TOWARDS NOISE-ROBUST SPEAKER RECOGNITION USING PROBABILISTIC LINEAR DISCRIMINANT ANALYSIS

Yun Lei, Lukas Burget, Luciana Ferrer, Martin Graciarena, Nicolas Scheffer

SRI International

ABSTRACT

This work addresses the problem of speaker verification where additive noise is present in the enrollment and testing utterances. We show how the current state-of-the-art framework can be effectively used to mitigate this effect. We first look at the degradation a standard speaker verification system is subjected to when presented with noisy speech waveforms. We designed and generated a corpus with noisy conditions, based on the NIST SRE 2008 and 2010 data, built using open-source tools and freely available noise samples. We then show how adding noisy training data in the current i-vector-based approach followed by probabilistic linear discriminant analysis (PLDA) can bring significant gains in accuracy at various signal-to-noise ratio (SNR) levels. We demonstrate that this improvement is not feature-specific as we present positive results for three disparate sets of features: standard mel frequency cepstral coefficients, prosodic polynomial coefficients and maximum likelihood linear regression (MLLR) transforms.

Index Terms— Speaker Recognition, noise, robustness, i-vector, PLDA

1. INTRODUCTION

Recently, the speaker verification community has seen a significant increase in accuracy from the successful application of the factor analysis framework. In this framework, the i-vector extractor paradigm [1] along with a Bayesian back end such as probabilistic linear discriminant analysis, has become the state of the art in speaker verification systems. An i-vector extractor is generally defined as a transformation where one speech utterance with variable duration is projected into a single low-dimensional vector, typically of a few hundred components.

The low rank of the i-vector itself opened up new possibilities for the application of advanced machine learning paradigms that would have been otherwise too costly at the very high dimensionality most systems relied on earlier. Probabilistic linear discriminant analysis (PLDA) [2, 3] has been shown to be one of the most powerful techniques to produce a good verification score. In this model, each i-vector is separated into a speaker and a channel part, analogous to the formulation in the Joint Factor Analysis framework [4], PLDA is a probabilistic model modeling speaker and intersession variability in the space of i-vectors.

This work is focused on the robustness of speaker verification systems under noisy conditions, and how the proposed paradigm can help compensate for the observed degradation. Although the current state-of-the-art speaker recognition systems achieve very high performance on clean data, noisy conditions have been rarely experimented with. With the advance of technology, and the widespread use of mobile services, the need for noise-robust speaker recognition is on the rise. A lot of prior work on noise robustness focused on designing new robust acoustic features to mitigate noise degradation where the standard Mel Frequency Cepstrum Coefficients (MFCC) tend to fail [5, 6]. In the same vein, system combination using different acoustic features has been shown to improve accuracy under noisy conditions [7]. On the modeling side, degradation due to noisy data was studied on several state-of-the-art systems, including Gaussian mixture models (GMM) and MLLR [8].

2. NOISY SPEAKER RECOGNITION CORPUS

Data available from NIST SRE evaluations cannot be used to address noise robustness since it does not contain signals with low-enough SNR levels. Hence, in order to assess our systems' robustness to noise, we designed a corpus by adding real noise data to existing NIST data. (For a detailed description of the corpus, see [9].)

2.1. Original clean corpus

The noisy speech corpus is created by adding real noise (i.e., recorded noise samples) to data extracted from the SRE10 and SRE08 corpora. Only clean microphone data is selected from those corpora. Specifically, microphone 2 (lavalier microphones) waveforms are chosen from both interview and telephone conversations. Only SRE08 data is used for train-

This work was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), through the Army Research Laboratory (ARL). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI, or the U. S. Government.

ing, while SRE10 data and a small portion of SRE08 data is used to create trials (enrollment and testing). The clean trials are created as the cross-product of the sessions in the second set (except for same-session trials, which are discarded). That is, all possible target and impostor samples are created for the selected list of sessions

2.2. Additive noise

We selected 15 cocktail noise samples from the free sound repository Freesound.org [10]. These noise samples were collected in bars, cafeterias, offices, and airports. We inspected the samples to remove single-speaker foreground speech sounds and artifacts (like clicks, etc.). The noise samples vary in duration from 1 to 13 minutes.

The noise samples are labeled 1 to 15, and we added these 15 noise samples to the full waveforms from SRE08 and SRE10 at 20, 15, and 8 dB SNR ratios, using the publicly available tool called FaNT [11]. After noise addition we extracted the speech regions obtained from clean speech. This was done as a way to assess the performance of the speaker recognition systems independent of the quality of the voice activity detector (VAD) under noise. Hence, our results under noise can be interpreted as a best-case scenario that could be obtained in a real system if a very good VAD system trained for speech detection under noise were used.

2.3. Noisy corpus

In the noisy corpus, different noises are added to training, enrollment, and test samples. This avoids the highly optimistic matched case in which the same type of noise is observed when training the systems as in enrollment and/or test samples. Noises are separated into three disjoint sets: noise samples 1 through 4 used on enrollment signals, noise samples 5 through 8 used on test signals, noise samples 9 through 15 used on training signals. Within these sets, noise samples are added to the clean data randomly.

The noisy trials are created following the clean trial definition, where the clean enrollment sample has been degraded by one of the enrollment noises (at a certain SNR level) and the clean test sample has been degraded by one of the test noises (at a possibly different SNR level). Table 1 shows the number of target and impostor trials in all evaluation conditions of the noisy corpus.

3. BASELINE SYSTEM DESCRIPTION

For all the system in this work, fixed-length vectors are first extracted from feature sequences as a low-dimensional representations of speech segments. For each verification trial, the low-dimensional vectors are compared by means of PLDA [2] model to obtain verification scores. For two of our system, the low-dimensional speech representations are i-vector as proposed in [1], which are MAP point estimate of a latent vectors adapting GMM to an feature sequence of a given segment. More specifically, GMM mean supervector is constrained to live in a low-dimensional subspace and i-vectors are latent variables defining coordinates in this subspace. In the third system studied in the work, the i-vector like low-dimensional representation is derived from set of MLLR transformations adapting speech recognition system to speaker of a given speech segment [12]. In this work, all systems use linear discriminant analysis (LDA) and length normalization [13] before the PLDA back end. In PLDA, the speaker variability is described by a full rank matrix of eigenvoice bases. The training data for the LDA and PLDA models in all systems is the same and includes NIST SRE 04, 05, 06 and also data from Switchboard and Fisher, where multiple sessions are available per speaker. The i-vector extraction process with the PLDA back end for the different front ends is described below.

Table 1. *Number of target and impostor samples in each subset of the noisy corpus. The sets with matched SNR are 8 dB vs. 8 dB, 15 dB vs. 15 dB, 20 dB vs. 20 dB and clean vs. clean. The sets with mismatched SNR are created by matching data with SNR level of X for training and Y for testing and conversely. The all vs. all set is created by combining all of these sets.*

Eval. condition	# of Targets	# of Impostors
sets with matched SNR	2450	592,508
sets with mismatched SNR	4900	1,185,016
all vs. all	39,200	9,480,128

3.1. Cepstral i-vector system

In this front end, 19 cepstral coefficients and the energy with appended deltas and double deltas are used. A gender-dependent system with 2048-component diagonal covariance universal background models (UBM) trained on NIST SRE 04 and 05 telephone data, the i-vector extractor was trained on NIST SRE 04,05,06, Switchboard and Fisher, with i-vector dimensionality of 600. The i-vector dimension is further reduced to 150 by LDA.

3.2. Prosodic i-vector system

The prosodic features include 6th-order Legendre polynomial coefficients estimated from the energy and pitch tracks over regions 20 frames long with 5-frame shift. These 12 coefficients and the number of voiced frames in the region form the feature vector for each region. (For a detailed description of these features, see [14].) A gender-dependent 1024-component full covariance UBM and a 300-dimensional i-vector extractor are trained on the same data as the cepstral i-vector front end. The i-vector dimension is further reduced

to 200 by LDA instead of 150 as in the cepstral i-vector system.

3.3. MLLR system

For each speech segment, a total of 16 affine 39x40 MLLR transformation matrices are estimated to adapt the Gaussian means of HMM based LVCSR system to the speaker in the segment; eight transforms were estimated relative for each of the male and female recognition models, independent of the speaker's true gender. More details on the speech recognition system used can be found in [12]. The resulting vector is of dimension 24,960, and probabilistic principal component analysis (PPCA), trained on the same data as used in the i-vector extractor training of the cepstral i-vector system, is used to reduce the feature dimension to 800. The i-vector dimension is also reduced to 150 by LDA here.

4. NOISE-ROBUST SYSTEM

A straightforward approach to achieving noise robustness in our systems would be to add noisy data in all stages of the system: UBM, i-vector extractor, and LDA/PLDA. UBM and i-vector extractor training are computationally expensive stages taking lots of memory and CPU resources. On the other hand, LDA/PLDA training is very fast by comparison. Initial attempts at adding noisy data to UBM and i-vector training have shown very small gains. Hence, in this work, we only add noisy training data into the LDA/PLDA stage. Note that, as explained in the previous section, the noisy data added in training is not affected by the same noises as the test data. Furthermore, the data added to train LDA/PLDA includes all three SNR levels. We do not create SNR-specific models.

5. EXPERIMENTS

We evaluated systems trained on clean data under noisy conditions using the corpus described above. Furthermore, we show how the use of noisy data in training the speaker verification system can mitigate the effect of additive noise in the utterances. Performance is reported using equal error rate (EER) and decision cost function (DCF) as recently defined by NIST for the core condition of 2010 SRE [15].

5.1. Cepstral system

Figure 1 shows the performance of the cepstral system under different noise conditions, as well as the relative improvement obtained from the addition of noisy data in PLDA training.

As expected, the system accuracy degrades as the SNR decreases. Indeed, the EER degrades by about 13 times from the clean-clean condition to the 8 dB-8 dB condition. The improvement obtained from retraining PLDA with noisy data is significant and increases as the SNR decreases. The same

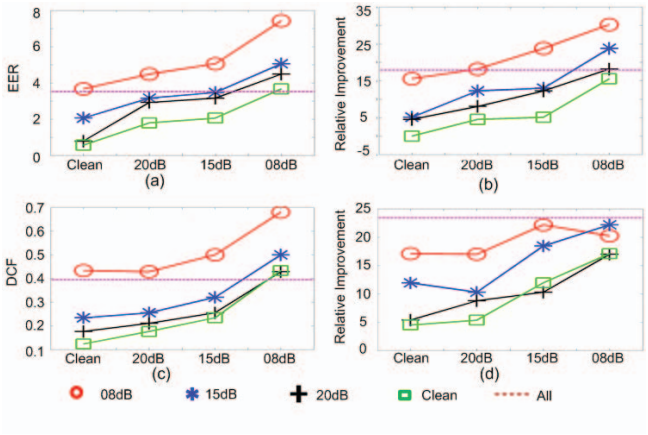

Fig. 1. A cepstral-based speaker verification system evaluated under different noise conditions. Figures (a) and (c) show the EER and DCF of the baseline system on different noise conditions. The x-axis corresponds to the SNR level in one of the sessions involved in the trials, while the color of the curve corresponds to the SNR level of the other session. Each marker then corresponds to the results on one of the sets in the noisy corpus. Figures (b) and (d) show the relative improvement when adding the noisy training data into the PLDA model relative to using only clean data (baseline system). The dashed line represents the performance in the *all vs. all* case

effect is observed on mismatched SNR sets. This is evidence that, in our systems, it is not essential to have a match between the SNR of the training data and that of enrollment and test data. This fact greatly simplifies the design of the system since a single LDA/PLDA model can be used for any SNR level in the data.

5.2. MLLR system

Figure 2 shows the performance of the MLLR system under different noise conditions and the relative improvement obtained from the addition of noisy data in PLDA training.

Similar conclusions as for the cepstral system described above can be drawn for the MLLR system: adding noisy data in PLDA improves the system's robustness under noise. However, there is a slight degradation of EER on the clean data set when using the noisy version of PLDA. Note also that, in this case, the degradation from the clean matched condition to the 8 dB matched condition is around 7 times, much less than that for the cepstral system.

5.3. Prosodic system

Figure 3 shows the performance of the prosodic system under different noise conditions and the relative improvement obtained from the addition of noisy data in PLDA training. The DCF metric is not shown here since it is always close to 1 for this system and no conclusions can be drawn from it.

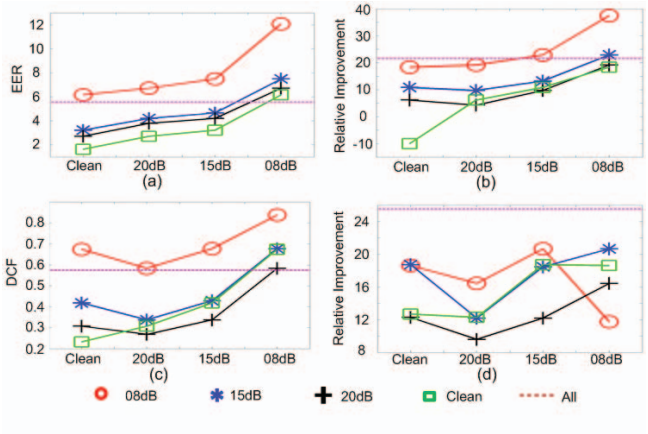

Fig. 2. An MLLR-based speaker verification system evaluated under different noise conditions. See caption of Figure 1 for an explanation of the format of the figures.

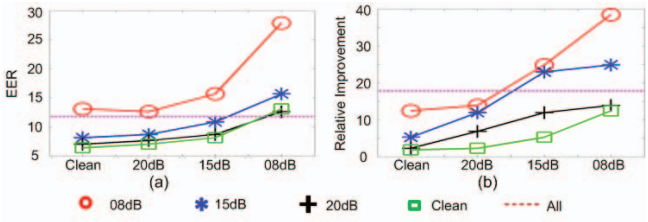

Fig. 3. A prosody-based speaker verification system evaluated under different noise conditions. See caption of Figure 1 for an explanation of the format of the figures.

In this case, the degradation from the clean matched condition to the 8 dB matched condition on the baseline approach is around four times, a much smaller degradation than for the other two systems. Similar conclusions can be drawn for the prosodic features as for the other feature sets described above with respect to the gains obtained from adding noisy data for PLDA training. However, in this case there is no performance degradation in the clean condition when the noisy training data is added into the PLDA training.

6. CONCLUSIONS

We show results on a newly designed noisy corpus for speaker recognition where real recordings of babble noise were added to original NIST SRE clean speech data. This corpus was designed so that the tools and data used for its creation are freely available, ensuring that the presented results can be replicated by other groups.

We show results for three different sets of features modeled using a state-of-the-art i-vector framework followed by PLDA modeling. The EER on the noisier sets of the corpus shows a range of degradations from 4 to 13 times with respect to the EER observed on the clean set. The level of degradation depends on the feature set, with higher-level features showing smaller degradations.

Finally, we show that adding nonmatched noisy data of several SNR levels to the PLDA training data gives improvements as large as 40% on noisy conditions with different SNR levels, with larger gains for noisier sets. These improvements are consistent across all feature sets tested. This observation supports the conclusion that current session variability compensation techniques can effectively deal with additive noise, and that a fairly noise-robust speaker verification system can be designed using state-of-the-art technologies as long as proper training data is available.

A natural next step for this work includes a combination of the presented systems. Initial results in this direction indicate that additional robustness to noise can be achieved by combination. These results will most likely be the topic of a follow-up paper. Another direction we plan to pursue is adding other kinds of real noises (apart from babble) to the noisy corpus, including, for example, traffic or office noises. Other robustness techniques are currently being explored, including class-dependent mean normalization of i-vectors.

7. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. ASLP*, vol. 19, May 2010.

[2] S.J.D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV-11th*. IEEE, 2007, pp. 1–8.

[3] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey 2010-The Speaker and Language Recognition Workshop*. IEEE, 2010.

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. ASLP*, vol. 16, July 2008.

[5] Y. Shao and D.L. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *ICASSP-2008*. IEEE, 2008.

[6] Y. Shao, S. Srinivasan, and D.L. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *ICASSP-2007*. IEEE, 2007, vol. IV.

[7] N. Thian and S. Bengio, "Noise-robust multi-stream fusion for text-independent speaker authentication," in *The Speaker and Recognition Workshop*, 2004.

[8] M. Graciarena, S. Kajarekar, A. Stolcke, and E. Shriberg, "Noise robust speaker identification for spontaneous Arabic speech," in *ICASSP-2007*. IEEE, 2007, vol. IV.

[9] L. Ferrer, H. Bratt, L. Burget, H. Cernocky, O. Glembek, M. Graciarena, A. Lawson, Y. Lei, P. Matejka, O. Plchot, and N. Scheffer, "Promoting robustness for speaker modeling in the community: the prism evaluation set," in *Proceedings of NIST 2011 Workshop*, 2011.

[10] "Freesound," http://www.freesound.org.

[11] "Fant," http://dnt.kr.hs-niederrhein.de/download.html.

[12] N. Scheffer and Y. Lei, "Factor analysis back ends for MLLR transforms in speaker recognition," in *Interspeech-2011*, 2011.

[13] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech-2011*, 2011.

[14] M. Kockmann, L. Ferrer, L. Burget, E. Shriberg, and J. Cernocky, "Recent progress in prosodic speaker verification," in *ICASSP-2011*. IEEE, 2011.

[15] "NIST SRE10 evaluation plan," http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf.