Intra-speaker variability compensation in speaker verification with limited enrolling data

Claudio Garreton, Nestor Becerra Yoma, Carlos Molina, and Fernando Huenupan

Speech Processing and Transmission Laboratory
Department of Electrical Engineering
Universidad de Chile, Santiago, Chile
nbecerra@ing.uchile.cl
Telephone: +56-2-678 4205 Fax: +56-2-695 3881

Abstract

In this paper a compensation method is proposed to address the problem of limited enrolling data in speaker verification. Instead of adapting the client HMM, the technique presented here modifies the verification speech signals by maximizing the *a posteriori* p.d.f. in order to optimize the reduction in intra-speaker variability. The proposed approach can lead to reductions of 38.9% and 61.8% in EER and in the integral below the false-acceptation / false-rejection ROC curve, respectively.
Index Terms: speaker verification, limited enrolling data.

1. Introduction

From the usability point of view, the enrolling procedure in speaker verification (SV) system should be fast and efficient. However, limited enrolling data leads to poorly trained models, which in turn seriously degrades the accuracy of SV engines. Moreover, additive and convolutional noise is usually one of the most important problems faced by speech and speaker recognition systems in real applications. On the other hand, several noise canceling technique have been proposed to handle additive and convolutional noise [1][2][3][4]. These noise cancellation techniques can substantially reduce the mismatch between training and testing conditions as far as additive and convolution distortion is concerned. However, they do not improve the generalization ability of trained models from the intra-speaker variability point of view.

The limited enrolling data problem in SV has been addressed by several authors using HMM adaptation methods. Those techniques adapt HMM parameters employing speech data that is input by the user in verification events after enrolling. The HMM parameters are usually re-estimated by means of applying maximum likelihood (ML) criteria [5][6], Bayesian Maximum a Posteriori (MAP) [7] adaptation, and Maximum Likelihood Linear Regression (MLLR) [8]. Those methods are classified as supervised or unsupervised depending on the requirement of human assistance to transcribe and label the adaptation data. Supervised adaptation techniques, although more effective than unsupervised approaches, are impractical on large-scale SV based services. On other hand, the unsupervised classification of adaptation data introduce an error in the HHM parameter re-estimation procedure, which in turn is propagated into further verification events.

In this paper an intra-speaker variability compensation is proposed to reduce the distortion between verification signals and the client HMM. Instead of adapting the client HMM, the approach described here modifies the verification signals using MAP estimation. The results presented show reductions of 38.9% and 61.8% in EER and the integral below the false-acceptation (FA) / false-rejection (FR) ROC curve, respectively. Due to the fact that the client HMM is not modified, the error caused by misclassification of adaptation data is avoided. Moreover, the proposed compensation scheme also leads to a noise removal effect. Finally this approach has not been found in the specialized literature.

2. Intra-speaker variability modeling

In the text-dependent SV task considered here, each utterance is processed with the forced-Viterbi algorithm in order to estimate the normalized log likelihood $\log L(O)$ [9]:

$$\log L(O) = \log P(O/\lambda_{SD}) - \log P(O/\lambda_{SI}) \quad (1)$$

where O is the observation sequence; and, $P(O/\lambda_{SD})$ and $P(O/\lambda_{SI})$ represent the likelihood related to the speaker dependent (λ_{SD}) and independent (λ_{SI}) models, respectively. Both models, λ_{SD} and λ_{SI}, correspond to the sequence of triphone HMM's that compose the testing sequence O. In order to estimate the false-rejection and false-acceptance error curves, the normalized log likelihood $\log L(O)$ is divided by the number of frames (T) in the verification utterance: $\log L(O)' = \frac{\log L(O)}{T}$. It is

worth highlighting that λ_{SD} is estimated with the enrolling data pronounced by the client, and λ_{SI} is estimated with a set of impostors.

Given a state (*s*) in λ_{SD} and the enrolling data, the intra-speaker variability is modeled in this paper as the averaged distance between the mean of the observation probability function associated to *s* and the frames allocated to this state as a result of the forced Viterbi alignment. This alignment associates a state within the HMM sequence to every frame. As a consequence, the state allocated to frame *O(t)* is denoted by *s(t)*. If $\mu_{s(t)}$ is the vector mean of the observation probability of state *s(t)*, the distance between frame *O(t)* and *s(t)*, *d(t)*, is expressed by:

$$d(t) = \sqrt{\sum_{n=1}^{N} \left(\mu_{s(t),n} - O(t,n)\right)^2} \tag{2}$$

where $\mu_{s(t),n}$ and $O(t,n)$ indicate the nth coefficient in $\mu_{s(t)}$ and $O(t)$, respectively; and, *N* is the number of parameters. In order to obtain the intra-speaker variability p.d.f., the histogram of *d(t)* was estimated using enrolling utterances from an evaluation database composed of 13 speakers after training the speaker dependent HMM's. The resulted histogram is shown in Fig. 1. As can be seen in this figure, the p.d.f. of *d(t)* can be modeled with a gamma distribution $\Pr(d)$ [10]:

$$\Pr(d) = A \cdot \exp(-\alpha d) \cdot d^{\,p-1} \tag{3}$$

where $\alpha = \frac{E[d]}{Var[d]}$; $p = \frac{E[d]^2}{Var[d]}$; *A* is a normalizing term; and, $E[d]$ and $Var[d]$ are the mean and variance of the histogram of *d(t)*, respectively. To simplify the notation, the argument *t* was withdrawn from *d(t)* in (3).

3. Feature Compensation

If $\tilde{O}(t)$ and $O(t)$ denote the compensated and observed frames, respectively, the compensation is expressed with:

$$\tilde{O}(t) = O(t) + \left[\Delta O(t)\right]^{optimal} \tag{4}$$

where $\left[\Delta O(t)\right]^{optimal}$ is the correction component at instant *t*. $\left[\Delta O(t)\right]^{optimal}$ is modeled here as a fraction of the multivariate vector distance between $O(t)$ and $\mu_{s(t)}$:

$$\left[\Delta O(t)\right]^{optimal} = \left[K(t)\right]^{optimal} \cdot \left[\mu_{s(t)} - O(t)\right] \tag{5}$$

where $\left[K(t)\right]^{optimal}$ represents the optimal fraction of the vector distance $\left[\mu_{s(t)} - O(t)\right]$.

The compensation component $\left[\Delta O(t)\right]^{optimal}$ is estimated by maximizing the a posteriori p.d.f. $\Pr\left[\Delta O(t)/O(t), s(t)\right]$. Using the Bayes theorem, the maximization of $\Pr\left[\Delta O(t)/O(t), s(t)\right]$ can be expressed as [11]:

$$\left[\Delta O(t)\right]^{optimal} = \arg\max_{\Delta O(t)} \left\{\Pr\left[\Delta O(t)/O(t), s(t)\right]\right\} = \arg\max_{\Delta O(t)} \left\{\frac{\Pr\left[O(t)/\Delta O(t), s(t)\right] \cdot \Pr\left[\Delta O(t)/s(t)\right]}{\Pr\left[O(t)/s(t)\right]}\right\} \tag{6}$$

As can be seen in (6), $\Pr\left[O(t)/s(t)\right]$ does not depend on ΔO_t. Also, $\Pr\left[O(t)/\Delta O(t), s(t)\right]$ is equivalent to $\Pr\left[O(t)/\tilde{\mu}_{s(t)} = \mu_{s(t)} - \Delta O(t), s(t)\right]$ due to the model in (4), where $\tilde{\mu}_{s(t)}$ would be the mean of the compensated observation probability of state *s(t)*. Consequently, $\Pr\left[O(t)/\Delta O(t), s(t)\right]$ can be written as $\Pr\left[O(t) + \Delta O(t)/\Delta O(t), s(t)\right]$, which in turn is equal to $\Pr\left[O(t) + \Delta O(t)/s(t)\right]$. Moreover, $\Delta O(t)$ is modeled with $\left|\Delta O(t)\right|$ in $\Pr\left[\Delta O(t)/s(t)\right]$, which in turn is supposed independent of *s(t)* and is replaced with $\Pr\left[\left|\mu_{s(t)} - \tilde{O}(t)\right|\right]$ as indicated in (3). Then, the optimization in (6) is reduced to:

$$\left[\Delta O(t)\right]^{optimal} = \arg\max_{\Delta O(t)} \left\{\Pr\left[\left|\mu_{s(t)} - O(t) - \Delta O(t)\right|\right] \cdot \Pr\left[\tilde{O}(t)/s(t)\right]\right\} \tag{7}$$

Replacing $\Delta O(t)$ with $K(t) \cdot \left[\mu_{s(t)} - O(t)\right]$ as shown in (5), the maximization expression (7) is equivalent to:

$$\left[K(t)\right]^{optimal} = \arg\max_{K(t)} \left\{\Pr\left[\left|\left(1 - K(t)\right) \cdot \left(\mu_{s(t)} - O(t)\right)\right|\right] \cdot \Pr\left[\tilde{O}(t)/s(t)\right]\right\} \tag{8}$$

In this paper the speaker-dependent observation probability $\Pr\left[\tilde{O}(t)/s(t)\right]$ is modeled with a single Gaussian with diagonal covariance matrices. If $\sigma^2_{s(t),n}$ corresponds to the variance of coefficient *n* in $\Pr\left[\tilde{O}(t)/s(t)\right]$, then (8) can be written as:

$$\left[K(t)\right]^{optimal} = \arg\max_{K(t)} \left\{ \begin{array}{l} A \cdot \left(\left\|\left[1 - K(t)\right] \cdot \left[\mu_{s(t)} - O(t)\right]\right\|\right)^{p-1} \\ \cdot \exp\left(-\alpha \left[\left\|\left[1 - K(t)\right] \cdot \left[\mu_{s(t)} - O(t)\right]\right\|\right]\right) \\ \cdot \exp\left[-\frac{1}{2} \sum_{n=1}^{N} \frac{\left(O(t,n) + K(t) \cdot \left[\mu_{s(t),n} - O(t,n)\right] - \mu_{s(t),n}\right)^2}{\sigma^2_{s(t),n}}\right] \end{array} \right\} \tag{9}$$

In the Log domain (9) can be expressed as:

$$\left[K(t)\right]^{optimal} = \arg\max_{K(t)} \left\{ \begin{array}{l} \log(A) + (p-1)\cdot\log\left(\left|\left[1-K(t)\right]\cdot\left[\mu_{s(t)} - O(t)\right]\right|\right) \\ -\alpha\cdot\left(\left|\left[1-K(t)\right]\cdot\left[\mu_{s(t)} - O(t)\right]\right|\right) \\ -\frac{1}{2}\sum_{n=1}^{N}\frac{\left(O(t,n)+K(t)\cdot\left[\mu_{s(t),n} - O(t,n)\right]-\mu_{s(t),n}\right)^2}{\sigma^2_{s(t),n}} \end{array} \right\} \quad (10)$$

Computing the partial derivate with respect to *K(t)* and setting it to zero:

$$\left[1-K(t)\right]\cdot\sum_{n=1}^{N}\left[\frac{\left(\mu_{s(t),n}-O(t,n)\right)^2}{\sigma^2_{s(t),n}}\right] + \alpha\cdot\left[\left|\mu_{s(t)}-O(t)\right|\right] - \frac{p-1}{1-K(t)} = 0 \quad (11)$$

This quadratic equation provides two solutions:

$$\left[K(t)\right]^{optimal} = 1-\frac{1}{2}\cdot\left(\frac{-\alpha\cdot\left(\left|\mu_{s(t)}-O(t)\right|\right)}{\Omega(t)} \pm \sqrt{\left[\frac{\alpha\cdot\left(\left|\mu_{s(t)}-O(t)\right|\right)}{\Omega(t)}\right]^2 + \frac{4\cdot(p-1)}{\Omega(t)}}\right) \quad (12)$$

where $\Omega(t) = \sum_{n=1}^{N}\frac{\left(\mu_{s(t),n}-O(t,n)\right)^2}{\sigma^2_{s(t),n}}$ and the solution $\left|K(t)\right| \geq 1$ was discarded. Finally, the compensation scheme was applied as follows:

$$\left[\Delta O(t)\right]^{optimal} = \begin{cases} \left[K(t)\right]^{optimal}\cdot\left[\mu_{s(t)} - O(t)\right], & \text{if } \left|\mu_{s(t)} - O(t)\right| \leq R \\ 0, & \text{otherwise} \end{cases} \quad (13)$$

where *R* is a threshold that defines a compensation region.

4. Experiments

The FA and FR curves [needed to compute the Equal Error Rate (EER)] were estimated with 31 speakers (11 males and 20 females) and the speaker-independent HMM, used in the likelihood normalization (1), was trained with 150 speakers. All the speech signals were recorded on the telephone line. The HMM's were trained with the Baum–Welch algorithm. Each speaker pronounced his first and family names eight times (3 for enrolling and 5 for verification) and the corresponding impostors repeated the client's first and family names one time each. The impostor universe for a given client is defined by the speakers with the same genre. As a consequence, every speaker-dependent HMM was trained with only three utterances. FR curves were estimated with (11 male-speakers + 20 female-speakers) x 5 verification utterances = 155 signals. FA curves were estimated with (11 male-speakers) x 10 impostor signals plus (20 female-speakers) x 19 impostor signals = 490 signals. The distribution of *d* defined in (2) and shown in Fig.1 was estimated with an evaluation database composed of 13 speakers that were different from the testing database.

Enrolling and verification utterances are decomposed as a sequence of triphones. Thirty-three cepstral coefficients are computed per frame: the frame energy plus ten static coefficients and their first and second time derivatives. Each triphone was modeled with a three-state left-to-right HMM topology without skip-state transition, with one multivariate Gaussian density per state in speaker-dependent models, and eight multivariate Gaussian densities per state in the speaker- independent model. Both models employed diagonal covariance matrices. The baseline system gave an EER equal to 11.4%. Results are presented in Table 1-2 and Figs. 2-4.

5. Discussions and conclusion

According to Fig. 2 and Table 1, the proposed compensation method can lead to reductions as high as 38,9% in EER. Although the reduction in EER is highly dependent on *R* in (13), Figure 2 shows that is a wide range of *R* where the scheme presented here provides significant improvements in speaker verification accuracy. As can be seen in Fig. 3 and Table 2, the integral below the ROC curve with *R* in (13) equal to 40 is 61,8% lower than the integral below the ROC curve given by the baseline system. The improvements in the discrimination ability can also be observed in Fig. 4 where the FR and FA curves obtained by the compensation method are compared with those provided by the baseline system. Besides the reduction in EER, Figure 4 also suggests that the separation between the FR and FA curves is increased by the technique described here.

The compensation scheme in (13) tends to reduce the distance between frames and states when *R* increases, which in turn makes the FR error rate decrease. On the other hand, there is a wide range of values for *R* where the FA error rate also decreases. This may be due to the fact that the compensation method also accounts for a reduction of the mismatch between training and testing conditions. To improve the accuracy of the proposed compensation method by including the dependence on phonetic class and by incorporating a model for the effect of *R* can be proposed as future research.

6. References

[1] Gong, Y. "Speech recognition in noise environments: A survey". Speech Comm, V.16, pp.261-291, 1995.

[2] Berouti, M. et al. "Enhancement of speech corrupted by acoustic noise". Proc. ICASSP, 1979.

[3] Furui, S. "Cepstral analysis technique for automatic speaker verification". IEEE Trans. on SAP, Vol. 29, No.2, pp.254-272, 1982.

[4] Hermansky, H. et al. "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)". Eurospeech 91, pp.1367-1370.

[5] Juang, B. "Maximum-likelihood estimation for mixture ultivariate stochastic observations of Markov chains". AT Bell Laboratories Technical Journal, pages 1235-1249, 1985

[6] Yu, Kin and Mason, J.S. "On-line incremental adaptation for speaker verification using maximum likelihood estimates of CDHMM parameters", In *ICSLP-1996*, 1752-1755, 1996.

[7] Gauvain, J. L. and Lee, C. H. "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains". IEEE Trans. Speech Audio Processing, vol. 2, pp. 291–298, 1994.

[8] Leggetter, C. and Woodland, P. "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171– 185, 1995.

[9] Furui, S. "Recent advances in speaker recognition", Pattern Recognition Letters 18, pp. 859-872, 1997.

[10] Rao, C. R. "Linear statistical inference and its applications", John Wiley and Sons, 1965.

[11] Yoma,N.B. et al. "Unsupervised reduction of intra-speaker variability in speaker verification". Submitted to IEEE Signal Processing Letters, 2006.

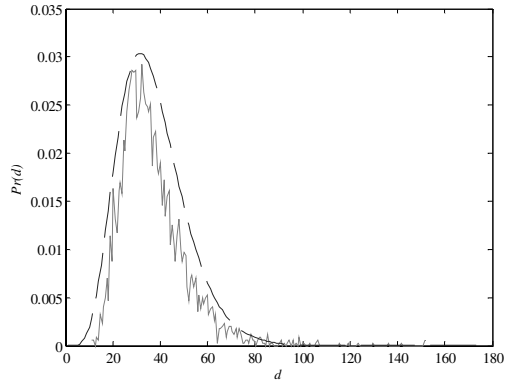

Figure 1: Distribution of *d(t)* as defined in (2): observed histogram (——); and, approximated gamma function Pr(*d*) (- - -).

R^2	EER	Improvement
0 (Baseline)	11,4%	0,0%
1000	8,0%	29,8%
1600	7,0%	38,9%
1800	8,0%	29,8%
2500	9,8%	14,0%

Table 1: EER (%) vs R^2 as defined in (13) employing the compensation method proposed here.

R^2	ROC Area	Reduction
0 (Baseline)	487,2	0,0%
1000	268,1	45,0%
1600	186,3	61,8%
1800	217,6	55,3%
2500	399,1	18,1%

Table 2: Integral below the ROC curve vs R^2 as defined in (13) employing the compensation method proposed here.

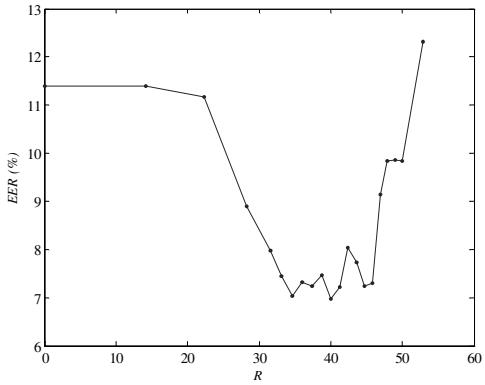

Figure 2: EER (%) vs *R* as defined in (13) employing the compensation method proposed here.

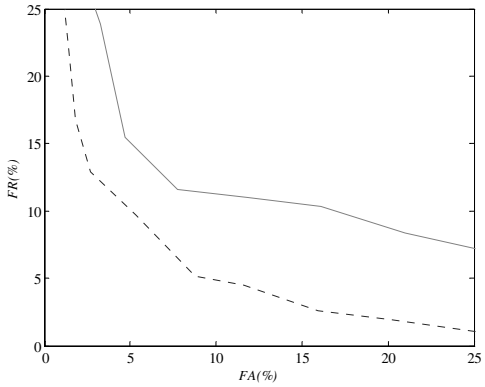

Figure 3: ROC curve given by the baseline system (——) and by the compensation method with *R*=40 (- - -).

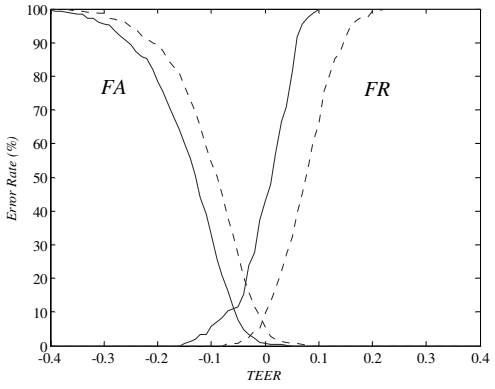

Figure 4: FA and FR curves given by the baseline system (——) and by the compensation method with *R*=40 (- - -).