interspeech 2008 incorporating SST 08

MAP and Sub-Word Level T-Norm for Text-Dependent Speaker Recognition

Doroteo T. Toledano[1], *Daniel Hernandez-Lopez*[1], *Cristina Esteve-Elizalde*[1], *Joaquin Gonzalez-Rodriguez*[1], *Ruben Fernandez Pozo*[2] *and Luis Hernandez Gomez*[2]

[1] ATVS Biometric Recognition Group, Universidad Autonoma de Madrid, Spain
[2] GAPS, SSR, Universidad Politecnica de Madrid, Spain

doroteo.torre@uam.es

Abstract

This paper presents improvements in text-dependent speaker recognition based on the use of Maximum A Posteriori (MAP) adaptation of Hidden Markov Models and the use of new sub-word level T-Normalization procedures. Results on the YOHO corpus show that the use of MAP adaptation provides a relative improvement of 22.6% in Equal Error Rate (EER) in comparison with Baum-Welch retraining and Maximum Likelihood Linear Regression (MLLR) adaptation. The newly proposed sub-word level T-Normalization procedures provide additional relative improvements, particularly for small cohorts, of up to 20% in EER in comparison with the normal utterance-level T-Normalization.
Index Terms: speaker recognition, text-dependent.

1. Introduction

Automatic Speaker Recognition (SR) aims to recognize the speaker that produces a particular speech utterance. It can be either text-independent or text-dependent depending on whether the linguistic content of the test speech utterance is unknown or known by the system. In the latter case the text can be a password set by the user or a random text prompted to the user (text-prompted). Despite its potential applications in interactive voice response systems, text-dependent SR has developed at a slower pace than text-independent SR, probably due to the lack of competitive evaluation campaigns such as NIST text-independent SR evaluations [1].

The most widely used modeling technique in text-dependent SR is Hidden Markov Models (HMMs) [2, 3, 4]. This paper also focuses on text-dependent SR using HMMs. Our previous work [5] compared Baum-Welch retraining versus Maximum Likelihood Linear Regression (MLLR) adaptation [6] for training the speaker models. In this paper we extend this comparison to the use of Maximum A Posteriori (MAP) [7] adaptation of the HMMs as a better way for obtaining the speaker models.

Besides this comparison, the other novelties in this paper are two new T-Norm procedures particularly designed for its use in text-dependent SR and an extensive experimentation with them. The main idea behind these new T-Norm procedures is to perform T-Norm on scores computed on smaller segments of speech (such as phonemes or HMM states) so that the averaging of the scores over the full utterance is performed on already normalized scores. This idea contrasts with the normal way of applying T-Norm in which first scores are averaged over the whole utterance and T-Norm is applied afterwards to these utterance-level scores. We call this normal way of T-Norm *Utterance-Level T-Norm* to distinguish it from the newly proposed schemes operating at the sub-word level, which we call *Phoneme-Level T-Norm* and *State-Level T-Norm*. We introduced these T-Norm schemes in [5], where we showed that using a single cohort composed of 10 male and 10 female speakers *Utterance-Level T-Norm* actually decreased performance, while *Phoneme-Level* and *State-Level T-Norm* yielded important improvements. Although the results were quite clear, some concerns could be raised about the generality of the conclusions given that the cohort included both genders (and therefore included gender-related variance), was small, and results included same-gender and cross-gender tests. This paper tries to give answer to these concerns by extending the experimentation to the cases of using two gender-dependent cohorts of 10 and 30 speakers and a male only test using a cohort of over 100 speakers. We also try to analyze the data in more detail to get insights into the reasons for the behavior observed. For the moment, all the experiments with T-Norm shown in this paper are performed on the well-known YOHO database [8]. We are currently working on extending these experiments to other databases [9].

The use of T-Norm for text-dependent SR has received little attention until very recently [4, 5, 10]. Of particular interest for this paper is the work in [10], where the authors propose the effect of the lexical mismatch as one of the reasons for the modest performance of T-Norm in text-dependent SR. In [10] the authors propose a technique for smoothing the normalization that yields improvements. Here we present an alternative way of improving the performance of normalization, by performing T-Norm at the phoneme or sub-phoneme levels instead of at the utterance level. This method, does not solve the problem of the lexical mismatch in the speech used in the enrollment of the models and in the utterance to verify, but we consider that by reducing the amount of the lexical content of the test segment used to compute the score before applying T-Norm to one phoneme or sub-phoneme the problem could be somewhat alleviated.

The rest of the paper is organized as follows: section 2 describes briefly the baseline algorithm used for text-dependent SR with HMMs. Section 3 describes the three different alternatives considered for performing T-Norm, section 4 describes the experimental protocol, section 5 presents experimental results, section 6 presents a discussion on the reasons for the behavior observed in the experiments, and finally, section 7 presents conclusions and future work.

2. General framework for text-dependent SR based on phonetic HMMs

The general framework used in this paper for text-dependent SR is defined by a common parameterization; a speaker-dependent *sentence* model of the utterance to be verified, a speaker-independent *sentence* model and a common way of scoring. This general framework is described in detail in [5], so we refer the interested reader to this article and will give here just a brief summary.

The front end starts with a pre-emphasis filter, after which the signal is windowed using 25 ms. Hamming windows with a window shift of 10 ms. From each window 13 Mel Frequency Cepstral Coefficients (MFCCs) are extracted (including C0), and their first and second-order differences are calculated, for a total of 39 features per frame.

A speaker-independent sentence model is built for each utterance to verify from a set of speaker-independent phonetic HMMs, a phonetic lexicon and the orthographic transcription of the sentence. The HMMs are 39 context-independent English phonetic HMM models previously trained on TIMIT. The phonetic models have 3 states, with a Bakis (left-to-right) topology with no skips.

This model will compete against a speaker-dependent sentence model built exactly in the same way but using speaker-dependent phonetic HMMs obtained from a small amount of speech (enrollment data) from that speaker. These speaker-dependent phonetic HMMs have exactly the same structure as the speaker-independent HMMs and can be obtained in different ways. We have explored three of them: performing Baum-Welch reestimation [11] of the speaker-independent phonetic HMMs on the enrollment data, adapting the speaker-independent HMMs using MLLR [6], and finally performing MLLR followed by MAP adaptation [7].

After the speaker-independent and the speaker-dependent models of the utterance have been built the utterance to verify is aligned to each of these two models using a Viterbi algorithm which produces the acoustic scores for each frame given the speaker-dependent and the speaker-independent models of the utterance. The final score is the ratio between the average score per frame obtained with the speaker-dependent model and the average score per frame obtained with the speaker-independent model. Assuming that the textual content of the utterance is the correct, the larger the score the larger the confidence the system has in verifying the speaker. This set-up models a text-prompted system where the text uttered normally coincides with the expected text.

In spite of the score normalization provided by the use of speaker-independent scores, which can be viewed as similar to a UBM (Universal Background Model), the speaker-dependent score variations and the need for speaker-independent decision thresholds usually requires the inclusion of further score normalization techniques (Z-norm, T-norm, …). In this sense we will consider that the scores obtained as described in this section are unnormalized scores. In next section we will describe three different ways to perform T-norm in this context.

3. T-Norm for text-dependent SR at the utterance, phoneme and state levels

In text-independent SR it is very common to use T-Normalization by comparing the score obtained with a test segment, not only to the model of the speaker in the test segment, but also against the models of other speakers (i.e. against a cohort of impostors).

The direct translation of this approach to text-dependent SR is what we call *Utterance-Level T-Norm*, to distinguish it from the novel T-Normalization schemes proposed in following sections. As with any T-Normalization scheme, we need to define a cohort of M speakers and compute the unnormalized scores (as described in Section 2) not only using the model of the speaker to verify but also the models for the M speakers in the cohort. After we have done this we T-Normalize the score in the usual way:

$$sc^{TNorm} = \frac{sc - \mu}{\sigma}, \qquad (11)$$

Where sc is the unnormalized score, μ and σ are the mean and the standard deviation of the scores obtained against the cohort of M speakers and sc^{TNorm} is the T-Normalized score.

With this T-Normalization scheme we T-Normalize the final scores after averaging over the whole utterance. In this sense, we are combining scores computed on very different parts of the test utterance (i.e. on different phonemes or different parts of the phonemes) which may produce scores with very different distributions. For that reason it seems to be a good idea to try to normalize the scores for similar segments before averaging the scores. We propose the use of sub-word level T-Normalization schemes in which we perform T-Normalization on averages of the acoustic scores over segments corresponding to phonemes or even HMM states within the phoneme before averaging the already T-Normalized scores over the whole utterance. We call these methods *Phoneme-Level T-Normalization* and *State-Level T-Normalization*. The idea behind these new T-Normalization schemes is relatively simple and we consider that a detailed description here is unnecessary. However, the interested reader can find a detailed description of these methods in [5].

4. YOHO experimental protocol

For the experiments we have used YOHO [3], probably the most widely used and well known benchmark for text-dependent SR system comparison and assessment. It consists of 96 utterances for enrollment collected in 4 different sessions and 40 utterances for testing collected in 10 sessions for each of a total of 138 speakers, 106 male and 32 female. Each utterance is a different set of three digit pairs (e.g. "12-34-56"). The results presented on YOHO are based on the following experimental protocol. Speaker models are trained using 6 utterances from session 1, the 24 utterances from session 1 or the 96 utterances from the 4 sessions. Our main focus was on the single session, 6 utterances, since it is the closest to what we expect to find in realistic operational conditions. Most experiments are referred to this condition. Speaker verification is performed using a single utterance from the test subset. The target scores are generated by matching each speaker model with all the test utterances from that user, leading to a total of 138 x 40 = 5,520 scores. The impostor scores are computed by comparing each speaker model with a single utterance randomly selected from those of all other users, which yields 138 x 137 = 18,906 trials. For all impostor trials the sentence models are produced using the actual text spoken to simulate a text-prompted system in which the impostors know what they have to say.

For experiments using T-Norm the experimental protocol has been slightly modified. We have considered 3 different cohort sizes for T-norm: 10 male and female speakers, 30 male and female speakers (this is the maximum we can reach with the 32 female speakers in YOHO) and all male speakers. For the 10 male and 10 female cohorts we have removed these speakers from the test. This way the number of target scores is reduced to 118 x 40 = 4,720, and the number of impostor scores to 118 x 117 = 13,806. For the 30 male and 30 female speakers and for the all male cohorts we cannot remove so many speakers from the test, so we have used Jackknife to use all trials and large (trial-dependent) cohorts with speakers not included in each trial.

5. Results

We have organized this section into three subsections. The first one compares results without score normalization using Baum-Welch and MLLR. The second presents results without normalization and with MAP. Finally, the third one focuses on the three proposed ways of performing T-Normalization, comparing them using several set-ups for the cohort.

5.1. Results with Baum-Welch and MLLR

In this section we compare MLLR adaptation and Baum-Welch re-estimation for different amounts of enrollment speech. In particular, we have compared the best results achieved by MLLR adaptation and Baum-Welch retraining for the condition of 6 utterances from the first training session, 24 utterances from the first training session, and of all 96 utterances in the 4 training sessions. Table 1 and Figure 1 show the best results obtained after an optimization performed on the number of Gaussians per state, the number of iterations of Baum-Welch re-estimation and the number of regression classes in MLLR adaptation. For Baum-Welch re-estimation the number of Gaussians per state was varied between 1 and 5 and the number of re-estimation iterations was either 1 or 4. For MLLR adaptation the number of Gaussians per state was varied between 5 and 80 in steps of 5 and the number of regression classes between 1 and 32 in power-of-2 steps. Our best results show that, even in the cases with the largest amount of data, MLLR adaptation outperforms Baum-Welch re-estimation in text-dependent speaker recognition. In fact, the difference in favour of MLLR tends to increase as the amount of enrollment material increases. The reason for this may be that the amount of enrollment material, even using the 96 utterances for training, is still very limited for Baum-Welch re-estimation. MLLR adaptation seems to be more adequate for the whole range of enrollment speech considered.

5.2. Results with MLLR plus MAP

After these experiments we tried to get more accurately speaker-adapted HMMs by performing MAP [7] adaptation after the MLLR adaptation. This yields increased speaker recognition performance (Fig. 1 and Table 1). The EER decreased by 1.04% absolute (22.6% relative improvement). This improvement comes at increased computational and storage costs (we need to store a whole new set of phonetic HMMs for each speaker, not only the transformation matrices) but in some applications we can take advantage of it. We have only performed experiments with MLLR followed by MAP for the 6 utterances enrollment condition because this is the most interesting condition for the applications we are considering currently.

5.3. Results with Utterance-Level, Phoneme-Level and State-Level T-Norm

In this section we make use of the method that produced the best results in the former sections, adaptation with MLLR followed by MAP, and focus on user enrollment with 6 utterances, which we consider the case most close to the applications we envisage. With these settings we have tested the three different schemes for T-Normalization described in section 3 with different set-ups of the cohort. Results from this extensive testing are summarized in terms of Equal Error Rate (EER) in percentage in Table 2.

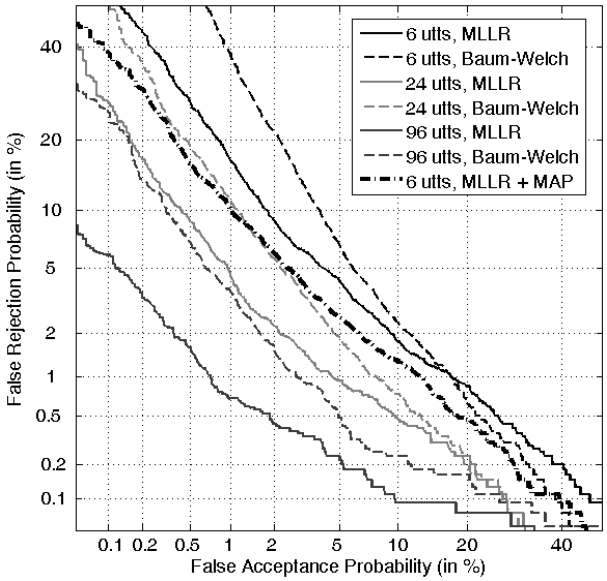

Figure 1: *DET curves with Baum-Welch re-estimation, MLLR adaptation and MLLR adaptation followed by MAP with 6, 24 and 96 utterances for enrollment.*

Table 1. *EERs (%) with Baum-Welch re-estimation, MLLR adaptation and MLLR adaptation followed by MAP with 6, 24 and 96 utterances for enrollment.*

Enrollment utterances (and sessions)	Baum-Welch	MLLR	MLLR + MAP
6 (1 session)	5,6	4,6	3,56
24 (1 session)	3,2	2,1	--
96 (4 sessions)	1,9	0,9	--

The first line of Table 2 presents results obtained with MLLR plus MAP adaptation without normalization, and serves as the baseline results. These correspond to Figure 1 but have been further detailed according to the gender in the trials. The last column of the table presents global results obtained by considering all trials, including same gender and cross gender trials.

The rest of the table is organized in blocks of three lines which represent results obtained with *Utterance-Level*, *Phoneme-Level* and *State-Level T-Norm* for the following cohorts of impostors:

- *G.I. 10m+10f:* A gender independent cohort including 10 male speakers and 10 female speakers.
- *G.D. 10m – 10f:* Two gender dependent cohorts obtained by dividing the previous cohort into two gender-dependent cohorts.
- *G.D. 30m – 30f:* Two gender-dependent cohorts with 30 speakers for each gender.
- *G.D. All male:* A male cohort including all speakers in YOHO except those involved in the trial.

For the two first cases we removed the speakers in the cohort from the test, while for the two last we used Jackknife and trial-dependent cohorts excluding speakers in the trial.

From the table we observe that *Phoneme-Level* and *State-Level T-Norm* clearly outperform *Utterance-Level T-Norm* for the smaller cohorts (10 male and 10 female), irrespective of whether the cohorts are gender-dependent or independent. In

Table 2. *T-Norm results (EERs in %) obtained on YOHO (with only 6 utterances from a single session as enrollment material) using MLLR and MAP adaptation. The table compares results obtained without normalization and with Utterance-Level, Phoneme-Level and State-Level T-Norm for different set-ups for the cohort.*

Cohort	Type of T-Norm	Gender Condition		
		Male	Female	All
NO	NO	3.90	7.26	3.56
G.I. 10m + 10f	Utterance	4.13	5.84	3.91
	Phoneme	3.21	4.76	2.98
	State	3.34	4.55	3.04
G.D. 10m – 10f	Utterance	3.53	13.85	3.64
	Phoneme	2.92	5.19	2.97
	State	3.02	4.55	2.91
G.D. 30m – 30f	Utterance	2.74	4.07	3.10
	Phoneme	2.52	4.13	2.98
	State	2.47	4.03	2.96
G.D. All male	Utterance	2.55	--	--
	Phoneme	2.43	--	--
	State	2.52	--	--

these cases, *Utterance-Level T-Norm* actually worsens the results obtained without normalization, while *Phoneme* and *State-Level T-Norm* produce important improvements. In the case of two gender-dependent cohorts with 10 male and 10 female speakers the relative improvement achieved by *State-Level T-Norm* over *Utterance-Level T-Norm* reaches 20.1% (0.73% absolute) in the all gender condition.

When we move to larger cohorts we observe that *Phoneme* and *State-Level T-Norm* still tend to perform better than *Utterance-Level T-Norm*. However, the increase of the cohort has a larger improvement effect on *Utterance-Level T-Norm* than on sub-word levels T-Norm. This reduces the difference between utterance and sub-word levels T-Norm.

6. Discussion

It is reasonable to consider that the different phonemes have different discrimination capabilities. In fact, this is the hypothesis of a recent work [12] in which the scores produced by different phonemes are combined with different weights using boosting for improved performance. In the context of T-Norm this will mean that the scores produced by different phonemes should be normalized in different ways. In fact, we have studied the impostor score distributions for different phonemes (not presented here due to space limitations) and have noticed important differences among them, which again suggest the convenience of sub-word level normalizations. Our experiments in this paper, however, have made that advantages clear particularly for small cohorts, pointing out other important advantage of sub-word score normalization schemes: their robustness to small cohorts.

7. Conclusions

In this paper we have experimented with three different ways of obtaining the speaker models from the enrollment material for a text-dependent SR system based on HMMs: Baum-Welch reestimation, MLLR adaptation and MLLR followed by MAP adaptation. Among them, we have found that MLLR followed by MAP tends to produce the best results, which are over 22.6% relatively better in terms of EER than those achieved by the second best, MLLR. We have also performed an extensive experimentation with T-Normalization methods, comparing the normal method, *Utterance-Level T-Norm*, with two novel methods, *Phoneme-Level T-Norm* and *State-Level T-Norm*. Experiments have been performed with different cohort set-ups, showing that *Phoneme-Level T-Norm* and *State-Level T-Norm* tend to perform better than *Utterance-Level T-Norm*. These differences are particularly noticeable (up to 20.1% relative improvements in EER) when small cohorts are used for T-Norm, probably due to the higher robustness to small cohorts of these new sub-word T-Norm methods compared to the normal, utterance-based T-Norm.

8. Acknowledgements

This work was funded by the Spanish Ministry of Science and Technology under project TEC2006-13170-C02-01.

9. References

[1] "National institute of standard and technology. Speaker Recognition Evaluation Home Page", http://www.nist.gov/speech/tests/spk/index.htm.

[2] T. Matsui and S. Furui, "Speaker Recognition Using Concatenated Phoneme HMMs," Proc. ICSLP, Banfl, Th.sA M.4.3 (1992).

[3] F. Bimbot, H. P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg and J. B. Pierrot, "Speaker verification in the telephone network: research activities in the CAVE project", in *Proc. Eurospeech 1997, pp. 971-974*.

[4] Hébert, M., "Text-Dependent Speaker Recognition", chapter 37 in Benesty, Sondhi and Huang (Eds.) "Handbook of Speech Processing", Springer, 2008.

[5] Toledano D. T., Esteve-Elizande C., Gonzalez-Rodriguez J., Fernandez-Pozo R. and Hernandez-Gomez L. "Phoneme and Sub-Phoneme T-Normalization for Text-Dependent Speaker Recognition", in Proc. IEEE Odyssey 2008.

[6] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation using maximum likelihood linear regression", in *Proc. Eurospeech 1995, pp. 1155-1158*.

[7] J. L. Gauvain and C. H. Lee, "Maximum a posteriori estimation for multivariate gaussian observations of markov chains," *IEEE Transactions on Speech and Audio Processing,* vol. 2, no. 2, pp. 291–298, 1994.

[8] J. Campbell and A. Higgins. Yoho speaker verification (ldc94s16). http://www.ldc.upenn.edu.

[9] Toledano D. T., Hernandez-Lopez D., Esteve-Elizalde C., Fierrez J., Ortega-Garcia J., Ramos D. and Gonzalez-Rodriguez J., "BioSec Multimodal Biometric Database in Text-Dependent Speaker Recognition", in Proc. LREC 2008 (to appear).

[10] M. Hébert and D. Boies, "T-Norm for text-dependent commercial speaker verification applications: effect of lexical mismatch", in *Proc. ICASSP 2005*, pp. 729-732.

[11] L. R. Rabiner, "A Tutorial on Hidden Markov Models", In *Proceedings of the IEEE*, vol. 77, n. 2, February 1989, pp. 257-286.

[12] Subramanya, A.; Zhengyou Zhang; Surendran, A.C.; Nguyen, P.; Narasimhan, M.; Acero, A.; "A Generative-Discriminative Framework using Ensemble Methods for Text-Dependent Speaker Verification" in Proc. ICASSP, 2007. Volume 4, 15-20 April 2007 pp. IV-225 - IV-228.