

VARIANCE FLOORING, SCALING AND TYING FOR TEXT-DEPENDENT SPEAKER VERIFICATION

H. Melin and J. Lindberg

KTH, Centre for Speech Technology
Drottning Kristinas väg 31, SE-100 44 Stockholm, Sweden
{melin,lindberg}@speech.kth.se
http://www.speech.kth.se/ctt

ABSTRACT

The problem of how to estimate variance parameters in client models from scarce data is addressed in the context of text-dependent, HMM-based, automatic speaker verification. Variance flooring and variance scaling are investigated as two alternative estimation techniques and are used with or without variance tying on the state level to reduce the number of parameters to estimate. The best results are achieved with no tying and a variance flooring method where the floor to a variance vector in a client model is proportional to the corresponding variance vector in a gender-dependent, multi-speaker, non-client model. Further, variance tying reduces storage requirements considerably without much loss in recognition accuracy. It is also confirmed from a previous study that re-using non-client variances has comparable performance to variance flooring and is much simpler. Comparisons are made on three large telephone quality speech corpora.

1. INTRODUCTION

In practical applications, Automatic Speaker Verification (ASV) systems are generally used in contexts where very few client enrollment data are available. One problem with using small training data sets is the risk of over-training, that is, parameters of the client model are over-fitted to the particular training data. Especially variance parameters are susceptible to over-fitting: a variance estimated from only a few data points can be very small and might not be representative of the underlying distribution of the data source.

The maximum likelihood (ML) principle is often used in training parameters of continuous density hidden Markov models (HMM). The most general implementation of that principle (the EM-algorithm) consists in optimizing all parameters of the HMM, including means and variances of state pdfs. With sparse training data from a client, speaker variances tend to be over-trained [1].

In previous work [2] we compared two modifications to the EM algorithm for HMM training on sparse data in the context of text-dependent speaker verification. The first approach used client-independent variances [3]. Variances were copied from a gender-dependent, multi-speaker non-client model and were kept fixed while the EM-algorithm was applied to means and mixture weights. In the second approach, variances were trained but they were floored after each iteration of EM. Three variants of the variance flooring method with different resolution were tried and it was found that the one with the highest resolution performed best. In this approach, the floor

for the variance vector of a given Gaussian mixture component is proportional to the corresponding variance vector in the non-client model. The optimal scaling factor for this kind of variance flooring was found to be around 1.1, which means that all variances were actually larger than the client-independent variances. It was also found that performance with client-independent variances was near that of floored variances and much simpler.

In this paper we first generalize the approach with client-independent variances to let client variances be proportional to non-client variances. We refer to this approach as *scaled variances*. The original approach is a special case with scale factor 1.0. We look empirically at verification error rate as a function of the scale factor to see if there is a minimum for some value. Secondly, we reduce the number of variance parameters by tying variances across mixture components within each state. The client models have eight mixture components per state, and by tying variances within states we reduce the number of variance parameters by a factor eight. We compare tied-variance models with the original ones for the various variance estimation methods to see if the smaller number of variances can be more robustly trained.

The experiments are made with three separate telephone quality databases: Gandalf [4], SESP [5] and Polycost [6]. The recognition tasks are slightly different, but are all some form of text-dependent task using digits. This paper is organized as follows: In section 2 the speaker verification system is described as well as the various variance estimation techniques. Section 3 describes the test protocols for the three databases and section 4 presents the results. Section 5 contains a summary and conclusions.

2. SYSTEM DESCRIPTION

The text-dependent ASV system based on word-level HMMs is built on a generic platform for speaker verification systems called GIVES (General Identity Verification System) [8]. The input signal is pre-emphasized and divided into one 25.6 ms frame each 10 ms and a Hamming window is applied. For each frame a 12-element cepstral vector and an energy term is computed, and they are appended with first and second order deltas. Cepstral mean subtraction is applied to the 13 static coefficients. Cepstral vectors are computed from a 24-channel, FFT-based, mel-warped, log-amplitude filterbank between 300-3400 Hz followed by a cosine transform and cepstral liftering. The energy term is the 0th cepstral coefficient. The total vector dimension is 39.

A speaker model has 10 word-level left-to-right HMMs, one for each digit. Each HMM have two states per phoneme, a mixture of eight Gaussians per state and diagonal covariances.

A non-client multi-speaker model is used for log-likelihood normalization on a per-word basis. Each word score is further divided by the number of frames in the word segment, and finally averaged over words in the utterance. Non-client model HMMs are also left-to-right and have the same size as the client HMMs.

The non-client model is selected individually for each client and each word during enrollment as one of two competing gender-dependent multi-speaker models, with no *a priori* information on the gender of the client. When training the client model, the best matching multi-speaker model is copied as a seed for the client model.

The HMM implementation is based on HTK [9]. Each client and multi-speaker HMM is trained independently, and the system depends on explicit segmentation of the input speech into words during both enrollment and test. The segmentation is produced by a speech recognizer working in forced alignment mode given the expected utterance [8].

2.1 Parameter estimation

Client model means and mixture weights are always estimated from enrollment data with the ordinary EM equations while transition probabilities are kept constant. Client variances are estimated with one of two alternative methods. To define those methods we denote as σ_{ijk}^2 the variance vector of client model i , state j and mixture component k ; as s_{ijk}^2 the corresponding variance vector of the seed model; and as α or γ a scalar, system-global scale factor. In the first method, referred to as *scaled variances*, client variances are inferred directly from the seed variances, Eq. (1), and no training on enrollment data is involved. The second method is *variance flooring* [2] where variances are trained from enrollment data with a constraint on the minimum variance as given by Eq. (2). Note that with $\gamma=0$ this method converges to the original EM algorithm.

$$\sigma_{ijk}^2 = \alpha \cdot s_{ijk}^2 \quad (1)$$

$$\sigma_{ijk}^2 \geq \gamma \cdot s_{ijk}^2 \quad (2)$$

2.2 Tied variances

To reduce the number of parameters to train, the variances of a set of state distributions can be tied to a single vector. We use a letter-pair $vs=a/b$ to indicate the “level” of tying, where a indicates tying in client model and b in non-client model. Letters a and b can take symbols in an ordered alphabet $\Omega=\{X,S\}$, where X indicates one variance vector per *mixture component* (no tying) and S one vector per *state*. With $vs=S/S$, equations (1) and (2) are still valid if we remove index k . If $vs=S/X$, on the other hand, we need to compute a state-variance from mixture component variances in the non-client model. While this can be done in many ways, we use a linear combination of component variances. Eq. (3) and (4) are then our modifications of Eq. (1) and (2) for the case $vs=S/X$, where c_k is the mixture weight for component k .

$$\sigma_{ij}^2 = \alpha \cdot \sum_k c_k \cdot s_{ijk}^2 \quad (3)$$

$$\sigma_{ij}^2 \geq \gamma \cdot \sum_k c_k \cdot s_{ijk}^2 \quad (4)$$

2.3 Tied variance floors

The “resolution” of a variance floor discussed in [2] can be described in the same framework as in the previous section if the variance floor vector is viewed as a tied vector. We can then define another letter-pair vf to denote the tying level of the variance floor. The variable vf takes values from the same alphabet Ω . Since it was concluded in [2] that higher resolution flooring is better than lower, only $vf=vs$ is used in this paper. Hence, when variances are tied within states variance floors are state-dependent, and when variances are not tied variance floors are mixture component-dependent.

3. DATABASE AND PROTOCOL

We have used the same three databases [4][5][6] as in [2]. The protocols are also the same, except with Polycost where we use version 2.0¹ of baseline experiment 2. With this protocol the enrollment set is smaller than with version 1.0 and contains only two repetitions of each digit instead of eight. The test set has also been reduced to have only impostor attempts with speakers from the same language group [7]. Table 2 summarizes the main features of databases and protocols. All databases contain digital telephony speech recorded through an ISDN interface. The notation used for enrollment sets is $NsMh-T$, where N is number of sessions, M number of handsets, and T is the approximate (effective) amount of speech in minutes. The norm for the amount of speech is Gandalf where 25 five-digit sequences are estimated to one minute of speech (one digit is then 1/2 second).

4. EXPERIMENTS AND RESULTS

Results are presented in terms of equal-error-rates (EERs) based on same-sex impostor attempts and a client-independent *a posteriori* threshold. In all presented figures results for the variance scaling case with $\alpha=1$ are included as a baseline for comparison across figures.

Figure 1 shows the error rate as a function of γ for variance flooring and three cases of variance tying, vs : X/X, S/S and S/X. To evaluate the performance of the variance flooring method relative to our baseline, we treat the Gandalf development set as our development database and determine an optimal scaling factor $\hat{\gamma}$ for each tying level from this data. We then treat the other three data sets as our evaluation database and compare the error rate achieved with $\hat{\gamma}$ on the three evaluation sets to the baseline. The average error rates on the evaluation data are shown in the right-most column of Table 1. In the same table we also include the corresponding error rates for an *a posteriori* optimal choice of γ for each individual data set.

The purpose of tying variances across a set of mixture components is to reduce the number of variance parameters so the remaining parameters can be robustly estimated from enrollment data. One can therefore expect the relative improvement from variance scaling to variance flooring to be larger for tied variances than for non-tied variances. This is not the case in Table 1 for $vs=S/S$ relative to X/X. In terms of absolute EERs, differences are small between S/S and X/X.

¹ <http://circwww.epfl.ch/polycost>

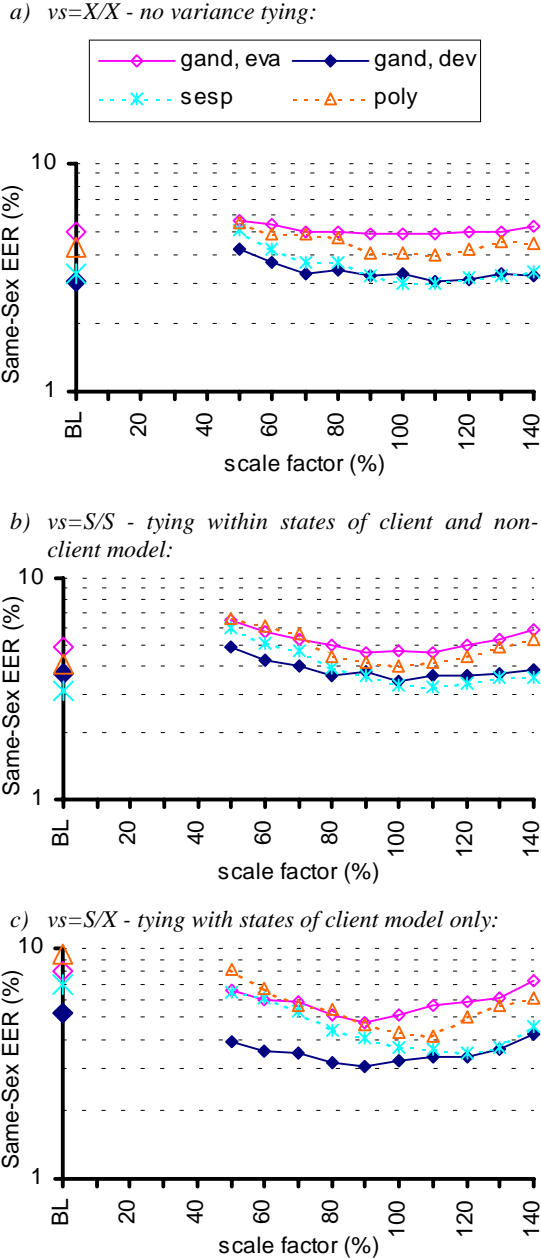


Figure 1. Same-sex EER as a function of the variance flooring factor γ for the four database sets and for three levels of variance tying (vs): a) X/X , b) S/S and c) S/X . In all charts the baseline case with client-independent scaled variances and $\alpha = 1$ is included at the left ('BL').

The motivation for $vs=S/X$ is to allow for a high modeling accuracy in the non-client model for which there is usually much data available, while having a more coarse but robust model for the client for which there are always little available data. The poor performance of scaled variances in Figure 1 indicates that the computation of state variances from the non-client variances in Eq. (3) is not good for predicting the state variances of the client model. An alternative approach would be to use a multi-speaker model with tied variances in parallel to the one with non-tied variances, and to take the seed variances (s_{ij}^2) from the former while using the latter for score normalization.

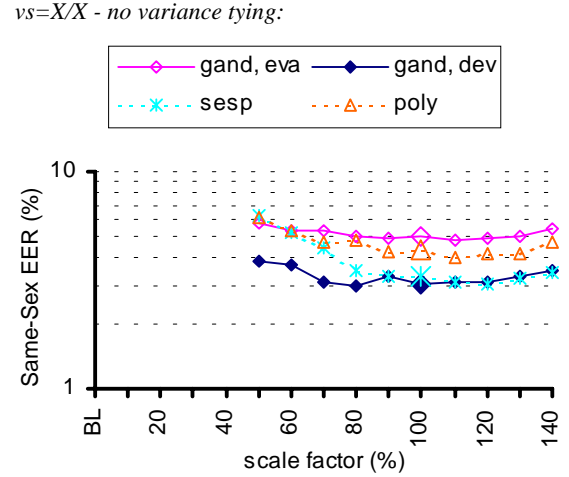


Figure 2. Same-sex EER as a function of scale factor α for client-independent scaled variances. The baseline case is $\alpha = 1$.

a) flooring

tying level	baseline	a posteriori γ	a priori $\hat{\gamma}$
X/X	4.22 %	3.94 %	3.95 % ($\hat{\gamma} = 1.1$)
S/X	8.11 %	3.97 %	4.53 % ($\hat{\gamma} = 0.9$)
S/S	4.01 %	3.95 %	4.00 % ($\hat{\gamma} = 1.0$)

b) scaling

tying level	baseline	a posteriori α	a priori $\hat{\alpha}$
X/X	4.22 %	3.96 %	4.44 % ($\hat{\alpha} = 0.8$)

Table 1. Average EER for a) variance flooring and b) variance scaling. Baseline is client-independent scaled variances with $\alpha = 1$. For the third column, the scale factor was chosen a posteriori for each individual data set. For the last column, a single scale factor was chosen based on the Gandalf development set and used as an a priori factor with the other data sets. All averages are taken over the three other data sets (Gandalf evaluation, SESP and Polycost).

There is a stronger correlation between error curves in Figure 1 for the two Gandalf sets than between those and the corresponding SESP and Polycost curves. This is reasonable since the choice of a good scaling factor for the various methods may depend on relationships between training data for non-client models and data for client enrollment and test; and since those are different for the databases we use. It is therefore likely that better predictions could have been made from development sets especially designed for the SESP and Polycost sets respectively. Table 1 includes results for an a posteriori, per-database choice of scaling factor that give a hint on what results could be achieved with such development sets. The table shows that results with a priori and a posteriori choices of γ are very similar in the X/X and S/S cases.

Figure 2 shows error rate as a function of scaling factor α for variance scaling and $vs=X/X$. Curves are similar to the corresponding curves in Figure 1a and in both figures there are minima for a scale factor around 1.1.

5. CONCLUSION

We have extended a previous study [2] on variance flooring techniques in text-dependent, HMM-based speaker verification to include tied variances and variance scaling. One advantage of using tied variances is reduced storage requirements. With variances tied across 8 mixture components within each state, 30% of the size is saved. Another expected advantage is that fewer parameters can be more robustly estimated, but no positive effect was observed from this, and recognition accuracy with and without tying was comparable. A possible further extension of this work is to look at more flexible sets of mixture components as targets for variance tying.

6. REFERENCES

- [1] Bimbot F., Hutter H.-P., Jaboulet C., Koolwaaij J., Lindberg J. and Pierrot J.-B. "An Overview of the CAVE Project Research Activities in Speaker Verification". Proc. Speaker Recognition and its Commercial and Forensic Applications, Avignon, France, pp 215-220, April 1998.
- [2] Melin H., Koolwaaij J., Lindberg J. and Bimbot F. "A Comparative Evaluation of Variance Flooring Techniques in HMM-based Speaker Verification". *International Conference on Spoken Language Processing*, Sydney, Australia, pp 1903-1906, Dec. 1998.
- [3] Matsui T. and Furui S. "Concatenated Phoneme Models for Text-Variable Speaker Recognition". *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Minneapolis, USA, pp 391-394, April 1993.
- [4] Melin H. "Gandalf - A Swedish Telephone Speaker Verification Database". *International Conference on Spoken Language Processing*, Philadelphia, USA, pp. 1954-1957, Oct. 1996.
- [5] Boves L., Bogaart T. and Bos L. "Design and Recording of large data bases for use in speaker verification and identification". *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, Martigny, Switzerland, pp 43-46, April 5-7, 1994.
- [6] Petrovska D., Hennebert J., Melin H. and Genoud D. "POLYCOST: A Telephone-Speech Database for Speaker Recognition". *Proc. Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France, pp 211-214, April 20-23, 1998.
- [7] Nordström T., Melin H., Lindberg J. "A Comparative Study of Speaker Verification Systems using the Polycost Database". *International Conference on Spoken Language Processing*, Sydney, Australia, pp 1359-1362, Dec. 1998.
- [8] Melin H. "On Word Boundary Detection in Digit-based Speaker Verification". *Proc. Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France, pp 46-49, April 1998.
- [9] Young S., Odell J., Ollason D., Valtchev V. and Woodland P. *The HTK Book (for HTK version 2.1)*. Entropic Cambridge Research Laboratory, 1997.

Test database Set		Gandalf		Polycost	SESP
		dev-set	eval-set		
Task	language	Swedish		English	Dutch
	native speakers	100 %		~15%	100 %
	enrollment	1s1h-1.0		2s1h-0.2	4s2h-0.9 ⁺
	password	2 x 4 digits		10 digits	14 digits
Test population	clients	22 / 18	24 / 18	61 / 49	21 / 20
	impostors	23 / 18	58 / 32	61 / 49	21 / 20
	total number of true-speaker tests	927	886	664	1658
	false-speaker tests (same-sex)	790	1926	824	763
Non-client population	off-line database	SpeechDat		Polycost	Polyphone
	speakers	399 / 561		11 / 11	24 / 24
	total time (approx.)	5 h		0.5 h	0.3 h
	examples per digit and speaker	4		19	5

Table 2. Summary of main features of the three databases and their protocols. Number of speakers are given as #male/#female. ⁺The number of handsets is an estimate. This enrollment set is referred to as *G* in previous literature [1].