Evaluation of the Vocal Tract Length Normalization Based Classifiers for Speaker Verification

https://doi.org/10.3991/ijes.v4i4.6544

W. B. Hussein [1], S. A. Essmat [1], N. B. Yoma [2], F. Huenupan[3]
[1] The British University in Egypt, Cairo, Egypt
[2] Universidad de Chile, Santiago, Chile.
[3]Universidad de La Frontera, Temuco, Chile.

Abstract—**This paper proposes and evaluates classifiers based on Vocal Tract Length Normalization (VTLN) in a text-dependent speaker verification (SV) task with short testing utterances. This type of tasks is important in commercial applications and is not easily addressed with methods designed for long utterances such as JFA and i-Vectors. In contrast, VTLN is a speaker compensation scheme that can lead to significant improvements in speech recognition accuracy with just a few seconds of speech samples. A novel scheme to generate new classifiers is employed by incorporating the observation vector sequence compensated with VTLN. The modified sequence of feature vectors and the corresponding warping factors are used to generate classifiers whose scores are combined by a Support Vector Machine (SVM) based SV system. The proposed scheme can provide an average reduction in EER equal to 14% when compared with the baseline system based on the likelihood of observation vectors.**

Index Terms—**Principal Component Analysis, Speaker Verification, Support Vector Machine, Vocal Tract Length Normalization.**

I. INTRODUCTION

Vocal Tract Length Normalization, VTLN, is a widely used method to compensate inter-speaker variation in speaker-independent automatic speech recognition (ASR) [1]. To achieve this, VTLN tries to compensate for the effects of speaker-specific vocal tract lengths by warping the frequency axis of the power spectrum of the observation vector sequence by employing a warping factor optimized for each speaker [2]. VTLN has extensively been employed in ASR but is limited work in Speaker Verification (SV). A simple and efficient implementation can be achieved by moving the center of the filter bank in the parameterization process via the inverse frequency warping function [3]. VTLN can also be applied in the cepstral domain by using linear transformation [2].

As mentioned above, with a few exceptions, VTLN has hardly been applied to SV. In [4] authors proposed a GMM-UBM SV with multiple background model (MBM) system based on VTLN criterion for UBM training data selection. An improvement of 8% in EER can be achieved if the UBM is trained with selected mean-VTLN data when compared with training with all the data. In [5-6] authors proposed the use of a background model per each group of target speakers that were clustered by employing their vocal tract length factor as well as MLLR super-vectors in a text-dependent SV with GMM. A different approach is presented in [7] where an ASR is employed to estimate the warping factor and combine it with a GMM-UBM SV system in order to improve the SV accuracy. This scheme provided an improvement of 23% in EER when compared with the baseline system. Despite the fact that VTLN based approaches could improve SV accuracy, it has not been explored further.

Text-dependent SV task with short testing utterances have an important presence in commercial applications and is not easily addressed with methods designed for long utterances such as Joint Factor Analysis (JFA) and i-Vectors. Otherwise, VTLN is a speaker compensation scheme and with just a few seconds of speech samples can lead to significant improvements in speech recognition accuracy.

In [8] it was suggested that the VTLN warping factor could be employed for gender classification. It is well known that men and women have different warping factors, with men in general showing a higher value [7]. These results imply that the VTLN warping factor could be a criterion for discriminating clients or target speaker from impostors or non-target speaker in a SV task. Besides, new feature vectors may be generated by compensating the input utterance with VTLN warping to obtain new classifiers, and using selection and combination techniques to improve the accuracy of the entire SV system.

This paper proposes a novel scheme for the generation of classifiers based on VTLN in a text-dependent SV task with short testing utterances. In order to fuse this new classifiers, a combination scheme is performed based on SVM to improve the accuracy of the SV System.

II. SPEAKER VERIFICATION SYSTEM

In a SV system, the task is to describe the identity that is claimed by a given user. Two classes are possible: client, C1; and, impostor, C2. In the enrolling process, each user is prompted to pronounce a given number of utterances that will be employed to generate the user's speaker dependent (SD) model. In verification, the speech signal from a user that claims a given identity is compared with the corresponding SD model associated to the claimed identity. In a HMM based system, the observation vector sequence is also compared with an impostor model

[10]. This impostor model is denominated speaker independent (SI) because it is usually trained with a wide variety of users.

Given an input vector sequence $X = \{X_1, ..., X_t, ..., X_T\}$, where T is the total number of frames, and $X_t = \{x_{t1}, .., x_{tk}, ..., x_{tK}\}$ is the feature vector in the k^{th} frame, where K is the total number of features. As an output of a SV system the alignments or the sequences of states associated to each frame are obtained, for the SD model $\lambda_{SD} = \{\lambda_{SD}^{(1)}, ..., \lambda_{SD}^{(t)}, ..., \lambda_{SD}^{(T)}\}$ and for the SI model $\lambda_{SI} = \{\lambda_{SI}^{(1)}, ..., \lambda_{SI}^{(t)}, ..., \lambda_{SI}^{(T)}\}$.

A log-likelihood score for each frame and his state is computed for each sequence, $Score_{SD} = \{S_1^{SD}, .., S_t^{SD}, ..., S_T^{SD}\}$ and $Score_{SI} = \{S_1^{SI}, .., S_t^{SI}, ..., S_T^{SI}\}$. Finally, the system estimates a log-likelihood score associated to the input observation vector sequence as the difference between the SD and SI log-likelihood scores for each frame , $S = \{S_1, .., S_t, ..., S_T\}$, and the final score for every utterance is computed as $LL_{Score} = \sum_{i=1}^{T} S_i$.

III. VOCAL TRACT LENGTH NORMALIZATION

VTLN attempts to compensate for the difference among speakers′ vocal tract lengths by warping the frequency axis of the speech signal power spectrum [2]. In general, the frequency axis is scaled by a warping function with a transformation parameter α.

$$\begin{aligned} & g_\alpha : [0,\pi] \to [0,\pi] \\ & \omega \to \hat{\omega}_m(\alpha) = g_\alpha(\omega) \end{aligned} \tag{1}$$

Consider that ω_m is the central frequency of filter m in a filter-bank composed of M filters. Then $\hat{\omega}_m$ is the warped central frequency of filter m. By using the linear piece-wise warping function proposed in [3], $\hat{\omega}_m$ can be written as

$$\hat{\omega}_m(\alpha) = \begin{cases} \alpha \cdot \omega_m & \omega_m \le \omega_0 \\ \alpha \cdot \omega_0 + \dfrac{\omega_{\max} - \alpha \cdot \omega_o}{\omega_{\max} - \omega_0}(\omega_m - \omega_0) & \omega_m \ge \omega_0 \end{cases} \tag{2}$$

Where ω_{max} corresponds to the highest filter-bank frequency, α is the warping factor or parameter, and ω_0 is defined as follows

$$\omega_0 = \begin{cases} \dfrac{7}{8}\omega_{\max} & \alpha \le 1 \\ \dfrac{7}{8 \cdot \alpha}\omega_{\max} & \alpha > 1 \end{cases} \tag{3}$$

Conventional VTLN is usually implemented by generating a filter-bank per each warping factor α to be evaluated. Then, the optimal α is that one that provides the maximum likelihood of a feature vector sequence transformed with the warping function, $g_\alpha(X_r)$, where X_r is the sequence of acoustic data and $g_\alpha(X)$ is the piece-wise function defined in (4)

$$\alpha_{optimal} = \arg\max_{\alpha} \{\log[\Pr(g_\alpha(X_r) \mid Wr)]\} \tag{4}$$

Where Wr is the recognized word sequence obtained by a first recognition pass

IV. VTLN BASED CLASSIFICATION

After the conventional operational procedure of a SV system explained in Section 2.1, supplementary information can potentially be obtained by incorporating a new observation vector sequence adapted with VTLN. This modified utterance can be used as an input to the SV system, obtaining different scores when compared with those computed with the original feature vectors without compensation. The new scores, combined with the one obtained with the original observation vector sequences, can be fused by using standard techniques of feature selection and classifier combination to improve the accuracy of the SV system.

Given a SD model state alignment resulting from the first forced Viterbi pass in a text dependent SV system, an optimal α that maximizes the following function could be estimated by employing VTLN

$$\alpha_{optimal} = \arg\max_{\alpha} \{\log[\Pr(g_\alpha(X) \mid \lambda_{SD})]\} \tag{5}$$

Where is $g_a\ (X)$ the piece-wise function defined in (1), X is the input feature vector sequence and λ_{SD} is the SD model state alignment. The estimation of $\alpha_{optimal}$ can be achieved using any VTLN technique. It is reasonable to assume that if the estimated warping factor $\alpha_{optimal}$ is distant from 1, the observation vector sequence had to be compensated more to increase its likelihood with respect to the SD model. In this case, the probability that the input feature vectors was an impostor should be higher. In contrast, if $\alpha_{optimal}$ is close or equal to one, it is sensible to suppose that it corresponds to a client or target speaker. The difference with respect to $\alpha_o=1$ is calculated with $\alpha_{optimal}$ using the following equation.

$$\Delta\alpha = |1 - \alpha_{optimal}| \tag{6}$$

Where $\Delta\alpha$ is the absolute value of this distance or difference.

The compensation of the observation vector sequence with VTLN, $X(\alpha_{optimal})$, can be used as a new input to the SV system in order to obtain a new sequence of aligned SD and SI models, and a new log-likelihood score that depends on $\alpha_{optimal}$, $LL_{Score}(\alpha)$. Similarly to the $\Delta\alpha$ case, if the input observation vector sequence corresponds to a client, the VTLN compensation over the signal should be lower when compared with that estimated with an utterance from an impostor speaker. Consequently, the difference between the log-likelihood scores obtained with the original and compensated observation vector sequence should be lower for a client speaker than for an impostor. The estimation of the difference between these log-likelihoods is computed as

$$\Delta LL_{score}(\alpha_{optimal}) = |LL_{score} - LL_{score}(\alpha_{optimal})| \tag{7}$$

It can be seen that the adapted feature vector sequence with VTLN $X(\alpha_{optimal})$ could possibly be utilized to generate new criteria for classification. In Fig.1., a scheme to obtain new classifiers using VTLN is shown. The SD model (λ_{SD}) alignment obtained by the first forced Viterbi pass and the input feature sequence X are employed to estimate $\alpha_{optimal}$ that maximize (5). The adapted observation vector sequence $X(\alpha_{optimal})$ is the input to the SV system, obtaining new scores such as $LL_{Score}(\alpha)$, $LL_{SD\text{-}Score}(\alpha)$ and $LL_{SI\text{-}Score}(\alpha)$. Finally, given any VTLN method, a difference is estimated between the new and the original scores, obtaining five new classifiers.

V. EXPERIMENTS

Experiments were carried out with the Yoho database [9]. The Yoho Speaker Verification Corpus supports development, training and testing of speaker verification systems that use limited vocabulary, free-text input. The vocabulary is composed of two-digit numbers spoken continuously in sets of three. The database is divided into "enrollment" and "verification" segments; each segment contains data from all of the 138 speakers. There are four enrollment sessions per speaker and each session contains 24 utterances. Each verification segment contains 10 sessions and each session contains four utterances per speaker. The database was divided in three groups: Yoho_A, Yoho_B and Yoho_C. 92 speakers were selected for Yoho A and Yoho B. 77 of these were randomly selected for Yoho A, used for testing. While the remaining 15 speakers for Yoho B to be used in the SVM classifier, explained later. The process of random selection of users for Yoho A and Yoho B was repeated 1000 times to generate an equal number of experiments. Finally, Yoho_C, composed of 41 speakers (29 males and 12 females), was used to train the SI model.

The TD-SV system is based on HMM with forced-Viterbi algorithm [10]. While the combination of scores is based on SVM with linear Kernel [11, 12]. The SVM parameters were estimated with Yoho_B.

The procedure for training the SVM curve and the evaluation of the classifier combinations was repeated 1000 times in order to obtain a more representative result. VTLN according to [8] was implemented in the experiments reported here.

VI. RESULTS AND DISCUSSION

Fig. 2. depicts histogram of $\Delta\alpha$ with VTLN. It can be seen that the values of $\Delta\alpha$ for the clients are in average lower than the impostor. This indicates that $\Delta\alpha$ may discriminate about the two classes and can be itself a criterion for classification in SV. This implies that the client's observation vector sequence compensated with VTLN will not be very different from the original one and the log-likelihood score will be similar to the one obtained in the first round of verification. Similar results are obtained for the histogram of the difference of log-likelihood (ΔLL_{Score}) by using VTLN: in average ΔLL_{Score} are higher for the impostors than for the clients. As mentioned above, this due to the fact that the VTLN compensation for the client's feature vectors is lower than for the impostor's observation vectors. Consequently, the log-likelihood score with the adapted features is similar compared with the signal without adaptation for client speakers.

In Table 1 the individual performance for the six classification criteria are shown: the baseline criterion and the five new criteria obtained with VTLN. It can be observed that the best performance corresponds to the baseline and $LL_{Score}^{VTLN}(\alpha)$. It can also be learned that $\Delta\alpha$ achieves a low discrimination between client and impostor, and the performance is very low when compared with the baseline.

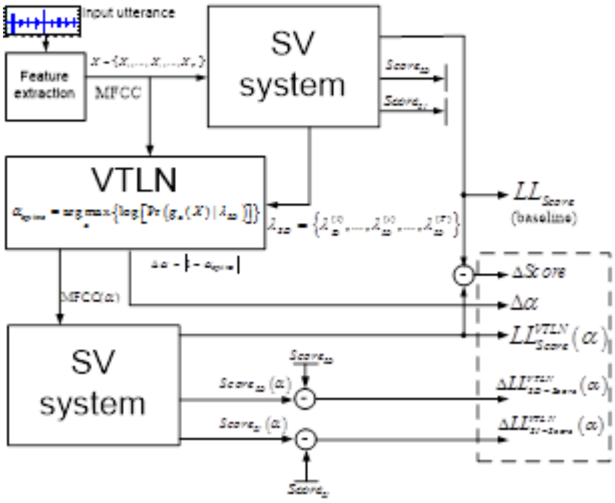
Input utterance
Feature extraction
$x=\{x_1,...,x_i,...,x_n\}$
MFCC
SV system
$Score_{SD}$
$Score_{SI}$
VTLN
$\alpha_{optimo}=\arg\max_{\alpha}\{\log[\Pr(g_\alpha(X)|\lambda_{SD})]\}$
$\lambda_{SD}=\{\lambda_{SD}^{(1)},...,\lambda_{SD}^{(i)},...,\lambda_{SD}^{(T)}\}$
LL_{Score} (baseline)
$\Delta\alpha=|1-\alpha_{optimo}|$
MFCC(α)
SV system
$\Delta Score$
$\Delta\alpha$
$LL_{Score}^{VTLN}(\alpha)$
$Score_{SD}$
$Score_{SD}(\alpha)$
$\Delta LL_{SD-Score}^{VTLN}(\alpha)$
$Score_{SI}(\alpha)$
$\Delta LL_{SI-Score}^{VTLN}(\alpha)$
$Score_{SI}$

Figure 1. The proposed scheme to obtain new classifiers with VTLN in speaker verification systems.

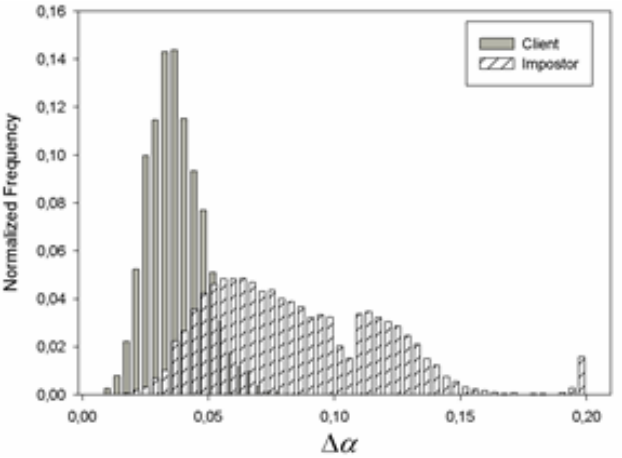

Figure 2. The Histogram for Δα using VTLN.

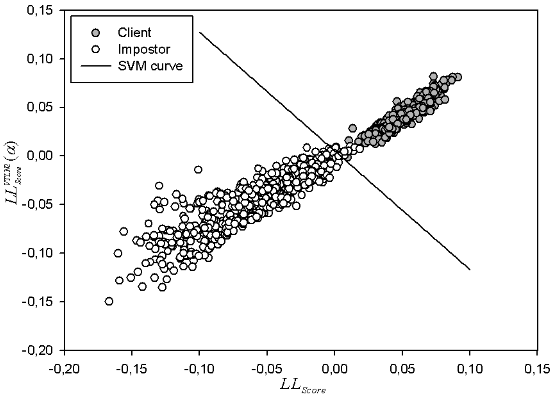

Figure 3. Discrimination between client and impostor with SVM by using LL_{Score} and $LL_{Score}^{VTLN}(\alpha)$.

Fig.3. shows the discrimination between the client and impostor using two criteria for classification, i.e. LL_{score} and $LL_{Score}^{VTLN}(\alpha)$. Results in Fig. 3 suggest that incorporating a second criterion to the baseline improves the discrimination between the client and impostor.

Fig.4. provides the histogram of improvement with respect to the baseline for the combination of LL_{Score}

(Baseline) with $LL_{Score}^{VTLN}(\alpha)$, with an average improvement of 13.9%. The results indicate that the improvement depends on the group of users selected to estimate the SVM parameters. In the worst scenario there is no improvement (0%), while in the best ones the reduction in EER can be as high as 24%.

VII. CONCLUSION

In this paper a scheme to generate new classification criteria based on VTLN in a text-dependent speaker verification task is proposed. Additional information was obtained by incorporating a new observation vector sequence computed from the original one by applying VTLN compensation. The modified observation vector sequence was used as an input to the SV system to estimate new scores regarding the original ones without the VTLN compensation. The new criteria based on VTLN were combined with the baseline one by using SVM. It is worth emphasizing that the proposed scheme can be employed with any VTLN method.

Experiments with YOHO database are presented. The performances for each classification criterion and for the selection and combination of these new classifiers with the baseline are discussed. The VTLN warping factor was also tested as a criterion for classification. However, its performance was found to be poor when compared with the baseline system (EER equal to 14.24% and 0.72%, respectively).

The propose method, using the combined scores LL_{Score} and $LL_{Score}^{VTLN}(\alpha)$ with SVM, provided an average reduction of 13.9% in EER when compared with the baseline. Also, a reduction as high as 24% in EER can be achieved for some speakers.

TABLE I.
EER FOR THE SELECTED CLASSIFICATION METHODS

Classifier	EER	Classifier	EER
Baseline	0.72	$\Delta LL_{Score}^{VTLN}(\alpha)$	13.82
$LL_{Score}^{VTLN}(\alpha)$	0.97	$\Delta LL_{SD-Score}^{VTLN}(\alpha)$	35.62
$\Delta\alpha$	14.24	$\Delta LL_{SI-Score}^{VTLN}(\alpha)$	17.03

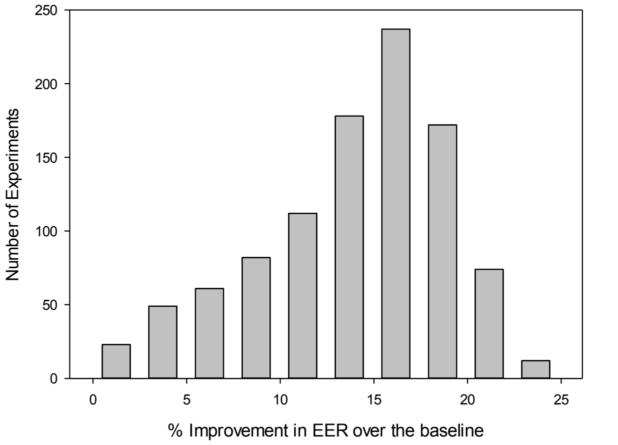

Figure 4. Histogram for improvements over the baseline system for the combination LL_{Score} and $LL_{Score}^{VTLN}(\alpha)$ with 1000 repetitions.

ACKNOWLEDGMENT

This research was partially funded by the Chilean national association for investigation in science and technology (conicyt), project Anillo Act1120, Chile.

REFERENCES

[1] J. Lööf, H. Ney, and S. Umesh, "VTLN Warping Factor Estimation Using Accumulation of Sufficient Statistics," ICASSP, Toulouse, France, pp. 1201-1204, 2006.

[2] M. Pitz, and H. Ney, 2005. "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space," IEEE Trans. Speech Audio Process., vol. 15(3), pp. 930-944, 2005. https://doi.org/10.1109/TSA.2005.848881

[3] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," IEEE Trans. Speech Audio Process., vol. 6(1), pp. 49-60, 1998. https://doi.org/10.1109/89.650310

[4] W.-Q. Zhang, Y. Shan, and J. Liu, "Multiple Background Models for Speaker Verification," IEEE Odyssey, The Speaker and Language Recognition Workshop, Brno, Czech Republic, pp. 47-51, 2010.

[5] K. Sarkar, and S. Umesh, "Investigation of Speaker-Clustered UBMs based on Vocal Tract Lengths and MLLR matrices for Speaker Verification," IEEE Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic, pp. 286-293, 2010.

[6] K. Sarkar, S. P. Rath, and S. Umesh, "Vocal Tract Length Normalization Factor Based Speaker-Cluster UBM for Speaker Verification," National Conference on Communications (NCC), Chennai, pp. 1-5, 2010. https://doi.org/10.1109/ncc.2010.5430207

[7] S. Grashey, and C. GeiBler, "Using a Vocal Tract Length Related Parameter for Speaker Recognition," IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, San Juan , pp. 1-5, 2006.

[8] N. B. Yoma, C. Garretón, F. Huenupán et al., "On Reducing Harmonic and Sampling Distortion in Vocal Tract Length Normalization," IEEE Trans. Speech Audio Process., vol. 21(1), pp. 110-121, 2013. https://doi.org/10.1109/TASL.2012.2215590

[9] J. Campbell and A. Higgins, Yoho Speaker Verification, in Linguistic Data Consortium. 1994: Philadelphia.

[10] S. Furui, "Recent advances in speaker recognition," Pattern Recognition Lett. vol. 18, 859-872, 1997. https://doi.org/10.1016/S0167-8655(97)00073-1

[11] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery. vol. 2, pp. 121-167, 1998. https://doi.org/10.1023/A:1009715923555

[12] W.M. Campbell, J.P. Campbell, D.A. Reynolds, D.A. Jones, and T.R. Leek, "High-Level Speaker Verification with Support Vector Machines," ICASSP, Montréal, Quebec, Canada. pp. 73-76, 2004. https://doi.org/10.1109/icassp.2004.1325925

AUTHORS

W. B. Hussein and **S. A. Essmat** are with the Faculty of Informatics and Computer Science, The British University in Egypt, Cairo, Egypt (e-mails: walid.hussein | sarah.akram@bue.edu.eg).

N. B. Yoma is with the Department of Electrical Engineering, Universidad de Chile, Santiago, Chile.

F. Huenupan is with the Department of Electrical Engineering, Universidad de La Frontera, Temuco, Chile.

Submitted 30 October 2016. Published as resubmitted by the authors 04 December 2016.