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

On the Influence of Text Content on Pass-Phrase Strength for Short-Duration Text-Dependent Automatic Speaker Authentication

Giacomo Valenti[1,2], *Adrien Daniel*[1] *and Nicholas Evans*[2]

[1]NXP Software, Sophia Antipolis, France
[2]EURECOM, Biot, France

giacomo.valenti@nxp.com, adrien.daniel@nxp.com, evans@eurecom.fr

Abstract

In the context of automatic speaker verification it is well known that different speech units offer different levels of speaker discrimination. For short-duration, text-dependent automatic speaker recognition, a user's pass-phrase bears influence on how reliably they can be recognized; just as is the case with text passwords, some spoken pass-phrases are more secure than others. This paper investigates the influence of text or phone content on recognition performance. This work is performed using the shortest duration subset of the standard RSR2015 database. With a thorough statistical analysis, the work shows how significant reductions in error rates can be achieved by preventing the use of weak passwords and that improvements in performance are consistent across disjoint speaker subsets. The ultimate goal of this work is to develop an automated means of enforcing the use of stronger or more discriminant spoken pass-phrases.

Index Terms: speaker recognition, text-dependent, short duration performance evaluation

1. Introduction

The performance of automatic speaker verification (ASV) technology is now sufficient to support mass-market, consumer applications [1]. Most of these, for instance smart phone, smart service applications and those within the sphere of the Internet of Things (IoT), call for short-duration enrolment and recognition, implying text-dependent recognition. While gaining momentum since the release of the RSR2015 [2] and RedDots [3] corpora, research in this area lags behind that in text-independent recognition.

The seminal work in [4] investigated differences in recognition performance at the speaker level, characterising four different speaker classes referred to as Doddington's menagerie. Later work in [5] investigated the influence on performance of specific training utterances. This work aimed to go beyond Doddington's menagerie and to investigate the role of phonetic content on ASV performance. With substantial variation in performance being observed, this raises the question of exactly what speech content is most relevant for speaker discrimination.

The work in [5] was extended in [6] which analysed the idiosyncratic information contained in French vowels. While perhaps offering greater insights relevant to the forensic branch of speaker recognition in terms of explaining results, the work points towards a mechanism for the selection or weighting of the most discriminant speech components for speaker modelling and recognition [7].

Most of the past work detailed above focuses on text-independent recognition where the tradition of speaker recognition evaluation (SRE) campaigns administered by the National Institute of Standards and Technology (NIST) generally dictates relatively long-duration training and testing. When speech data is plentiful, phonetic variation is naturally normalised to some extent. This is not the case for short-duration training and testing where speech data is sparse. In this case, phonetic variation can have a significant impact on recognition performance [8, 9]. Herein lies the contribution of our research.

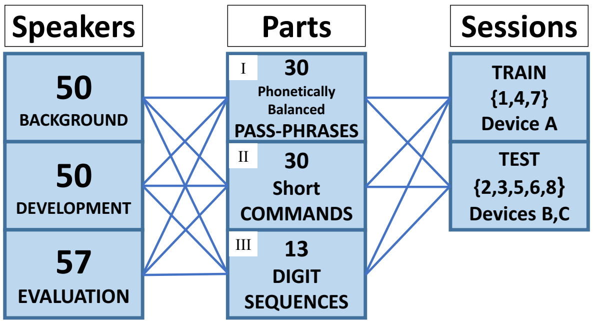

Figure 1: *RSR2015 Database partition for male speakers. The partition is identical for female speakers but with 43 speakers in the evaluation set instead of 57.*

This paper investigates the influence of text content on short-duration, text-dependent speaker recognition. The aim is to assess the variability in recognition performance and to determine the extent to which such variability is consistent across speakers. This work calls for a thorough statistical analysis which is reported here.

The remainder of this paper is organised as follows. Section 2 expands on the motivation for this work and identifies the database and protocols used for it. Section 3 describes the ASV system and results. The statistical analysis of command strength is described in Section 4.

2. Database and protocols

This section describes the database and text-dependent ASV system used for the work reported in this paper.

2.1. Database

The ultimate goal of this work is to develop a system to detect and prevent automatically the use of weak spoken passwords. Such a system would necessarily draw upon the use of speech data collected from other speakers; the only speaker-specific data available at enrolment would be one, or a small number of repetitions of the speaker's chosen password.

http://dx.doi.org/10.21437/Interspeech.2016-1115

Table 1: *The four possible kinds of trials for a text-dependent speaker verification system. They involve different combinations of matching speakers and text.*

Match	Speaker	Text
Target Correct (TC)	Yes	Yes
Target Wrong (TW)	Yes	No
Impostor Correct (IC)	No	Yes
Impostor Wrong (IW)	No	No

As a consequence, weak passwords are thus assumed to be universally weak, that is to say not specific to a given speaker. Required to support this work then, is a corpus collected from different speakers with multiple repetitions of the same set of sentences. The so-called sly impostor subset and associated protocol of the RSR2015 database [10] is ideally suited and is used for all work reported in this paper. The RSR2015 database partition is illustrated in Fig. 1. The sly impostor condition involves matched content impostor trials, sometimes referred to as the impostor-correct (IC) condition. This is one of four possible trials illustrated in Table 1.

The RSR2015 database contains phonetically-balanced sentences (part I), short commands (part II) and random digit trials (part III) (see Fig. 1). Since the target application of this work involves short spoken passwords, all experiments reported here are based upon the short commands condition (part II) where utterances contain in the order of 0.5 seconds of speech.

2.2. Protocols

As illustrated in Fig. 1, there are 50 male and female speakers in the background subset and 50 male and female speakers in the development subset. The evaluation subset is comprised of 57 male speaker and 43 female speakers. Each speaker provides recordings in 9 sessions. Data collected from 3 of the 9 sessions are set aside for training while the remaining 6 are used for testing. When experimenting on Part II, only Part I data is used for the learning of background information and there is no overlap between speakers or phrases between the data used for background modelling and that used for training and testing.

The sly impostor subset of Part II of the RSR2015 corpus contains 8990 TC (target) and 440510 IC (impostor) trials for the development set and 10250 TC and 574000 IC trials for the evaluation set. These numbers differ slightly from those reported in [11][1]. Since the literature focuses on results for the phonetically-balanced pass-phrases of Part I – this is the standard protocol distributed with the RSR2015 database – this paper also reports results for the same standard protocol. The Part I protocol dictates speaker-specific models which are trained with all 30 pass-phrases across the 3 training sessions, giving a total of 90 utterances. Speaker and pass-phrase models are trained with 3 utterances.

3. ASV system

Reported here is the ASV system architecture including details of the modelling and features together with results. While the contribution of this paper is not linked to advances in ASV technology, results are included as a means of illustrating performance relative to the state of the art.

[1] The authors became aware of the standard protocols for RSR 2015 Part II only after most of the work reported here was already completed.

Table 2: *Comparison of results for Part I of the RSR2015 database. Results shown for our implementation of the HiLam system with original results reported in [12]. Results are reported in terms of EER.*

Speaker set	Ours	Larcher et al. [12]
Part I Development	1.74%	1.43%
Part I Evaluation	1.93%	1.33%

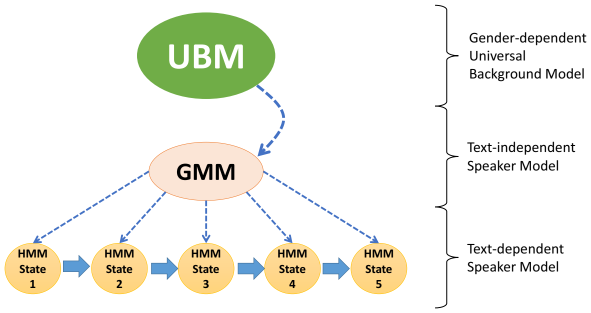
UBM
GMM
HMM State 1
HMM State 2
HMM State 3
HMM State 4
HMM State 5
Gender-dependent Universal Background Model
Text-independent Speaker Model
Text-dependent Speaker Model

Figure 2: *HiLam system architecture, reproduced from [10].*

3.1. Architecture

The baseline text-dependent ASV system used for all work reported in this paper is our own implementation of the so-called HiLam system originally reported in [11]. As illustrated in Fig. 2, the system is comprised of 3 layers: (i) a gender-dependent universal background model (UBM); (ii) speaker-specific Gaussian mixture models (GMMs) and (iii) speaker-and-text-specific hidden Markov models (HMMs).

The speaker-specific GMM model is obtained from the maximum a posteriori (MAP) adaptation of the UBM. The former is text-independent and does not model any time-sequence information; this is reflected only in the lower text-dependent level. Each HMM state is initialised with the same, second-level GMM model before Viterbi realignment and retraining. The full HiLam training and testing procedures are described in the original work [11].

In our implementation, GMM models have 64 components. MAP adaptation is applied with relevance factors of 19 and 3 for the second and third layers respectively. Scores are conventional log-likelihood ratios calculated between the claimed model and the UBM.

3.2. Feature extraction

The original RSR speech files are pre-processed with silence removal, by calculating the speech active level as recommended in ITU-T P.56 and by thresholding at 15.9 dB. This typically labels in the order of 64% of data for further processing; the remaining high-energy speech data is then frame blocked into 20ms frames with 10ms overlap. Standard MFCC features are then extracted in the usual way. They are comprised of 18 coefficients, without C0, which are appended with deltas and double deltas to produce features of 54 coefficients.

3.3. Performance

Table 2 shows a comparison of ASV results obtained with our implementation of the HiLam system with those reported in the original work [12] for Part I (phonetically-balanced pass-

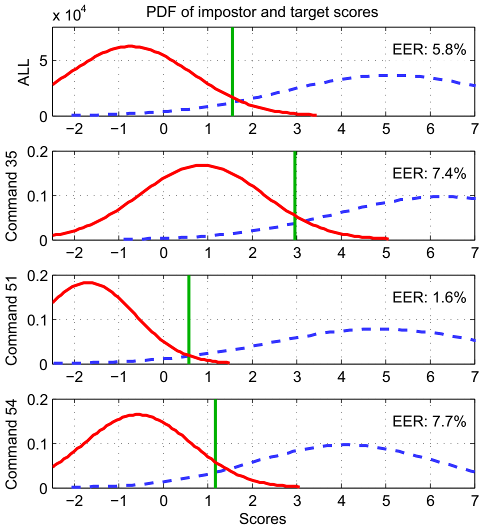

Figure 3: *Impostor (solid red) and target (blue dashed) score distributions and EER thresholds (green vertical lines). Plots illustrated separately for all commands trials (top) and for 3 command-specific trials.*

phrases). Results are reported in terms of EER. All results correspond to the IC condition and show a respectable level of performance; our results are only marginally worse than those reported in [12].

4. Statistical analysis of password strength

Both speaker characteristics and text content influence ASV score distributions. Example target and impostor distributions are illustrated in the top row of Fig. 3. Accept and reject decisions are made according to a *global threshold* illustrated by the vertical green line between the modes of each distribution. The amount of overlap between the two will then determine the *global EER*. The threshold is an inevitable compromise between the 'inner' target and impostor distributions related to an array of different factors, e.g. speaker-dependency, device-dependency and, in this case, text-dependency.

In the case of the IC condition, the influence of text is quantifiable from the target and impostor score distributions for subsets of same-text trials. These distributions are referred to as *text-dependent distributions* and the corresponding distribution overlap as the *text-dependent overlap*. As illustrated in Fig. 3 for commands 35, 51 and 54 of the RSR2015 database there is thus a *text-dependent EER* obtained with a *text-dependent threshold* for each command. The global EER is thus affected by both the text-dependent overlaps and the variation in the text-dependent thresholds. In contrast, text-dependent EERs are affected only by the text-dependent overlaps.

The following sections describe a statistical analysis that illustrates the potential to improve ASV performance through the selection of strong spoken sentences. It furthermore demonstrates that the notion of password strength is consistent across disjoint sets of speakers.

4.1. Variable strength command groups

The following describes a process to rank commands in terms of strength. This is needed in order to simulate a text-dependent ASV system that would eventually include password strength recommendation. On the assumption that a strong password is characterised by a relatively small text-dependent overlap, commands are first ranked by decreasing text-dependent EER. This process is performed separately for the development and evaluation sets thus yielding two rankings. From each of these rankings, groups of commands are formed by selecting 10 with the closest strength starting at every rank position, thereby producing 21 groups in total. The first group is comprised of the 10 weakest commands ranked #1 to #10, the second group is comprised of those ranked #2 to #11 and so on until the last group which contains the 10 strongest commands ranked #20 to #30. It is stressed that, while the groups obtained for the development and evaluation sets are similar, they are not identical.

4.2. Sampling distribution of the EER

ASV performance is assessed independently for each group in terms of the global EER (encompassing all commands in each group). The significance of the difference in recognition performance obtained for each group is measured with the following bootstrapping procedure [13].

For each group, a thousand populations of 30 commands are generated by picking at random from the 10 commands in the group. This procedure is known as resampling with replacement [13]. Each resampling of 30 commands out of 10 produces a population whose size is the same as that of the full dataset in terms of the number of trials. Each of these sampled populations yields an EER value which is computed from the target and impostor trials of the 30 commands of the population. These 1000 EERs form a sampling distribution of the global EER for each group.

The sampling distributions were visually inspected for normality, allowing for 95% confidence intervals of 1.96 times the standard deviation of the distribution, thereby removing 2.5% of the observations at each end of the distribution. This interval around the mean EER of the distribution has a high probability of encompassing the true value of the EER for each group. Differences in performance obtained for groups with non-overlapping confidence intervals can hence be considered as being statistically significant.

The bootstrapping procedure is applied using four combinations of different ranking and trial sets: (i) ranking and trials both for the development set, (ii) ranking and trials both for the evaluation set, (iii) ranking for the development set and trials for the evaluation set, and (iv) ranking for the evaluation set and trials for the development set. Combinations (iii) and (iv) are necessary in order to illustrate whether or not command strength is consistent across disjoint speaker sets. Statistics obtained for combinations (i) and (ii) are depicted in Fig. 4(a) and 4(b) by solid symbols in each plot. Statistics obtained for combinations (iii) and (iv) are depicted by unfilled symbols.

4.3. Isolating the influence of overlap

ASV performance estimated for each group is the consequence of the variation in text-dependent overlaps and text-dependent thresholds in each group. To illustrate the dependence on overlap in isolation from threshold effects, the experiments described above are repeated with all trial scores normalised according to the text-dependent threshold. The text-dependent EER for each command is then obtained with a score threshold of zero. Results for this experiment are reported in Fig. 4(c) and (d).

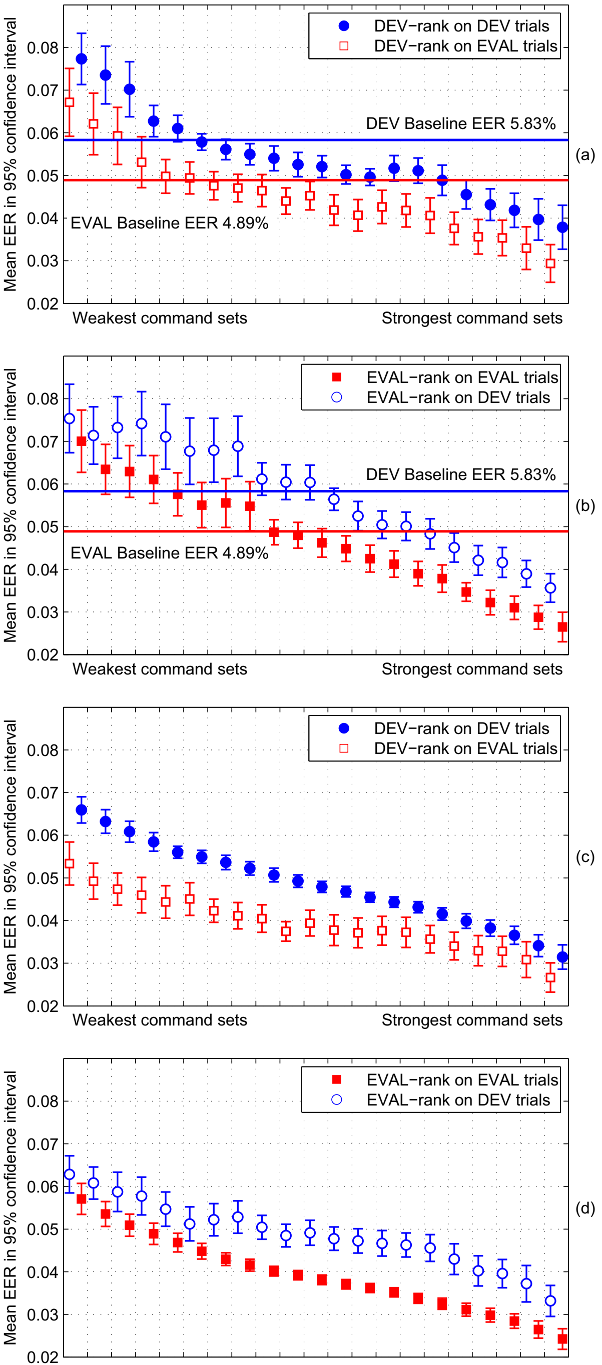

Figure 4: *ASV performance with (c,d) and without (a,b) text-dependent threshold adjustment. Each point represents the mean EER over 1000 resamplings of 30 commands chosen with replacement among the 10 commands of each sub group. The horizontal lines in (a,b) represent the baseline performance of the system for both sets with all 30 commands.*

4.4. Results interpretation

When using their own ranking, EER results for both the development and evaluation sets show significant decreases as the group contains increasingly stronger commands – solid-symbol plots in Fig. 4(a) and 4(b). When using threshold-adjusted scores (solid-symbol plots in Fig. 4(c) and 4(d), decreases are strictly monotonic. This observation confirms that the spread of text-dependent thresholds also affects performance.

Other observations concern results for cross-set rankings – unfilled-symbol plots in Fig. 4(c) and 4(d). Rankings made on the development set translate well to the evaluation set and vice-versa. For the evaluation set, results illustrated in Fig. 4(a) show that only 6 groups have an EER which is not significantly different to the overall EER (4.89%). For the development set, results illustrated in Fig. 4(b) show only 4 groups with a non-significantly different overall EER (5.83%). The significant global decrease in EER (albeit non-monotonic) shows that, with negligible differences in ranking, some commands are consistently ‘weak’ across different speakers. According to these results, a system including a password strength acceptance criterion could halve the error rate by choosing stronger sentences over weaker ones (from 5.34% to 2.67% on the development set, and from 6.28% to 3.32% on the evaluation set). Finally, we note that the visible offset of the evaluation set EERs is inherent to the RSR2015 database and consistent with results presented by others [11, 12].

The factors responsible for the ranking of command strength are not addressed in this paper, thus a solution to identify automatically weak short sentences is left for future work. Some intuitive, high-level observations are nonetheless offered. Consistent to both development and evaluation sets is the higher ranking of longer duration sentences. This is not surprising. Other observations are more intriguing. While commands such as ‘Turn on light’, ‘Watch Cartoon’ and ‘Volume Down’, all of similar duration, all perform well across both subsets, others of similar length such as ‘Door Open’, ‘Volume up’ and ‘Aircon off’ performed poorly across both subsets. Given the similar duration, it is assumed that the first three commands have more discriminative phonetic content. ‘Volume up’ and ‘Volume down’ vary only by the last two phonemes but are ranked among the weakest and strongest commands respectively. These observations are consistent with the discriminative power of nasal sounds studied in [7]. Clearly these factors warrant further attention in future work.

5. Conclusions and future work

This paper investigates short-duration, text-dependent automatic speaker authentication. The contribution relates to a thorough statistical analysis of the influence of text content on command strength. This not only influences the optimum system threshold, but also the degree of overlap between target and impostor score distributions. As a result, some spoken commands are stronger than others.

In order to examine the impact of text on the overlap between target and impostor score distributions and hence ASV performance, the influence of the threshold is compensated for *a posteriori*. Automatic means to compensate or normalise for this influence is an issue for future work. The ranking of commands according to their strength reveals considerable differences in their impact on system performance. The next stage of this work is to develop an automatic means of identifying weaker spoken short sentences. The intention is to develop such a system for a real-use case scenario in which the user of an ASV system may be encouraged to use a *strong* spoken sentence, namely one which offers a high level of discrimination among different speakers.

6. References

[1] K. A. Lee, B. Ma, and H. Li, “Speaker verification makes its debut in smartphone,” IEEE SLTC Newsletter, February 2013.

[2] A. Larcher, K.-A. Lee, B. Ma, and H. Li, “RSR2015: Database for text-dependent speaker verification using multiple pass-phrases.” in *INTERSPEECH*, 2012, pp. 1580–1583.

[3] K. A. Lee, A. Larcher, G. Wang, P. Kenny, N. Brummer, D. 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 *INTERSPEECH*, 2015, pp. 2996–3000.

[4] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, “Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation,” in *DTIC Document*, 1998.

[5] J. Kahn, S. Rossato, and J.-F. Bonastre, “Beyond doddington menagerie, a first step towards,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4534–4537.

[6] J. Kahn, N. Audibert, J.-F. Bonastre, and S. Rossato, “Inter and intraspeaker variability in french: an analysis of oral vowels and its implication for automatic speaker verification,” in *International Congress of Phonetic Sciences (ICPhS)*, 2011, pp. 1002–1005.

[7] K. Amino, T. Sugawara, and T. Arai, “Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties,” in *Acoustical science and technology*, vol. 27, no. 4, 2006, pp. 233–235.

[8] B. Fauve, N. Evans, and J.Mason, “Improving the performance of text-independent short duration SVM-and GMM-based speaker verification,” in *Odyssey*, 2008, p. 18.

[9] G. Soldi, S. Bozonnet, F. Alegre, C. Beaugeant, and N. Evans, “Short-duration speaker modelling with phone adaptive training,” in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[10] A. Larcher, K.-A. Lee, B. Ma, , and H. Li, “RSR2015: Database for text-dependent speaker verification using multiple pass-phrases,” in *Interspeech 2012*.

[11] A.Larcher, K. Lee, B. Ma, and H. Li, “Text-dependent speaker verification: Classifiers, databases and RSR2015,” *Speech Communication*, vol. 60, pp. 56–77, 2014.

[12] A. Larcher, K. A. Lee, P. L. S. Martnez, T. H. Nguyen, B. Ma, and H. Li, “Extended RSR2015 for text-dependent speaker verification over VHF channel,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[13] D. Howell, *Statistical methods for psychology, 7th edition*. Cengage Learning, 2010.