Content matching for short duration speaker recognition

Nicolas Scheffer, Yun Lei

Speech Technology and Research Laboratory, SRI International, California, USA

{nicolas.scheffer, yun.lei}@sri.com

Abstract

This work attempts to tackle the problem of content mismatch for short duration speaker verification. Experiments are run on both text-dependent and text-independent protocols, where a larger amount of enrollment data is available in the latter. We recently proposed a framework based on a deep neural network that explicitly utilizes phonetic information, and showed increased performance on long duration utterances. We show how this new framework can also yield significant improvements for short duration. We then propose an innovative approach to perform content matching, i.e. transforming a text-independent trial into a text-dependent one by mining content from a speaker's enrollment data to match the test utterance. We show how content matching can be effectively done at the statistics level to enable the use of standard verification backends. Experiments – run on the RSR2015 and NIST SRE 2010 data sets – show relative improvements of 50% for cases where the content has been said during enrollment. While no significant improvements were observed for the general text-independent case, we believe that this work might pave the way for new research for speaker verification with very short utterances.

1. Introduction

The speaker verification community has shown great interest in text-independent (TI) speaker verification (SV) for the past decade, and dramatic improvements have been obtained at various durations. Still, the ability to match the content between the enrollment and the test for a trial achieves the highest performance in short durations testing [1]. Unlike TI systems, text-dependent (TD) systems employ a different structure where the content and the speaker are modelled jointly [2], and thus cannot be used in a text-independent scenario or easily leverage the advances in TI research. TD verification has an additional major issue: speakers need to be cooperative and say a specific sentence or prompt.

This work aims at improving TI systems on short duration testing, where content plays an important role. While we are first interested in the TD protocol, where the content in enrollment and test is the same, our main focus is the scenario where models are estimated using all data available for a speaker regardless of what was said (a realistic assumption in the case of a deployed solution). Ideally, if the test prompt was pronounced during enrollment, a system should be able to maintain the performance achieved on a TD protocol. We however show that state-of-the-art TI systems fail to do so, as the additional data creates a large distortion not mitigated by the current modeling methods.

To tackle the challenges above, we propose to use approaches that explicitly take into account phonetic information for speaker verification. First, we look at the performance of a recently proposed paradigm [3] that involves a deep neural network (DNN) trained for automatic speech recognition (ASR). We argue that this framework is a better fit for a verification task where lexical variability is yielding the highest distortion. Second, based on this framework, we propose a method to perform content matching, i.e. a method that mines the lexical content from the enrollment data to match it to the one from the test utterance, somewhat like speech synthesis. We propose an efficient method to perform content matching at the statistics level before i-vector extraction by matching the zero-order statistics of the enrollment and the test. We show how the DNN system is an ideal tool as content is naturally factored out in the frame posteriors produced by the network. This approach can drastically improve performance if the content was already said by a speaker during enrollment without the need of explicit selection or labelling.

2. Baseline system and protocol

We describe below the experimental protocol, the baseline system as well as the performance issues that arise when speaker models are built with additional data n a text-independent fashion.

2.1. Experimental protocol

For reference, we first compare systems on the original RSR2015 protocol [4], restricted to the female speakers. The protocol defines 30 unique prompts, where speaker models are trained using 3 utterances from the same device (1, 4, 7). Test data is taken from sessions recorded on 2 other devices among the 6 available. Next, we define a new set of conditions to compare TD and TI verification performance with different definitions for the speaker models:

- **Match**: TD verification: a subset of the RSR2015 protocol restricted to 15 prompts only (1-15) and where trials that measure liveness detection are discarded (TARwrong). All impostor trials involving different prompts are also discarded (IMPwrong).
- **Seen**: TI verification where the test prompt pronounced in the enrollment data: A single, prompt-independent, model is built per speaker using all data available (15 prompts x 3 sessions = 45 utterances). The test data is

This material is based on work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of DARPA or its contracting agent, the U.S. Department of the Interior, National Business Center, Acquisition & Property Management Division, Southwest Branch. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. "A" (Approved for Public Release, Distribution Unlimited)

Table 1: *Baseline results on the RSR protocol, female speakers, with different PLDA scoring methods. EER (%) / mDCF x 100. Extracting a single i-vector for multiple sessions mimics a single enrollment session with different prompts and gives competitive results.*

Cond.	**Multi**	**Average**	**Single**
DEV.f	4.4 / 23.3	4.7 / 25.0	4.3 / 23.9
EVAL.f	3.6 / 18.2	3.9 / 20.4	3.4 / 16.6

the same as in **Match**.

- **Unseen**: TI verification where the test prompt does not appear in the enrollment data: Same enrollment as in **Seen** but the test data is drawn from the remaining prompts (16-30).

We use the same metric as for the RSR2015 protocol, i.e. the equal error rate (EER) and the minimum decision cost function (mDCF) with $P_{tar} = 0.01, C_{miss} = 10, C_{fa} = 1$.

We also show results on the NIST SRE 2010[1], 8-side 10-seconds (8s-10s) condition, where more than 15 minutes of enrollment data is available, as well as on a modified version where the test utterances were restricted to the first 3 seconds of speech (8c-3s).

2.2. Baseline speaker recognition system

Our baseline TI system is a UBM/GMM i-vector system [6, 7], similar in many aspects to the system presented in [5, 3]. 400 dimensional i-vectors are first extracted, followed by a reduction through linear discriminant analysis (LDA) to 200, and a probabilistic linear discriminant analysis (PLDA) backend. Features are standard mel frequency cepstrum coefficient (MFCC) with deltas of dimension 40. Training was based on the PRISM protocols [8].

We employ a scoring method that mimics a single long utterance for enrollment, by extracting a single i-vector for speaker models even when multiple sessions are available. This is because the problem of selecting data could be greatly simplified in the presence of separate sessions. Table 1 show the results with the different scoring techniques based on the PLDA model. Namely, extracting a single i-vector (*single*), averaging i-vectors across sessions (*average*), and multi-session scoring (*multi*). As extracting a single i-vector is giving the best results, this method will be used in the rest of the document.

Finally, we observe that our baseline system does not match the performance reported in [4]. This can be explained by the fact that none of the development data from the RSR data set was used for training the system. It has been shown that using short duration data for UBM or PLDA training can increase performance [9]. Note also that we didn't use the DEV results to improve results on EVAL. Our primary goal was not to design an optimal system for the RSR data set, but rather to use one that is employed for NIST SRE evaluations.

2.3. Problem highlight

Table 3 shows the performance of the baseline system on the conditions defined in 2.1. The TI condition **Unseen** is 5 times worse than the TD one (**Match**), highlighting the importance of the content mismatch between enrollment and test, even if more data is used to estimate the speaker models. Interestingly, and while systems are text-independent, testing against a prompt that was seen during enrollment – the **Seen** condition – yields large improvements ($\sim 40\%$) over **Unseen**. This condition is nonetheless 3 to 4 times worse than the **Match** one, and highlights the fact that current TI systems fail to detect that the added data is detrimental to the verification performance. By explicitly modeling content in the speaker recognition process, this work aims at solving that particular problem. As we strive for a solution, nothing should preclude our approach to be suited to the more general **Unseen** condition.

[1]National Institute of Standards and Technology, NIST speech group web, www.itl.nist.gov/iad/mig//tests/sre/2010/index.html

3. A DNN phonetic system for short duration verification

Our first attempt at tackling content mismatch is by using a recently proposed framework that leverage phonetic information for speaker verification.

3.1. The DNN / i-vector framework

The work in [3] shows how a deep neural network (DNN) trained for (tri)phone recognition can be used to produce the frame alignments for the speaker recognition system. Specifically, the posterior probabilities for each class are provided by the output of the DNN instead of the standard GMM model. Each output node in the DNN represent a tied state of context-dependent phones, defined by a decision tree and called *senones*. This system benefits from the recent successful research in deep learning for ASR [10, 11]. An added benefit is that DNN systems, by using a long window analysis, are able to normalize for speaker-specific pronunciations. By using speaker-independent frame alignments, two speakers saying the same content will more likely have their frames assigned to the same states regardless of their pronunciation.

The inputs of the net are formed from a 15-frame context window, from which the logarithm of the mel-filterbank output – of dimension 40 – of each frame is stacked to form a vector of dimension 600. Neurons used for the output layer are defined by the decision tree of the ASR system. Given an utterance, the sum of the posterior probabilities for each node over all frames is used as the zero-th order statistics for the i-vector model. The first order statistics can be computed with the same features as in 2.2, which are then whitened by means and covariances estimated on training data. The i-vector / PLDA framework can then be applied without any changes with the same configuration as the baseline system. The reader can refer to [3] for more details. In the experiments, the DNN is composed of 5 hidden layers each of dimension 1200. The DNN is trained using a cross-entropy criterion using alignments produced by a HMM-GMM ASR system, trained with maximum likelihood and 200K Gaussians. Both systems are trained on roughly 1300 hours of clean English telephone speech from Fisher, Callhome, and Switchboard data sets. We show results using two different DNN configurations, a large one with 3500 states and a smaller one with 400 states.

3.2. Experimental results

Table 2 shows the performance of the proposed systems on the RSR protocol. Compared to the GMM system, the DNN / i-vector system brings a relative 40% improvement for both metrics. Even with a low number of target trials, the improvements on the NIST SRE conditions are also significant, at 45% relative on the 8c-3s condition. This tends to show how the system fully leverage the large amount of enrollment data available for

Table 2: *Performance (EER (%) / mDCF x 100) of text-independent i-vector systems on RSR2015 (female speakers). The DNN-based system consistently improves over the baseline*

	GMM	DNN 400	DNN 3500
DEV.f	4.3 / 23.9	4.0 / 21.0	2.7 / 15.0
EVAL.f	3.4 / 16.6	2.8 / 13.8	2.0 / 10.4

a speaker. A similar conclusion as in [3] can be drawn for short duration testing where this additional phonetic information is helpful for speaker verification. However, the difference in performance between the **Seen** and **Match** conditions shows that the additional enrollment data is still detrimental to the performance.

4. Content matching

We now propose an approach for *content matching* by mining the enrollment data to match the lexical content of a given test utterance. In a way, we attempt to transform a text-independent trial into a text-dependent one, aiming at reaching the level of performance of TD verification using a TI protocol. This approach is, by nature, a trial-based solution and its success depends the ability to mine and find the relevant phonetic units in the data available for a speaker for a given prompt.

4.1. Frame alignments for content selection

The first solution that comes to mind for content matching would be mimic a speech synthesis approach where one would i) run speech recognition on the enrollment data and the test utterance and ii) select the appropriate frames in enrollment using the transcript to recreate the content of the test utterance. We propose an integrated solution that avoids explicit selection or 1-best hypothesis by using the frame alignments produced by the system to select data. For instance, the authors in [12] use the UBM posteriors to leverage conversational data in order to adapt their system to a specific TD task. For content matching however, there is a clear need to produce frame posteriors that are as speaker-independent and content-dependent as possible. Ideally, the same sequence of phonemes from two different speakers should produce the same alignment (modulo the duration effects). We argue that for content matching the UBM is, by nature, not well suited to factor out the content from the speaker. Since the UBM is purely acoustically driven, two speakers saying the same phoneme are likely to reside in different parts of the acoustic space if their pronunciations are far from each other. However, in the DNN / i-vector framework, the DNN is trained specifically for phone recognition, which, by nature, makes its output more speaker-independent.

Table 4 illustrates the ability of the frame alignments from the DNN and GMM systems to model the prompt and the speaker. We ran experiments by producing verification scores using the system's zero-order statistics, in the same vein as in [13]. We first run the systems on the **Match** condition, which is a speaker modeling task, and we added a new condition called **PromptID** in which target trials are designed to be from the same prompt regardless of the speaker identity. The results show that, compare to a GMM system, the DNN zero-order statistics are producing much worse speaker verification accuracy on the **Match** condition (more speaker-independent), while they can be used to effectively distinguish between different prompts (more content-dependent).

Table 4: *EER of the GMM and DNN systems based on a distance between the sum of frame posteriors (female speakers, DEV). Posteriors from the DNN system are more speaker-independent and can better factor out lexical information from the speaker.*

	Match (SpeakerID)	**PromptID**
GMM	5.9	24.7
DNN 3.5K	12.7	3.5

Table 5: *Interpreting content matching at the statistics level. N_e^s and N_t^s are the state s zero-order statistics for enrollment and test respectively.*

Condition	Content matching	Interpretation
$N_e > N_t$	Eq. 1, $\beta < 1$	Data selection
$N_e \leq N_t$	Eq. 1, $\beta \geq 1$	Reuse frames
$N_e > 0, N_t = 0$	$\beta = 0$	Discard unit
$N_e = 0, N_t > 0$	$\beta = 0$	Data synthesis (not addressed)

4.2. Content matching at the statistics level

As illustrated in Figure 1, we propose to define content matching as the process of matching the zero-order statistics of the enrollment to the test, and scaling the first order statistics appropriately. This results in an efficient process that does not need explicit selection and does not preclude this approach to be applied on more general TI protocols. While abstract for a GMM-based system, this process has a natural interpretation for the DNN system where each state has a phonetic meaning. Content matching is indeed equivalent to matching the count of senones in enrollment and test, before computing the i-vector.

Let us assume, for illustration purposes, that one of the DNN output node represents the senone $c[a]r_0$, and that this senone is present in the enrollment data but not in the test. To perform content matching, one could discard the occurrences of this senone in enrollment by leaving out the frames with highest probability for that senone before computing statistics. This can be however be done equivalently by putting 0 for this senone in the sufficient statistics. Unlike speech synthesis where a particular unit is selected, and because speaker verification does not take into account the frame ordering, the average statistics over the enrollment data is used. In practice, if N_t, F_t, N_e, and F_e are the zero- and first-order statistics of the test and enrollment respectively, then the new statistics $\hat{N}_e, \hat{F}_e$ are given by:

$$\begin{aligned} \hat{N}_e &= \beta * N_e \approx N_t \\ \hat{F}_e &= \beta * F_e \\ \text{with } \beta &= op\left(\frac{N_t}{N_e}\right) \end{aligned} \quad (1)$$

where *op* is an operator that numerically conditions the scaling factor between the statistics, as described in Table 5. Note that we do not address the case of senones that appear in the test utterance but not in enrollment (data synthesis). This is likely to be the focus of future research.

An added benefit of this approach is that the i-vector extraction can be done very efficiently as all models now share the same zero-order statistics and thus the same uncertainty for the i-vector point estimate, which is the most costly computation.

Table 3: *Performance (EER (%) / mDCF x 100) of text-independent GMM and DNN systems on different protocols and conditions (female speakers). Adding enrolment data to the model is always detrimental. The DNN system shows a consistent improvement over the GMM baseline. The proposed content matching (CM) is particularly effective for the Seen condition.*

	DEV.f			EVAL.f			NIST SRE 10 f	
	Match	**Seen**	**Unseen**	**Match**	**Seen**	**Unseen**	8c-10s	8c-3s
# tar / # imp	4223/194258	4223/388608	4225/388608	4403/211344	4403/422784	4405/422784	201/10313	201/10313
GMM	4.7 / 26.8	11.8 / 61.0	18.6 / 78.9	3.8 / 18.8	12.2 / 57.3	19.7 / 80.1	5.0 / 23.2	18.4 / 80.5
– CM	5.7 / 30.1	9.0 / 44.3	20.4 / 83.7	4.3 / 20.9	8.2 / 17.3	21.6 / 86.8	3.9 / 19.2	18.0 / 80.6
DNN 400	4.5 / 23.0	14.5 / 63.7	21.2 / 81.8	3.2 / 15.5	14.4 / 65.6	22.5 / 84.9	3.5 / 15.4	11.0 / 48.9
– CM	4.3 / 22.8	11.0 / 51.3	21.8 / 84.9	2.8 / 13.6	10.4 / 48.4	23.4 / 88.2	3.0 / 15.5	10.4 / 48.7
DNN 3.5K	3.3 / 17.2	10.2 / 52.2	17.9 / 80.0	2.9 / 11.1	10.4 / 50.1	20.1 / 81.6	3.0 / 14.2	12.4 / 47.1
– CM	3.2 / 16.2	4.6 / 25.4	20.7 / 88.4	1.9 / 9.7	4.2 / 21.6	23.1 / 91.4	2.9 / 13.3	10.4 / 44.8

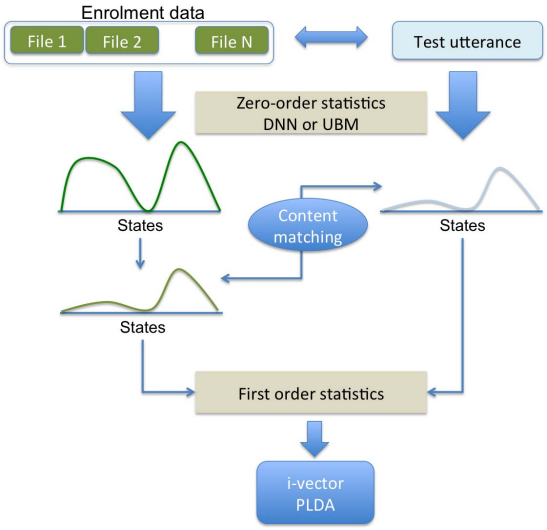

Figure 1: Content matching at the statistics level. The zero-order statistics of the enrollment data set are scaled to reflect the one from the test utterance. Speaker models are re-estimated for each test utterance.

4.3. Experimental Results

Table 3 includes the results of the proposed content matching (CM) approach on the **Seen** and **Unseen** conditions which maximizes the amount of enrollment data.

We observe that for the **Seen** condition, our approach yields significant improvements for all systems. While GMM posteriors are less content-dependent than the DNN ones, a 10% improvement is obtained for the baseline system. More interestingly, content matching using the large DNN setup show a 50% improvement with a EER at 1-2% absolute to the **Match** performance. While more modest, improvements with the DNN-400 are quite significant considering the few number of senones. This highlights the fact that the DNN system has inherently a better capability to model the content and factor out the speaker than the GMM system. However, we do not observe any improvement for the **Unseen** condition. This can be due to the lack of data present in the enrollment data, as only 15 different prompts per speaker are available as a whole. As a result, a large number of states in the test prompt do not appear in enrollment (*data synthesis*). Zeroing out this statistics in the current approach is clearly sub-optimal. However, the same trend is still observed on the NIST conditions where plenty of enrollment data is available. As we cannot produce text-dependent results on the NIST data, there is a possibility that the ASR-DNN system implicitly performs content matching.

There is a clear need of further research to explain why this method would not generalize to the real text-independent scenarios. Some research directions could consist in designing a more targeted decision tree, using a larger number of senones, or finding a more robust solution to the data synthesis scenario where the states in the test data have not been observed in enrollment.

5. Conclusions

This work aims at tackling the problem of content mismatch for short duration speaker verification using text-independent systems. We propose to use approaches that explicitly leverage phonetic information in order to increase accuracy for text-dependent protocols, but also for short duration testing where a large amount of text-independent enrollment data is available.

The current systems show degraded performance on short duration conditions if data from different content is added to the enrollment of a text-dependent model. We first show how the recently proposed DNN / i-vector system yields around 40% relative improvement on the RSR protocols but also on conditions where more enrollment data is available. This makes this system a suitable alternative to the current state-of-the-art.

Then, based on this framework, we propose a new approach to perform *content matching* where the enrollment data is transformed to be phonetically matched to a given test prompt. We show how the DNN system is a natural fit for this task compared to the UBM, and we propose an efficient approach by scaling the sufficient statistics such that no change is required in the current speaker recognition pipeline. Experiments show relative improvements of 50% for cases where the test prompt has been said by the speaker in the enrollment data. However, no significant improvement can be observed for the more general text-independent condition. While this approach might not solve the entire problem, we believe it can pave the way for future research in the community. To that end, future research directions are aimed at leveraging methods from the speech synthesis field which can help us refine the selection of data for every trial.

6. References

[1] Matthieu Hébert, "Text-dependent speaker recognition," *Springer Handbook of Speech Processing*, pp. 743–762, 2008.

[2] Anthony Larcher, Jean-François Bonastre, and John SD Mason, "Reinforced temporal structure information for embedded utterance-based speaker recognition.," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2008, pp. 371–374.

[3] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2014.

[4] Anthony Larcher, Kong-Aik Lee, Bin Ma, and Haizhou Li, "Rsr2015: Database for text-dependent speaker verification using multiple pass-phrases.," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2012.

[5] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013.

[6] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 788–798, 2011.

[7] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing.*, vol. 16, no. 5, pp. 980–988, 2008.

[8] Luciana Ferrer, Harry Bratt, Lukas Burget, Honza Cernocky, Ondrej Glembek, Martin Graciarena, Aaron Lawson, Yun Lei, Pavel Matejka, Olda Plchot, et al., "Promoting robustness for speaker modeling in the community: the prism evaluation set," in *Proceedings of NIST 2011 Workshop*, 2011.

[9] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7673–7677.

[10] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[11] George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[12] Hagai Aronowitz and Oren Barkan, "On leveraging conversational data for building a text dependent speaker verification system," *Annual Conference of the International Speech Communication Association (Interspeech)*, 2013.

[13] Nicolas Scheffer and J-F Bonastre, "Ubm-gmm driven discriminative approach for speaker verification," in *Odyssey, the Speaker and Language Recognition Workshop*. IEEE, 2006, pp. 1–7.