T-Norm for Text-Dependent Commercial Speaker Verification Applications: Effect of Lexical Mismatch

Matthieu Hébert and Daniel Boies

Nuance Communications
{hebert,boies}@nuance.com

Abstract

In this paper we describe a test-time score normalization technique (T-norm) for text-dependent speaker verification that is robust to lexical mismatch. The main challenge to the deployment of T-norm in a text-dependent task is the mismatch between the lexicon of the target speaker model in the application and that of the cohort speaker models. We show the negative effect of that mismatch in controlled experiments and propose a hybrid scoring scheme (T-Norm and background model) to remedy it. In a lexically mismatched scenario, which is inherent to the deployment of T-Norm in a text-dependent system, we show a 31% relative error rate reduction using the hybrid scoring over T-Norm alone. A 22% relative error rate reduction is measured over the baseline (no T-Norm) system.

1. INTRODUCTION

Test normalization, commonly known as T-Norm [8], was introduced as a natural extension to cohort normalization [9]. The technique proves very effective in normalizing verification scores. A variant called HT-Norm was designed to handle transmission channel effects. The bulk of the work on T-Norm and the related litterature is centered on its application to text-independent tasks. To our knowledge, a single study has applied T-Norm to a text-dependent task, even though it was not the main focus [12].

Although text-dependent and text-independent speaker recognition tasks are inherently different in nature, the last years have seen a significant convergence of the two fields. A trend in the current state-of-the-art text-independent systems now concentrates on frequently occurring words in conversational speech and perform speaker recognition on text-constrained, lexically restricted sets of words [10, 1]. This technique, in essence, puts a text-independent system in a text-dependent mode. The main effect is to reduce the lexical mismatch associated with a standard (non text-constrained) text-independent system. On the other hand, renewed interest in gaining flexibility in the possible prompted phrases [5] in text-dependent speaker recognition has culminated in a few studies on the effect of lexical mismatch in a text-dependent task [4, 2]. The main challenge, as recently pointed out in an invited lecture at the Odyssey workshop [3], is to improve the text-dependent speaker recognition performance when lexical mismatch is present. The text-dependent tasks are very well suited for controlled, in-depth analysis of lexical mismatch, and we suggest without proof that the conclusions can largely be applied to text-independent speaker recognition tasks especially in the text-constrained mode.

The deployment of a text-dependent system that uses T-Norm as a score normalization scheme represents technical challenges. From the point of view of memory usage and I/O to the speaker model database, it is currently illusory to have a specific cohort for each and every speaker. Furthermore, for verification dialogs that allow a password phrase unique to each user, the system would require a set of cohort speaker models that would have been trained using the target user's password phrase to achieve a lexically competitive cohort. A way to circumvent this limitation is to use the same password phrase for all users of the application. Examples of common password phrases are : *"My Voice is My Password"* and *"One Two Three Four Five Six Seven Eight Nine"*. The cohort is then easy to construct, but this type of dialog is restrictive and may not be suited for most applications since the claim of identity needs to be done in a separate step in the dialog and a common phrase represents a loss in overall security of the application.

This paper presents the results of a series of experiments on T-Norm in a text-dependent speaker recognition system. The emphasis is on the effect of lexical mismatch between the target password's lexicon and the cohort speaker model's lexicon. We will show that the aforementioned lexical mismatch has a negative effect on T-Norm. We will introduce a back-off to the standard background model score normalization to address the lexical mismatch and improve significantly the efficiency and robustness of T-Norm.

2. ALGORITHM DESCRIPTION

The baseline system used in this study has been described in [6] and [7]. In essence, it uses the recognizer's alignments to perform verification on a per-phoneme basis using Gaussian mixture models (GMM) as the underlying modeling unit. Channel-dependent background modeling and speaker model synthesis are used for robustness to cross-channel effects [11]. The scoring of cohort speaker models during T-Norm related computations involves the same processing.

Let $j(t)$ be the frame-level phonetic alignment given by the recognizer; it states that at time t' the frame is aligned to phonetic class j. Also, let us define $\lambda^{t,c}$ and $\bar{\lambda}^c$ as the target speaker and background models for channel c. Then the likelihood ratio involved in the standard decision scheme takes the form

$$\begin{aligned} L(\mathbf{X}|\lambda^{t,c}) &= \log p(\mathbf{X}|\lambda^{t,c}) - \log p(\mathbf{X}|\bar{\lambda}^c) \\ &= \frac{1}{T}\sum_{t'}\left[\log p(\mathbf{x}_t|\lambda^{t,c}_{j(t')}) - \log p(\mathbf{x}_t|\bar{\lambda}^c_{j(t')})\right] \end{aligned} \quad (1)$$

where $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ is the set of feature vectors extracted from the utterance and $\lambda^{t,c}_j$, $\bar{\lambda}^c_j$ are respectively the GMMs representing the user t (speaker model) and background model for channel c and class j. The channel is identified at test time as the one corresponding to the $\bar{\lambda}^c$ with the highest likelihood.

The baseline T-Norm implementation used in this paper is

defined by

$$L_C(\mathbf{X}|\lambda^{t,c}) = \frac{\log p(\mathbf{X}|\lambda^{t,c}) - \mu(\mathbf{X}, C)}{\sigma(\mathbf{X}, C)} \quad (2)$$

where the first moment of the cohort C score distribution is defined by

$$\mu(\mathbf{X}, C) = \frac{1}{N_C} \sum_{s \in C} \log p(\mathbf{X}|\lambda^{s,c}) \quad (3)$$

and the standard deviation $\sigma(\mathbf{X}, C)$ is defined in a consistent fashion. The number of cohort speaker models is N_C. Again, the channel c is selected in the same manner as described above. In this case, the cohort C is selected based on the detected channel at testing time and we will name it C^t in the rest.

A variant of the T-Norm algorithm uses a cohort that is defined by the channel as detected during enrollment of user t (C^e). We can re-formulate Eq. 2 for this case as

$$L_{C^e}(\mathbf{X}|\lambda^{t,c}) = \frac{\log p(\mathbf{X}|\lambda^{t,c}) - \mu(\mathbf{X}, C^e)}{\sigma(\mathbf{X}, C^e)}. \quad (4)$$

In this case, the testing time channel detection affects only the likelihood computations and not the composition of the cohort. Although this T-Norm scheme is expected to perform poorly in cross-gender attempts (where the $\mu(\mathbf{X}, C^e)$ will become large and negative), it will serve as an illustrative example of our proposed back-off to standard background model mechanism.

In this paper, we used gender-dependent cohorts only. As such, there are only two possible C^t and C^e, namely C^{male} and C^{female}. The gender is determined offline, prior to the experiments using the aforementioned channel detection algorithm. We also have the real gender for the speaker models present in the cohort: the real and detected gender are consistent.

Without justifying the necessity of a back-off mechanism for now, let us introduce the formalism here. This back-off to the standard background model mechanism is implemented as a smoothing between the standard verification score (1) and T-Norm score (2 or 3 depending on the particular experiment). Define α as

$$\alpha(\mathbf{X}, C) = \frac{1}{1 + \exp\left[\left(\mu(\mathbf{X}, C) - \log p(\mathbf{X}|\bar{\lambda}^c) - \theta\right)/\beta\right]} \quad (5)$$

where C can be C^e or C^t depending on the experiment. Parameter θ and β are free parameters that we can optimize. The smoothed T-Norm score is defined as

$$L'_C(\mathbf{X}|\lambda^{t,c}) = \frac{\log p(\mathbf{X}|\lambda^{t,c}) - \mu'(\mathbf{X}, C)}{\sigma'(\mathbf{X}, C)} \quad (6)$$

where

$$\mu'(\mathbf{X}, C) = \alpha(\mathbf{X}, C) \log p(\mathbf{X}|\bar{\lambda}^c) + (1 - \alpha(\mathbf{X}, C))\mu(\mathbf{X}, C) \quad (7)$$

$$\sigma'(\mathbf{X}, C) = \alpha(\mathbf{X}, C) + (1 - \alpha(\mathbf{X}, C))\sigma(\mathbf{X}, C) \quad (8)$$

As shown in (7) and (8), the proposed smoothing scheme is an interpolation of the normalizing statistics μ and σ between standard T-norm ($\alpha = 0$) and background model normalization ($\alpha = 1, \sigma = 1$). The proposed adaptive normalization reverts to background model normalization when the T-norm μ is much lower than the standard background model score. This situation occurs when, for a given test attempt, the impostor models in the cohort are not competitive enough.

3. DATA SET DESCRIPTION

To benchmark T-Norm on a text-dependent task, we have used one of our internal databases. The data was collected in September of 2003 from 142 unique speakers (70 males and 72 females). Speakers were requested to read a sheet with 3 repetitions of the phrase "1 2 3 4 5 6 7 8 9" which we'll call the E utterances and 3 repetitions of a random 9-digit string which we'll call the S utterances. There were only 8 unique S digit strings in the database in order to use round-robin impostor attempts, and a given speaker was assigned only 1 S string. The speakers were requested to complete several calls on a variety of channels in realistic noise conditions. The fact that for every call, we have 3 E utterances and 3 S utterances is very convenient: we can substitute E $\Leftrightarrow$ S and perform very controlled experiments because all utterances are from the same call.

We have randomly set aside 25 male and 25 female speakers to compose our cohort. From the remaining 92 speakers, we can construct a testset with 13172 true speaker attempts and 8657 impostor attempts. Note that the impostors in the case of the S utterance were picked ONLY from the pool of speakers that were assigned the same S utterance out of the possible 8 to get lexically challenging impostors. The enrollment dialog is composed of the 3 repetitions of either E or S in a call and the verification is composed of 2 repetitions of E or S in a different call. The testset was designed to have no cross-gender attempts based on the true gender of the participants in the database.

We have targeted 100 speaker models for each gender in our cohort. This means that 4 different calls for each speaker were used to construct 4 different speaker models for each speaker in the cohort. Due to the fact that not all speakers had completed the target 4 calls, we ended up with 93 speaker models in our male cohort and 99 in our female cohort. The channels are evenly represented in the cohort.

Let us define the notation eXXX_vYY_cZZZ to describe an experiment. It defines the enrollment as 3 repetitions of X, the verificaton attempts as 2 repetitions of Y and the cohort speaker models as enrolled with 3 repetitions of Z. If the _cZZZ part is omitted, then the experiment doesn't use T-Norm (verification score is computed using Eq. 1).

4. RESULTS AND ANALYSIS

4.1. Justification for back-off mechanism

An interesting way to show the necessity of a back-off mechanism is by looking at the main argument of Eq. 5 namely $\mu(\mathbf{X}, C) - \log p(\mathbf{X}|\bar{\lambda}^c)$ which we'll call "mu-bm" for short. Figure 1 presents two special cases. On the left pane, the histogram of mu-bm shows a very long tail for the case where C^e is used: this is due to tagged cross-gender attempts. Even if the testset was not designed to have any, the gender tagger makes a few mistakes which have a very negative effect on the overall performance, as we will see later. We can circumvent this shortcoming by using the appropriate T-Norm formulation Eq. 2 with C^t as the cohort. On the right pane of this Figure, we can see another, potentially more dangerous situation. The histograms presented show the effect of lexical mismatch of the cohort speaker model's enrollment lexicon and test-time lexicon. In the case of a lexically mismatched cohort (eSSS_vSSS_cEEE), we can see that the mu-bm distribution is shifted towards large negative values. We'll see later that this has a significant negative effect on the T-Norm performance.

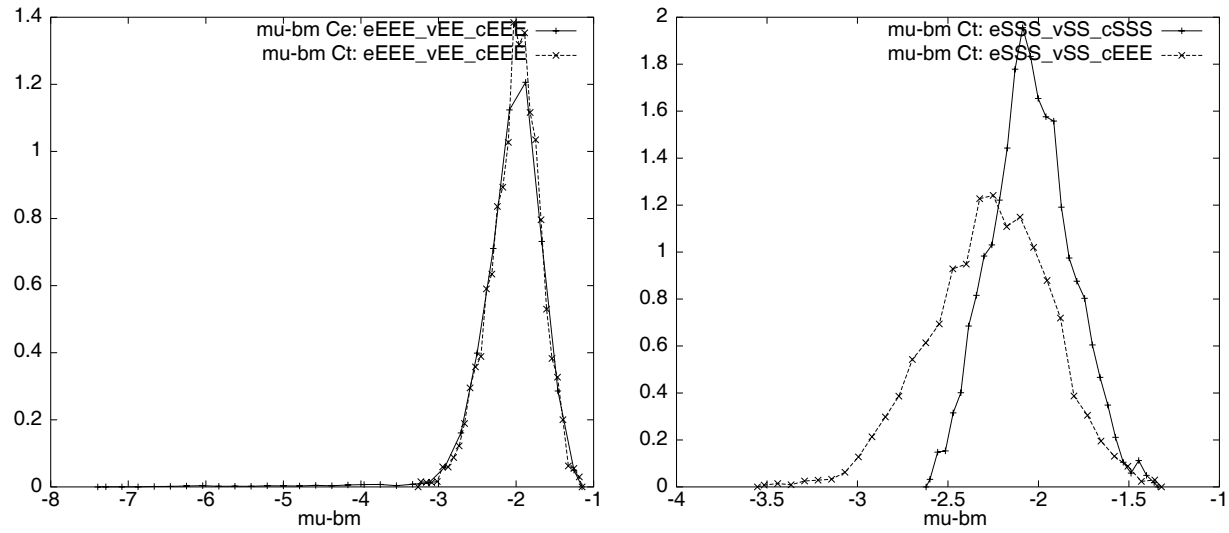

Figure 1: Histograms mu-bm for 2 different illustrations of the need for a back-off to standard verification scores via smoothing by Eq. 6. In the legends, Ce stands for a cohort that is selected by the gender of the target speaker model (C^e in the text) and Ct is for a cohort selected using the detected gender of the utterances of the verification trials (C^t in the text).

4.2. Effect of T-Norm on performance

It is well known that T-Norm induces a counter-clockwise tilt in DET curves [8]. Therefore, we will not use the Equal Error Rate (EER) to compare system performance but rather the False Rejection rate (FR rate) at fixed False Acceptance rate (FA rate). The target FA rate is set to 1%. As can be seen in Table

Exp. set-up	Baseline (no T-Norm)	T-Norm C^t cEEE	T-Norm C^t cSSS
eEEE_vEE	17.10%	14.96%	14.74%
eSSS_vSS	14.44%	16.39%	10.42%

Table 1: Table showing the FR rates at FA = 1% for various configurations. Based on the lower number of trials (impostor in our case), the 90% confidence interval on the measures is 0.6%.

1, T-Norm in the context of a text-dependent task clearly improves the performance (reducing the FR rate by 20% relative) especially when the cohort is chosen to match the target user's password phrase. In the case of a lexically poor cohort (_cEEE) and lexical mismatch with the test lexicon, we can see a significant degradation over the baseline (13% relative degradation). This is an indication that the standard T-Norm algorithm is sensitive to lexical mismatch between the target user's password and the cohort speaker model's enrollment lexicon. Due to this fact, a text-dependent system using T-Norm would be very hard to deploy, as mentioned in the Introduction, unless a restrictive dialog is used (namely the one that uses the E utterances for everybody). For a text-independent system, a similar effect is certainly present especially in the text-constrained paradigm. The remaining sections of this paper will focus on the eSSS_vSS scenario which is the most appealing from the deployment security and flexibility point of view.

4.3. Effect of Smoothed T-Norm on performance

Let us first benchmark the effect of smoothing the T-Norm normalization and the background model normalization in the context of lexically matched experiments (see Table 2). Note here that optimization of β and θ was performed independently in all experiments with smoothed T-Norm. The optimization was done on the experiment's results, but values of β and θ are found to be very stable across experimental conditions. As can be seen from the Table, in the case of the cohort selection done at testing time (C^t), smoothing seems to still improve the performance in the all-E experiment, whereas it has little effect on the all-S experiment. For the case of the cohort selection at enrollment time (C^e), the smoothing seems to help systematically. It essentially addresses the long tail seen in Figure 1. All of these results point in a single direction: smoothed scores can be seen as a safeguard against $\mu(\mathbf{X}, C)$ (Eq. 3) becoming too large and negative. In such case, we rely more and more on the background model score, as is the case for lexically mismatched experiments presented here. We think that the proposed smoothing can also have a beneficial effect when faced with other types of mismatch (channel, environmental noise, etc).

Figure 2 shows a series of DET curves that illustrate the benefit of the smoothed T-Norm scores. In the case of a lexically mismatched cohort we can see that performance of T-Norm degrades; when smoothing is applied, we observe a substantial gain in performance and almost recover the lexically matched cohort. The smoothing mechanism can thus be seen as increasing the robustness of T-Norm.

5. CONCLUSION

In this paper, we have studied the use of T-Norm in a text-dependent system. We have shown its efficiency in improving the performance, but also its fragility to lexical mismatch. A smoothing between the T-Norm normalization and the background model normalization was introduced as a means to alleviate the lack of robustness of T-norm to lexical mismatch. We also suggest that the smoothing can also be applied to make T-Norm more robust to other sources of mismatch and that the conclusions therein may be applicable to text-independent speaker verification especially in the text-constrained mode. In a

Exp. set-up	Baseline (no T-Norm)	T-Norm C^e	T-Norm C^e smooth	T-Norm C^t	T-Norm C^t smooth
eEEE_vEE_cEEE	17.10%	16.95%	13.74%	14.96%	13.46%
eSSS_vSS_cSSS	14.44%	11.35%	10.53%	10.42%	10.45%

Table 2: Table showing the FR rates at FA = 1% for various configurations.

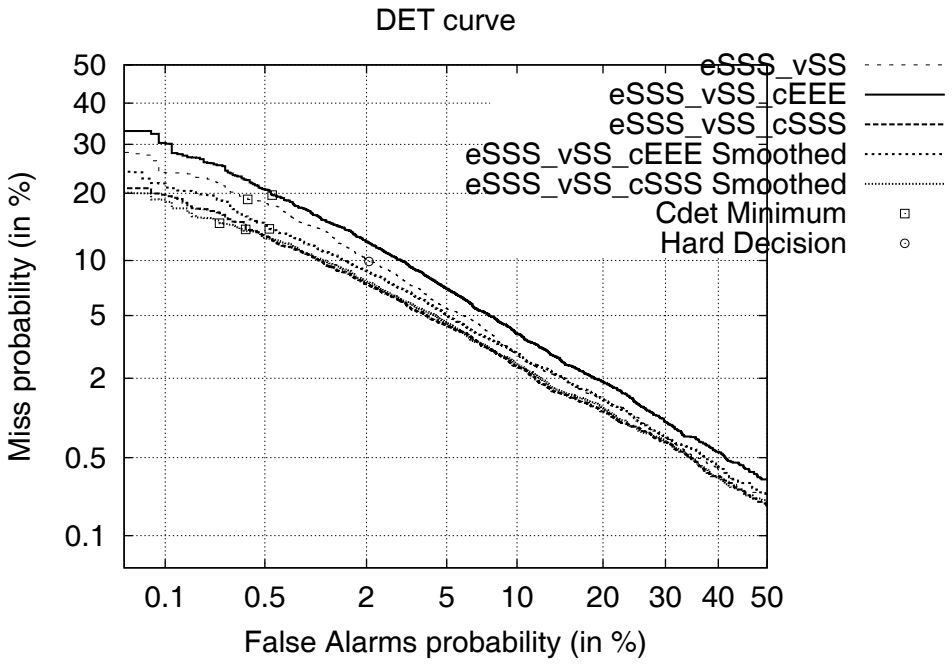

Figure 2: DET curve showing the T-Norm (Eq. 2) and its Smoothed variant (Eq. 6) in the case of eSSS_vSS with the cohort selected at testing time (C^t). In this Figure, $\theta = -3.4625$ and $\beta = 0.6975$ are common to both smoothed curves.

lexically mismatched scenario which is inherent to the deployment of T-Norm in a text-dependent system, we show a 31% relative error rate reduction (FR @ FA=1%) using the smoothed T-Norm over standard T-Norm. A 22% relative error rate reduction is measured over the baseline (no T-Norm) system.

6. References

[1] K. Boakye and B. Peskin. Text-constrained speaker recognition on a text-independent task. *Proc. Odyssey Speaker Recognition Workshop*, 2004.

[2] D. Boies, M. Hébert, and L.P. Heck. Study on the effect of lexical mismatch in text-dependent speaker verification. *Proc. Odyssey Speaker Recognition Workshop*, 2004.

[3] L.P. Heck. On the deployment of speaker recognition for commercial applications: Issues and best practices. *Proc. Odyssey Speaker Recognition Workshop*, 2004.

[4] T. Kato and T. Shimizu. Improved speaker verification over the cellular phone network using phoneme-balanced and digit-sequence-preserving connected digit patterns. *ICASSP*, II:57–60, 2003.

[5] T. Matsui and S. Furui. Concatenated phoneme models for text-variable speaker recognition. *ICASSP*, II:391–394, 1993.

[6] M.Hébert and L.P. Heck. Phonetic class-based speaker verification. *EUROSPEECH*, pages 1665–1668, 2003.

[7] N. Mirghafori and M. Hébert. Parametrization of the score threshold for a text-dependent adaptive speaker verification system. *ICASSP*, I:361–364, 2004.

[8] M. Carey R. Auckenthaler and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. In *Digital Signal Processing*, volume 10, pages 42–54, 2000.

[9] A.E. Rosenberg, J. Delong, C.-H. Lee, B. Juang, and F.K. Soong. The use of cohort normalized scores for speaker recognition. In *ICSLP*, pages 599–602, Banff, Canada, 1992.

[10] D.E. Sturim, D.A. Reynolds, R.B. Dunn, and T.F. Quatieri. Speaker verificaiton using text-constrained gaussian mixture models. *ICASSP*, I:677–680, 2002.

[11] R. Teunen, B. Shahshahani, and L.P. Heck. A model-based transformational approach to robust speaker recognition. In *ICSLP*, Beijing, China, 2000.

[12] R.D. Zilca, J.W. Pelecanos, U.V. Chaudhari, and G.N. Ramaswamy. Real time robust speech detection for text-independent speaker recognition. *Proc. Odyssey Speaker Recognition Workshop*, 2004.