A Comparative Evaluation of Variance Flooring Techniques in HMM-based Speaker Verification

Melin H.[*], Koolwaaij J.W.[+], Lindberg J.[*], Bimbot F.[#]

[*]KTH, Dept. of Speech, Music and Hearing, Stockholm, Sweden; {melin,lindberg}@speech.kth.se
[+]KUN, Dept. of Language & Speech, Nijmegen, The Netherlands; koolwaaij@let.kun.nl
[#]IRISA / CNRS & INRIA – Sigma2, Rennes, France; bimbot@irisa.fr

ABSTRACT

The problem of how to train variance parameters on scarce data is addressed in the context of text-dependent, HMM-based, automatic speaker verification. Three variations of variance flooring is explored as a means to prevent over-fitting. With the best performing one, the floor to a variance vector of a client model is proportional to the corresponding variance vector in a non-client multi-speaker model. It is also found that adapting the means and mixture weights from the non-client model while keeping variances constant works comparably to variance flooring and is much simpler. Comparisons are made on three large telephone quality 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 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].

One way to modify the EM-algorithm is to impose a lower bound on variance parameters, a *variance floor*. With this method any given variance value will have its corresponding floor value as a lower bound during iterations of EM. A new problem is then how to compute this floor value. Within the CAVE-project a method to compute variance floors is suggested [1]. With this method, all variance vectors of all HMMs in a speaker model share one flooring vector. This vector is estimated as the variance over some calibration data set multiplied by a constant variance-flooring factor. The calibration data set can be for instance the same data used to train non-client models.

Variance flooring can be implemented with several levels of "resolution" in up to three "dimensions". The first dimension is the *vector index*, where resolution can range from a scalar floor where all components of a variance vector share a floor value, to a floor vector where each component has its own floor value. The second dimension is *time* (represented by a state sequence in a left-right HMM) where a unique floor can be shared by variance vectors within all states in all models, ranging to each state having its own floor. The third dimension is *feature space*, where different parts of the feature space may have their own floor. An example of the latter is when each Gaussian term within a composite pdf has its own floor value.

An alternative modification to the EM algorithm is to keep variances fixed while updating means and transition probabilities [2]. In the context of speaker verification where a non-client model is often used for likelihood normalization, the variances of the client model can be copied from the non-client model. A non-client model is often trained on a lot of data from many speakers and all parameters of the model can be reliably estimated with the original EM-algorithm. If non-client model variances are used systematically in client models, client variances become *client-independent*.

In this paper we compare several variations of the two principle modifications to the EM-algorithm mentioned above. The comparison is made on three separate telephone quality databases: Gandalf [5], SESP [6] and Polycost [7]. The recognition tasks are slightly different, but are all some form of text-dependent task using digits.

From the variety of possible variance flooring methods we try three variants with gradually increasing resolution: model-dependent, state-dependent and mixture component-dependent vector floors. The various floor vectors are computed as an empirical constant times a basis vector, like in [1]. The basis vector is derived from speech data or directly from a multi-speaker model.

Since the variance flooring technique involves the setting of an empirical constant, its usefulness depends on to which extent the choice of an optimal scaling factor will generalize from development data to new evaluation data. In this paper we describe a series of experiments to investigate on such generalization properties.

2. SYSTEM DESCRIPTION

A text-prompted ASV system based on word-level HMMs [3] is built on a generic platform for speaker verification systems called GIVES (General Identity Verification System). 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 those are appended with first and second order deltas. Cepstral mean subtraction is applied to the 13 static coefficients. In most experiments MFCC ceptral vectors are used. They are computed from a 24-channel, FFT-based, mel-warped, log-amplitude filterbank between 300-3400 Hz followed by a cosine transform. The energy term is the 0'th cepstral coefficient. In the end of section 4 the MFCCs are also replaced with LPCCs, where parameters from a 16-pole linear prediction filter are computed with the autocorrelation method and are transformed to 12-element cepstrum. The energy term is then the raw log-energy within each frame of samples, normalized within each utterance to have constant maximum amplitude for every utterance. All cepstral vectors are liftered to equalize the component variances. 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 and a mixture of eight Gaussians per state. 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 dimensions 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. Depending on the variance estimation method to be used, client model training proceeds in one of the following ways:

a) *Client-independent variances:* the client model means and mixture weights are re-estimated from enrollment data while variances are kept constant.

b) *Variance flooring:* client model means, mixture weights and variances are re-estimated from enrollment data. Variances are floored with one of three alternative methods.

Transition probabilities are kept constant in both cases. The three flooring methods are implemented as follows. Basis vectors with the same resolution as the flooring vectors are derived either directly from speech data or from a non-client model. Each floor vector is then set proportional to the corresponding basis vector. The scaling constant is unique for the entire system. The basis vectors are derived in one of the three following ways:

- *model-dependent floor:* the basis-vector for word model w is the variance of all vectors within segments identified as this word
- *state-dependent floor:* the basis-vector for a state s in word model w is computed as a linear combination of variance parameters of the mixture Gaussian in state s in a non-client model
- *mixture component-dependent floor:* the basis-vector for a mixture component i in state s of model w is the variance of mixture-component i in the non-client model.

The HMMs are implemented with HTK [4] with minor modifications to allow for training models on sparse data. The parameters of HMMs in multi-speaker models are estimated with EM-algorithm, with a crude fixed floor of 0.01 for all variance parameters. Initial parameters for a single-Gaussian model is first computed from Viterbi alignment of training data and are further trained with Baum-Welch re-estimation. The Gaussian terms are then split in two and the resulting mixture-Gaussian is again re-estimated. This procedure is repeated until there are eight Gaussians per state. This procedure is done independently for each HMM in the client model. The system depends on explicit segmentation of the input speech into words during both enrollment and test, the segmentation being produced by a speech recognizer (see Table 2).

3. DATABASE AND PROTOCOL

Three database [5,6,7] have been used in the tests. One of them, Gandalf, has been divided into two separate parts which are used as if they were two different databases in this paper. Table 2 summarizes the main features of the databases. They are all digital telephony databases recorded through ISDN. 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 ½ second).

Some additional facts not included in the table: Polycost test was baseline experiment 2 as defined in [6]; enrollment set on SESP is referred to as G in previous literature [1]; segmentation into words is made with a speech recognizer operating in forced alignment mode given the prompted text.

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 each figure, the left-most data points, labelled 'fixv', indicate client-independent variances. The remaining data points show error rate as a function of the scaling factor of some variance flooring method. This way, performance of client-independent variances and variance flooring can be compared within each figure.

Figure 1 compares the three different variance flooring methods on all three databases. It can be seen that the higher resolution in flooring, the less critical is the choice of scaling factor, since the minima in those curves are much wider and the position of the minima are closer to each other than for low resolution flooring. To investigate in detail on the lowest achieved error rates, Table 1 shows the average improvement when going from client-independent variances to each of the three variance flooring methods. Two cases are shown: First, the scaling factor has been chosen as the *a posteriori* best one for each of the databases. Second, a global scaling factor for all databases was chosen as the *a posteriori* best one for Gandalf development set. This corresponds to using that database as development data and testing the resulting system on the other three databases. There is a clear trend that higher resolution in variance flooring is better than lower, and only for the mixture component-dependent floors is the average error-rate lower than with client-independent variances.

Since the variance flooring method is applied to avoid under-training of variances on sparse training data, it can be expected that for a given recognition task and database, the need for flooring would systematically decrease with increased size of the enrollment set. The more training data the more should variances need to be floored. Hence, one can expect the optimal scale factor in variance flooring to decrease with larger enrollment sets. Such a trend is clearly visible in Figure 2 where we compare enrollment sizes from 0.3 to 1 minutes (3 to 12 training examples per digit).

One could further expect that with larger enrollment sets, variance flooring should be better relative to client-independent

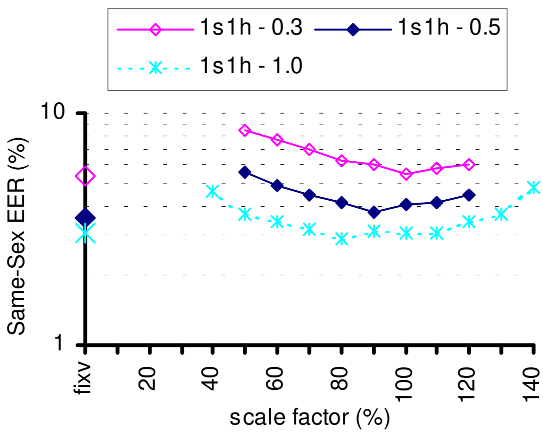

Figure 2. Comparison between different enrollment set sizes and 1-session, 1-handset (1s1h) enrollment on Gandalf (development set). Variance floor is state-dependent and the fixv-points shows results for client-independent variances.

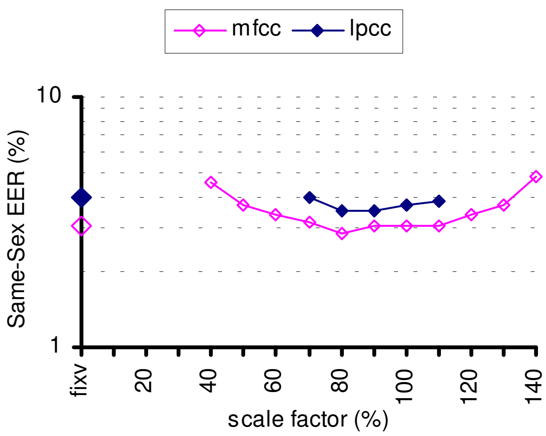

Figure 3. Comparison between MFCC-based and LPCC-based features. Each curve contains results with fixed, client-independent variances (fixv) and with variance flooring with *state-dependent* floors. Experiment is done on Gandalf (development set).

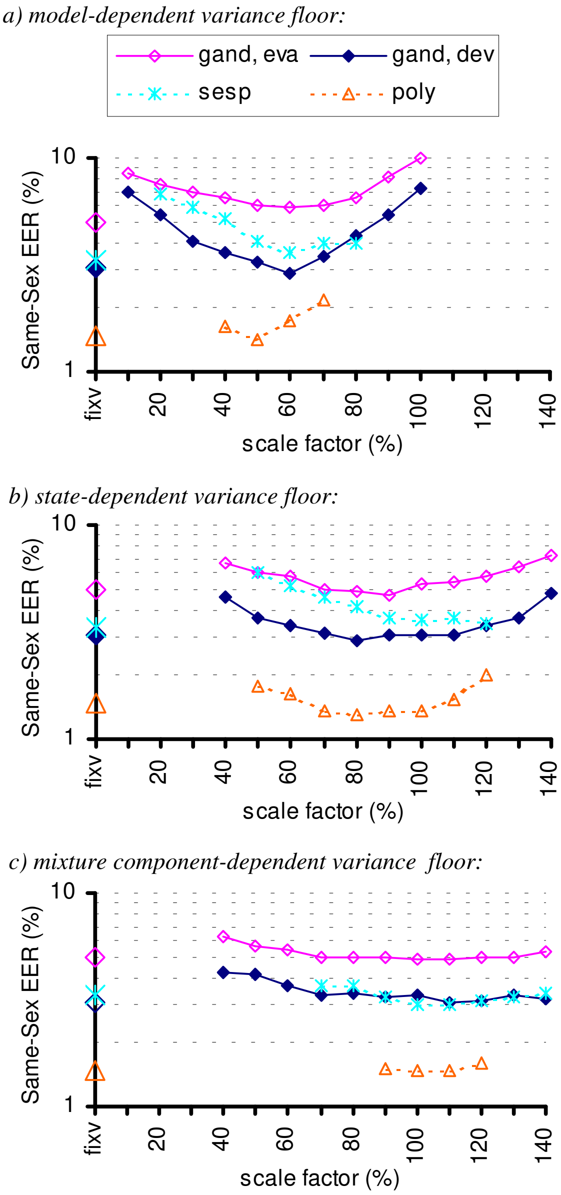

Figure 1. Comparison between same-sex EERs with fixed, client-independent variances (fixv) and with the three alternative flooring methods for the three databases: a) model-dependent floor, b) state-dependent floor and c) mixture component-dependent floor.

variances than with smaller enrollment sets. There is no clear evidence for this in the figure.

Finally we compared the MFCC-based features we used so far with LPCC-based features. Figure 3 shows error-rates for state-dependent floors on the Gandalf development set. On other databases too we observe that the optimal scaling factor is different for the two parameterizations and it seems that a scaling factor optimized for one parameterization may not be reusable for another.

Flooring level	individual scale factor	global scale factor
Model	-8%	-11% (0.60)
State	4%	-3% (0.80)
Mix-comp	3%	3% (1.10)

Table 1. Average improvement (negative values indicate a deterioration) for variance flooring over client-independent variances. The average is taken over all databases. The global scale factor (shown in parentheses) was chosen from the Gandalf development set.

5. CONCLUSIONS

We have compared two modifications of the EM-algorithm for HMM training on sparse data in the application of text-dependent speaker verification. The first is to copy variances from a non-client multi-speaker model and then keep them fixed while the EM-algorithm is applied to means and mixture weights. In the second method, variances are trained but they are 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, i.e. when the floor for the variance vector of a given Gaussian is proportional to the corresponding variance vector in the non-client model, was the best performing one. The optimal scaling factor for this kind of variance flooring was found to be around 1.10, which means that all variances are actually larger than with the client-independent variances.

Compared to the best performing variance flooring method, speaker-independent variances seem to work comparably without the need to estimate an empirical scaling factor for a variance floor. This trend is observed on three different databases, with two distinct parameterizations.

These results consolidate similar observations made in [2] and at recent NIST evaluations in text-independent ASV [8] that client models trained as adaptation of multi-speaker models with keeping covariance matrices constant brings a significant advantage, especially in the case of very scarce enrollment data.

The results open new tracks in the search for improved procedures and models in estimating client variances in the context of scarce enrollment data.

6. REFERENCES

1. Bimbot, Hutter, Jaboulet, Koolwaaij, Lindberg, Pierrot, "An Overview of the CAVE Project Research Activities in Speaker Verification", Proc. RLA2C, Avignon, France, April 20-23, pp 215-220, 1998.
2. Newman, Gillick, Ito, McAllaster, Peskin. "Speaker Verification Through Large Vocabulary Continuous Speech Recognition", Proc. ICSLP, Philadelphia, USA, 1996.
3. Rosenberg, Lee, Gokcen, "Connected Word Talker Verification Using Whole Word Hidden Markov Models", Proc. ICASSP, Toronto, Canada, pp 381-384, 1991.
4. Young, Odell, Ollason, Valtchev, Woodland, "The HTK Book (for HTK version 2.1)," Entropic Cambridge Research Laboratory, 1997.
5. Melin, "Gandalf - A Swedish Telephone Speaker Verification Database," Proc. ICSLP, Philadelphia, USA, pp. 1954-1957, 1996.
6. Petrovska D., Hennebert J., Melin H., Genoud D., "POLYCOST: A Telephone-Speech Database for Speaker Recognition", Proc. RLA2C, Avignon, France, April 20-23, 1998.
7. Boves L., Bogaart T., 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, April 5-7, pp 43-46, 1994.
8. NIST, Speaker Recognition Workshop. Informal proceedings, College Park, Maryland, USA, 31 March - 1 April, 1998.

	Test database Set	**Gandalf** dev	eval	**Polycost**	**SESP**
Task	language	Swedish		English	Dutch
	native speakers	100 %		~15%	100 %
	enrollment	1s1h-1.0*		2s1h-0.6	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	6012	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
System	speech recognizer for segmentation	HTK		Nuance	Phicos

Table 2. Summary of main features of the three databases and their protocols. Number of speakers are given as #male/#female. *In Figure 2 the enrollment set is varied between 0.3-1 minute length. +The number of handsets is an estimate.