second experiment we compare three different systems with different forms of channel compensation and speaker modeling: a simple dot-scoring system with channel compensation, a standard i-vector system using LDA followed by WCCN and cosine distance scoring, and finally a state-of-the-art PLDA i-vector system. In a third experiment we study whether our systems favor "matched" noise conditions for the training and test segment, or whether they perform better if any of the two segments contains clean speech. In a final experiment, we investigate whether a Wiener filter at the front-end can alleviate some of the drop in performance due to added noise.

The paper is structured as follows. Section 2 details the baseline speaker recognition systems presented in this paper. Section 3 defines databases used for our experiments, and how our noisy speech database was generated. The experiment results and analysis are given in Section 4.

2. BASELINE SPEAKER RECOGNITION SYSTEMS

2.1. Feature extraction and UBM training

All systems use the same feature extraction stage. We extract 20 MFCCs (including C_0) using a 20 ms analysis window with 10 ms overlap. Augmented with delta and double-delta coefficients, we form 60 dimension features. A two-Gaussian energy-based SAD algorithm (see Section 2.2) discards the silence frames, and then each

This research was funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 238803.

feature dimension is feature-warped [7] using a 5-second analysis window. Gender-conditioned, 2048-component UBMs were trained using telephony speech from Switchboard II: Phase 3, Switchboard Cellular (1&2), Fisher English and NIST SRE 2004–2006 corpora.

2.2. Speech Activity Detection (SAD)

An energy-based SAD algorithm [8, 9] was used to determine the speech frames in an audio signal. This process involved training a two-component Gaussian Mixture Model (GMM) from the log-energy of the audio signal. Samples below the mean of the non-speech Gaussian were iteratively removed and the GMM retrained until the variance of the speech Gaussian was less than fives times that of the non-speech Gaussian. Speech frames were found by applying a threshold of 1.3 standard deviations below the speech Gaussian mean to the log-energy signal. The audio was considered to contain no speech if no samples remained or the difference between Gaussian means was less than 4.

2.3. Dot-scoring system

The first reference system was based on a fast, linear scorer to approximate GMM likelihoods using a simple inner product of a model and a test vector in a similar manner to SDV's NIST SRE-2008 submission [10]. This system reflects the behavior of the more traditional GMM-UBM approach. Techniques that were incorporated to improve robustness of the dot-scoring system included ZT-score normalisation and channel compensation using 50 dimensions estimated from the NIST SRE 2004 and Switchboard data sets.

2.4. I-vector extraction and LDA system

The i-vector speaker recognition system follows the framework proposed in [11]. An i-vector is a representation of speech utterance in a low dimensional Total Variability (TV) subspace in which both speaker and channel variation reside. I-vectors were extracted from a 400-dimensional TV space that was trained using the same collection of speech databases as used in UBM training. Our second reference system was used in previous work [5], and consists of a traditional i-vector approach, using 200-dimensional LDA for separating speakers followed by Within-Class Covariance Normalization (WCCN) [12] and utilizing normalized cosine distance scoring [13].

2.5. Probabilistic Linear Discriminant Analysis (PLDA)

The third and most state-of-the-art system is based on Probabilistic Linear Discriminant Analysis (PLDA) modelling, following a similar approach to [14]. PLDA is a probabilistic approach that models the distribution of i-vectors as a multivariate Gaussian. Thus, the model can be used to directly compute the likelihood ratio of two i-vectors originating from the same speaker versus originating from different speakers. Our system incorporates WCCN and i-vector length normalisation [15] prior to PLDA which has been shown to dramatically improve recognition performance. This system framework is capable of obtaining the best performance in the NIST 2010 SRE conditions involving telephone, interview and microphone speech.

2.6. Wiener filter

We investigate the use of Wiener filtering to help reduce noise in the audio signal. Wiener filtering has been shown to be beneficial to the performance in NIST conditions involving 'microphone' and 'interview' speech while not harming performance when applied to telephone speech [9]. The filter relies on SAD in order to obtain an estimate of the stationary noise spectrum. We use the same SAD algorithm as detailed in Section 2.2, however, a non-iterative process is used and the criteria to determine a silent audio segment were not applied to ensure that part of the signal was defined as noise. Further, the mean of the speech Gaussian was set as the noise threshold. The noise spectrum for Wiener filtering was estimated from frames with log-energy below this threshold.

3. DATABASES

3.1. Evaluation data

As evaluation data, we use NIST SRE-2010 with extended core trials as detailed in the evaluation protocol [6] to characterize the noise robustness of our speaker recognition systems. The experiments focus on the condition 5 which involves 416119 trials using telephony speech, as this condition is most suitable to simulate forensic cases for speaker recognition. In this paper, the system discrimination performance is reported in the terms of equal error rate (EER) and $C_{\text{det}}^{\text{min}}$, with $C_{\text{miss}} = 10, C_{\text{FA}} = 1$ and $P_{\text{tar}} = 0.01$.

3.2. Noise data

The NIST evaluation data is typically recorded in clean conditions. We therefore obtain our noisy speech segments by adding noise from a noise database, NOISEX [16], a commonly used noise database in speech technology. Since we are motivated by the type of speech encountered in forensic application, we use 'babble' and 'interior car noise'. These noise types are considered to be the among the most commonly encountered in the speech recording from forensic speech trace. In particular, babble noise is perhaps the hardest type noise to deal with in speaker recognition and other speech technology fields in general, because it has a very similar spectral characteristics as the speech from the target speaker. The babble noise in NOISEX was recorded in a canteen room with average sound level of 88 dBA. The interior car noise 'volvo' was recorded inside a car driving at a speed of 120 km/h on the asphalt road in rainy conditions.

3.3. Adding noise to audio signals

The noisy utterances used in this paper are made by adding a noise signal from NOISEX database into the utterance trials in the NIST SRE-2010 database for signal-to-noise ratios (SNR) 0, 5, 10, 15 and 20 dBA. Because the noise can be spectrally shaped quite differently from the signal it is masking, it does not suffice to simply use the linear level of the signal and noise. Rather, we need to spectrally weight the noise spectrum according to the signal it is masking in order to compute the level of the noise. To this end, we use the standard A-weighting of sound level, which is a filter based on the characteristics of the human ear and includes the spectral region where the majority of speech energy is observed. The dBA level of a speech utterance was computed only from active speech frames after having applied SAD. The noise files from NOISEX were truncated to the same length as the target speech utterance and a starting point defined based on the utterance name to induce randomness (thus preventing any potential algorithm development depending on the exact shape of the noise). The noise signal was then scaled to reach the desired SNR before it was added to the speech signal.

4. RESULTS

The following experiments aims to characterize the noise-robustness of modern speaker recognition systems. Firstly, the effect of noise of automatic SAD is analysed after which we compare the general system performance of our current PLDA i-vector system to the alternate classifiers. Experiments then analyze the use of both clean and noisy speech in trials to determine the effect of mismatch. Finally, Wiener filtering is employed in an attempt to improve system performance in the context of noisy speech.

SNR (dBA)	SAD Speech	Babble C_{det}^{min}	Babble EER(%)	Car C_{det}^{min}	Car EER(%)
0	Noisy	.0960	30.37	.0670	15.18
	Clean	.0991	31.54	.0266	5.20
10	Noisy	.0946	26.26	.0498	10.39
	Clean	.0602	13.92	.0187	3.60
20	Noisy	.0371	7.23	.0440	7.64
	Clean	.0276	5.30	.0164	3.14

Table 1. PLDA i-vector results comparing the effect of Noisy or Clean speech on speech activity detection for a range of SNRs.

4.1. Speech activity detection algorithm evaluation

In the first experiment, we analyze the noise-robustness of our SAD algorithm. To this end, we extracted features from a noisy speech sample (see Section 3.3) and then used the clean or the same noisy speech sample for the purpose of SAD. Table 1 presents the recognition results from the PLDA i-vector system comparing Clean and Noisy SAD options in babble and car noise conditions. It should be noted that in the presence of severe babble noise, our SAD algorithm failed to detect speech in a considerable number of utterances, namely due to the criteria enforced in Section 2.2. Scores from trials involving these utterances were excluded from the results reported in Table 1.

Comparing the two SAD options in Table 1 it can be observed that the $C_{\text{det}}^{\min}$ and EER when using SAD based on noisy speech was considerably higher than SAD based on clean speech. In the 0 dBA babble noise condition, however, similar performance was obtained between the two SAD conditions. It is expected that this occurred due to the exclusion of a number of trials for which features were not extracted leaving only features in which speech was more readily detectable due to a higher dynamic range in speech relative to the babble noise. These findings indicate that our SAD algorithm is not robust against noise, with result that it gives a large contribution to worsen the system performance in the noisy speech condition.

Based on the above analysis of our SAD algorithm, the experiments in Section 4.2 and Section 4.3 are presented for the SAD based on clean speech signal. This decision is motivated by a forensic scenario in which the manual segmentation or speech labeling may be performed when encountering noisy speech.

4.2. The effect of noise on modern recognizers

In this section we compare the behavior of three modern speaker recognition systems on various noisy speech conditions. Figure 1 depicts the EER performance metric of the state-of the art PLDA and LDA i-vector systems and the traditional dot-scoring system across a range of signal-to-noise ratios. All experiments were carried out using SAD based on clean speech during feature extraction.

In Figure 1, the solid lines correspond to the babble noise conditions, while the dashed lines correspond to the car noise conditions. It is clear from Figure 1 that all systems performed better in the presence of car noise as opposed to the more challenging babble noise. This can be explained in consequence of the fact that babble noise has same spectral shape as speech signal, which made this type of noise harder to deal with.

Focusing on the babble noise condition, the performance of the dot-scoring system dropped severely from the reference point even at a relatively high SNR of 20 dBA. In contrast, the i-vector systems offered a more gradual degradation in performance at the high SNRs. In the babble noise condition, both i-vector systems with PLDA and LDA classifiers were found to have same performance trends across all SNR levels, with the PLDA classifier consistently

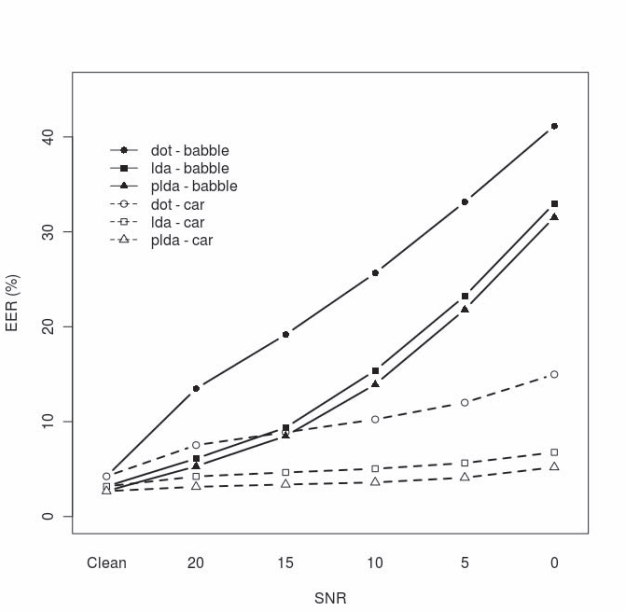

Fig. 1. System performance in terms of EER for dot-scoring, PLDA and LDA i-vector systems in Babble and Car noise conditions.

offering superior performance. The effect of car noise across all three systems was found to offer similar performance trends to those observed with babble noise, albeit to a lesser degree. In the presence of car noise, the relative EER drop of i-vector systems was approximately 10–20 % for every -5 dB step, compared to 40–60 % in the babble noise condition. It can observed from the plot that i-vector based speaker recognition systems are relatively robust to car noise while babble noise present a considerable problem to the system. This was not the case, however, for the dot-scoring system where it can be observed that the EER for both i-vector systems in the noisiest condition (0 dBA) was lower than the Dot-scorer with a SNR of 20 dBA. This analysis indicates that the noise robustness of automatic speaker recognition technology progresses along with improved algorithms and computational efficiency.

4.3. Mismatch noise conditions

In the previous section we presented results when both sides of a trial were degraded by noise. For clarity, we refer to this as a 'matched' noise condition. In this section, we present results on 'mismatched' noise conditions in which a clean speech signal was used for the train side and noisy speech for the test side. Motivation for mismatched evaluation comes from the forensic scenario in which a suspect interview can be recorded in a controlled environment while the conditions of the speech trace recording are typically uncontrolled. The EER from PLDA i-vector and the dot-scoring systems under matched and mismatched noise conditions are presented in Table 2.

Table 2 shows that the EER from the PLDA system for mismatched trials was consistently and considerably lower than for matched trials in both noise conditions. Similar trends were observed in the dot-scoring results despite the presumption that speaker recognition systems tend to perform worse in mismatch noise condition trials [3]. In fact, the relative performance improvement from mismatched conditions over matched conditions in the Dot-scorer was greater than that observed in the PLDA system. However, the PLDA system maintained a considerable improvement over dot-scoring performance. These results indicate that mismatched noise conditions do not adversely affect recognition performance. Rather, the presence of clean speech in one of the trial sides can significantly

	SNR	Babble		Car	
	(dBA)	Match	Mismatch	Match	Mismatch
PLDA	**0**	31.54	26.48	5.20	3.26
	10	13.92	9.29	3.60	2.59
	20	5.30	3.42	3.14	2.60
Dot	**0**	41.16	28.97	14.98	5.38
	10	25.65	12.39	10.24	4.30
	20	13.48	5.80	7.53	4.21

Table 2. EER (%) for matched and mismatched noise condition in trial sides for PLDA and dot-scoring (Dot) systems.

Wiener	Feature	Babble		Car	
SAD	SAD	C_{det}^{min}	EER(%)	C_{det}^{min}	EER(%)
N/A		.0602	13.92	.0187	3.60
Noisy	Clean	.0550	12.11	.0181	3.42
Clean		.0553	12.22	.0187	3.54
N/A	Noisy	.0946	26.26	.0498	10.39
Noisy	Filtered	.0994	23.28	.0180	3.59

Table 3. Results when using Clean and Noisy speech to obtain speech frames for Wiener filtering and feature selection at a SNR of 10 dBA for the PLDA i-vector system.

improve system performance.

4.4. Noise reduction through Wiener filtering

Results in previous sections illustrated the performance deterioration of the PLDA i-vector speaker recognition system in noisy speech conditions. In this section we investigate whether the straightforward approach of Wiener filtering can help reduce this deterioration. Given an estimation of noise component of an audio signal, Wiener filtering has the ability to reduce the perceived noise in a signal by removing the spectral average of the noise. As detailed in Section 2.6, we employ our SAD algorithm to obtain this noise estimation. Consequently, two independent SAD processes are involved in the front-end feature extraction process: Wiener filtering SAD and Feature SAD to select features corresponding to speech frames. We investigate the effect of noise on both of these SAD processes.

Table 3 details results from the PLDA i-vector system when evaluating noisy (SNR of 10 dBA) and subsequently Wiener-filtered speech. Note that 'N/A' indicates that Wiener filtering was not applied. In the case of Feature SAD using clean speech, it was observed that Wiener filtering offered a marginal improvement over un-filtered noisy speech in the babble noise scenario, irrespective of the speech used for Wiener SAD. In the case of car noise, however, the effect of Wiener filtering was negligible. Results using Feature SAD based on noisy or filtered speech are indicative of an automatic speaker recognition system in which speech labels nor clean speech are provided. In this scenario, the babble noise results were only marginally improved through the application of Wiener filtering. It is expected that the effect of Wiener filtering is limited in this case due to the perceived noise spectrum being closely representative of the speech spectrum. In the case of car noise, however, Wiener filtering provided a significant improvement. In fact, using noisy speech for Wiener SAD provided comparable results to those obtained used Feature SAD based on clean speech. These results indicate that Wiener filtering improves the robustness of our energy-based SAD algorithm in the presence of car noise. However, Wiener filtering is not sufficient to reduce the effects of noise in the PLDA i-vector based system motivating further research into more advanced noise reduction techniques to address the detrimental effects of noise.

5. CONCLUSIONS

This paper evaluated the recent i-vector framework for speaker recognition based on PLDA in various noise conditions in comparison to other previous systems. Results indicate that the state-of-the-art i-vector framework is more noise robust than traditional GMM-UBM (i.e., dot scoring) methods. The i-vector framework was found to offer some robustness to added car noise, in which the EER doubled under very noisy 0 dBA SNR condition. Babble noise, however, posed a more significant problem. The application of Wiener filtering provided little benefit, thus motivating further research into noise-robust modelling techniques for speaker recognition.

6. REFERENCES

[1] 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.

[2] M.S. Deshpande and R.S. Holambe, "Am-fm based robust speaker identification in babble noise," *environments*, vol. 6, no. 10, pp. 19, 2011.

[3] S. Kim, M. Ji, and H. Kim, "Robust speaker recognition based on filtering in autocorrelation domain and sub-band feature recombination," *Pattern Recognition Letters*, vol. 31, no. 7, pp. 593–599, 2010.

[4] Y. Shao and D.L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 1589–1592.

[5] M.I. Mandasari, M. McLaren, and D. van Leeuwen, "Evaluation of i-vector speaker recognition systems for forensic application," in *Submitted to the 12th Annual Conference of the International Speech Communication Association*, 2011.

[6] National Institute of Standards and Technology, *NIST 2010 Speaker Recognition Evaluation Plan*, Available at http://www.itl.nist.gov/iad/mig/tests/sre/2010/.

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001, pp. 213–218.

[8] M. McLaren and D. van Leeuwen, "Source-normalised LDA for robust speaker recognition using i-vectors," *In print, IEEE Trans. Audio Speech and Language Processing*, 2011.

[9] M. McLaren and D.A. van Leeuwen, "A simple and effective speech activity detection algorithm for telephone and microphone speech," in *accepted into Proc. NIST SRE Workshop*, 2011.

[10] A. Strasheim and N. Brümmer, "Sunsdv system description: NIST SRE 2008," in *Proc. NIST SRE Workshop*, 2008.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *In print IEEE Trans. Audio, Speech and Language Processing*, 2010.

[12] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *Proc. Ninth Int. Conf. on Spoken Language Processing*, 2006, pp. 1471–1474.

[13] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2010.

[14] O. Glembek, L. Burget, N. Brummer, O. Plchot, and P. Matejka, "Discriminatively trained i-vector extractor for speaker verification," in *Proc. IEEE ICASSP*, 2011.

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

[16] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.