ADVANCES IN CHANNEL COMPENSATION FOR SVM SPEAKER RECOGNITION*

Alex Solomonoff, W. M. Campbell, Ian Boardman

MIT Lincoln Laboratory, Lexington, MA
E-mail: {als,wcampbell,ian}@ll.mit.edu

ABSTRACT

Cross-channel degradation is one of the significant challenges facing speaker recognition systems. We study the problem for speaker recognition using support vector machines (SVMs). We perform channel compensation in SVM modeling by removing non-speaker nuisance dimensions in the SVM expansion space via projections. Training to remove these dimensions is accomplished via an eigenvalue problem. The eigenvalue problem attempts to reduce multisession variation for the same speaker, reduce different channel effects, and increase "distance" between different speakers. We apply our methods to a subset of the Switchboard 2 corpus. Experiments show dramatic improvement in performance for the cross-channel case.

1. INTRODUCTION

Cross-channel effects occur when a speaker has been enrolled on one type of channel and recognition occurs on a different channel. Many methods have been proposed to mitigate the problem—new features, transformation methods for standard cepstral features, score normalization, model transformation, etc. Feature-based compensation methods such as cepstral mean subtraction (CMS), stochastic matching [1], variance normalization, and feature mapping [2] have the advantage that they can be applied to any speaker modeling technique. Score-based normalization such as Tnorm [3] can be applied to any system that produces scores resembling a log-likelihood ratio. Model-based methods are powerful also, but have typically focused on GMM methods, e.g. SMS [4]. Our goal in this paper is to explore model-based methods for support vector machines (SVMs).

A speaker ID system that has shown good results is based upon sequence kernels and SVMs [5]. It uses a generalized linear discriminant (typically polynomials) as a kernel. We discuss this more in Section 2.

*This work was sponsored by the Department of Defense under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

In [6], a method of channel compensation was presented, which we call Nuisance Attribute Projection (NAP). The basic idea of NAP is to remove dimensions from the SVM expansion that are irrelevant to the speaker recognition problem. These initial results were promising [6]. Here we expand upon this work showing additional optimization criteria and extensive experiments.

We note that the closest technique to our channel compensation method is [7]. This work focuses on factor analysis of a general supervector of parameters. It is quite distinct from our techniques in that it is targeted toward GMM speaker recognition, does not use a SVM expansion space or SVM methodology, does not use a supervised optimization criterion, and is based upon a MAP criterion.

2. SUPPORT VECTOR MACHINES

At the most basic level, SVMs are two-class hyperplane-based classifiers operating in a (usually) high-dimensional space related nonlinearly to the original (usually lower-dimensional) input space. Given an observation $x \in X$ and a kernel function K, an SVM, $f(x)$ is given by

$$\begin{aligned} f(x) &= \sum_{i=1}^{N} \lambda_i \xi_i K(x, x_i) + b \\ &= \sum_{i=1}^{N} \lambda_i \xi_i \phi(x) \cdot \phi(x_i) + b \end{aligned} \tag{1}$$

We have assumed the Mercer condition [8]: $K(x, y)$ is an inner product expressible as $\phi(x) \cdot \phi(y)$ where $\phi : x \mapsto y \in Y$ for some expansion space Y. We compare the output of the SVM in (1) to a threshold in order to produce a decision. The x_i, ξ_i, and $\lambda_i > 0$ are obtained through a training process. The x_i are called support vectors and the ξ_i are the target class values: $+1$ for in-class and -1 for out-of-class.

3. NAP CHANNEL COMPENSATION

In [6] we introduced Nuisance Attribute Projection. Using NAP requires a corpus labeled with channel and/or speaker information. We created a projection matrix $P = I - XX^t$ which projects points in the n-dimensional expansion space

Y onto a subspace that is hopefully more resistant to channel effects. Then we project all background, test and target points onto this subspace.

The $n \times k$ matrix X (which could have a single column, or many columns) has orthonormal columns. In our work, k is at most a few hundred, compared to the dimension of the expansion space, which was typically $\approx 10,000$.

We chose X to minimize the average value of cross-channel distances:

$$\delta = \sum_{ij} W_{ij} \|P(\phi(x_i) - \phi(x_j))\|^2.$$

where the elements of the symmetric matrix W are positive for pairs of training points we want to pull together, negative for pairs we want to push apart, and zero for pairs we don't care about. We have tried several different versions of W.

The objective function δ is minimized by the k eigenvectors with largest eigenvalues of the eigenvalue problem

$$AZ(W)A^tX = X\Lambda. \tag{2}$$

where the matrix $Z(W) = \text{diag}(W\mathbf{1}) - W$.

Some definitions:

The corank of a matrix is the rank of its nullspace.

$\mathbf{1}$ is the column vector of all ones.

$\mathbf{1}_e$ is the vector with a one for each electret point in the training corpus and a zero for each carbon-button point.

$\mathbf{1}_c$ is the vector with a one for each carbon-button point in the training corpus and a zero for each electret point.

$\text{diag}(x)$ is the square diagonal matrix whose diagonal elements are the elements of the vector x.

$A = [\phi(x_0), \phi(x_1), \ldots, \phi(x_n)]$ is a matrix whose columns are the points in expansion space representing the training corpus.

Kernel Space: This eigenvalue problem occurs in expansion space. It has an alternative version in the space spanned by the training points, which we call kernel space. In this version.

$$X = AV,$$

where V is the matrix containing the k eigenvectors with largest eigenvalues of the symmetric generalized eigenvalue problem

$$KZ(W)KV = KV\Lambda. \tag{3}$$

with $K = A^tA$. This is the version of the equation we used for all of our experiments.

Weight Matrix: In the simple channel compensation case we used the weight matrix

$$W_{ij} = (W_{\text{channel}})_{ij} = \begin{cases} 1 & \text{channel}(x_i) \neq \text{channel}(x_j) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

This is the simplest type of weight matrix for minimizing cross-channel distances, but does nothing to increase cross-speaker distances, which might also result in an increase in performance. A weight matrix for this is

$$W = \alpha W_{\text{channel}} - \gamma W_{\text{speaker}}$$

where

$$\left(W_{\text{speaker}}\right)_{ij} = \begin{cases} 1 & \text{if } x_i \text{ and } x_j \text{ have different speakers} \\ 0 & \text{otherwise} \end{cases}$$

3.1. Difference of Channel Means

An alternative, much simpler version of channel compensation is to use the difference between the channel means. This is a corank-1 projection: $P = I - xx^t$, $\|x\| = 1$, with

$$x_0 = \frac{1}{n_c} A\mathbf{1}_c - \frac{1}{n_e} A\mathbf{1}_e.$$

The vector x_0 is not properly normalized, so $x = x_0/\|x_0\|$. Performance is quite similar to corank 1 results as discussed in Section 4.

As stated, it can only do corank-1 projection. If the training data has $m > 2$ different channel types, then a more elaborate process is possible. One could form the $m-1$-dimensional subspace spanned by all the pairwise differences between channel means and project those directions away.

3.2. The 3-Matrix Version of Channel Compensation

As mentioned in [6], a problem with the use of W_{channel} is that it tries to move together utterances that have different channels (good) but come from speakers that might sound very different even without channel differences. (bad)

Speaker Shrinking. Certainly we want to move a given speaker's different-channel utterance pairs together, but what about same-channel pairs? They include systematic differences which are also unrelated to speaker identity, such as the speaker's mood or health or differences between two different carbon-button handsets.

We really should be moving every same-speaker utterance pair together:

$$(W_{\text{ss}})_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ have the same speaker} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

We still want to move different-channel utterances together and different-speaker utterances apart, which gives us the following 3-term weight matrix:

$$W = \alpha W_{\text{channel}} + \beta W_{\text{ss}} - \gamma W_{\text{speaker}}, \tag{6}$$

with some positive coefficients α, β, and γ.

4. EXPERIMENTAL RESULTS

We performed several experiments based upon three corpora from Switchboard (SWB):

- (eval03) The 2003 NIST extended data task evaluation (using the "v1" lists), with single utterance enrollment. See [9] for a detailed description of the corpus. A background for the SVM was created from the unused splits. For EER results, the 95% confidence interval is about $\pm 0.6\%$. This corpus emphasized calls with the same telephone number from enrollment.
- (dev1) SWB 2 parts 2 and 4 and (dev2) SWB 2 parts 3 and 5. We used 3 session enrollment. Verification emphasized calls with different telephone numbers from enrollment. A background for the SVM was created from the SWB 2 phase 1 corpus and the unused portions of SWB 2 phases 4 and 5. For dev1, there were $3,790$ true trials and $54,514$ false trials. For dev2, there were $1,798$ true trials and $25,189$ false trials. For EER results, the 95% confidence intervals are $\pm 0.7\%$ for dev1/M, $\pm 0.7\%$ for dev1/F, $\pm 1.4\%$ for dev2/M, and $\pm 1.0\%$ for dev2/F.

The baseline system was a text-independent generalized linear discriminant kernel [5] using monomials of up to degree 3. Input features for eval03 were 18 LP cepstral coefficients (LPCC) and deltas (for consistency with prior work [6]), and for dev1/dev2 were 19 Mel filterbank cepstral coefficients (MFCC) and deltas. Standard channel-compensating measures were applied to the cepstral coefficients–cepstral mean subtraction and variance normalization. The dimension of the SVM expansion space was approximately $10,000$.

Since the extended data task is landline telephone, we used carbon button (CB) and electret (EL) as our two channels. We used a GMM channel classifier [2]; other choices of channel are possible.

4.1. Projection Rank

Significant error-rate reduction is obtained with increasing corank (see Section 3.2). Figure 1 presents results for male and female speakers for eval03 and dev1/dev2. The baseline system corresponds to the corank $= 0$ points. On eval03, using corank=128, EER was reduced by 20% below baseline for male speakers and by 30% for female speakers. In some tests, the effect of increasing corank is non-monotone, so for example, statistically significant improvement over baseline EER for male speakers in dev1 was achieved only for corank $>\approx 32$.

4.2. Tuning Weight Matrix Scalars

The coefficients α, β, and γ on the three weight matrices in (6) are tunable system parameters. We surveyed this parameter space seeking a local minimum EER. The results shown in Table 1 suggest a parameter space favoring equal or greater values of α and/or β over γ.

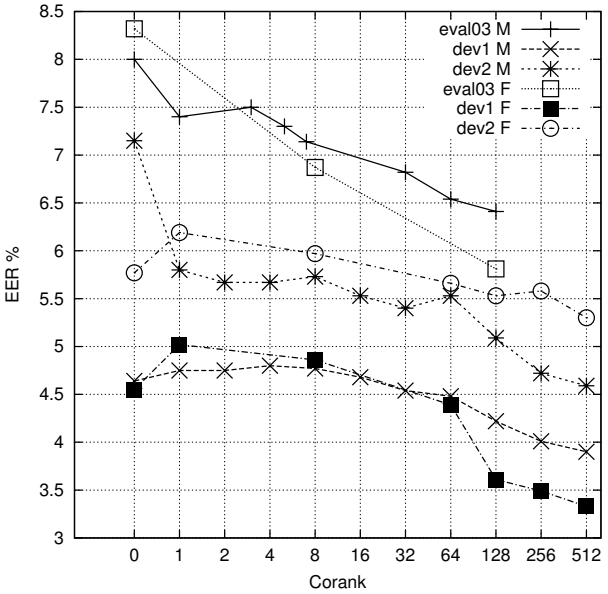

Fig. 1. EER vs. projection corank for male and female test sets, $(\alpha, \beta, \gamma) = (1, 8, 1)$.

Table 1. Parameter survey: EER dependence on weight matrix scale factors. Projection corank = 128. Male speakers only.

Parameter			Equal-error Rate (%)		
α	β	γ	dev-1	dev-2	comb.
0	1	0	3.61	4.72	3.98
1	1	0	3.69	4.72	4.02
10	1	1	3.73	4.59	4.02
10	10	1	3.74	4.59	4.02
10	1	0	3.71	4.70	4.04
1	0	0	3.69	4.66	4.06
1	10	0	3.74	4.72	4.06
1	1	1	4.15	5.05	4.38
10	1	10	4.22	4.94	4.42
1	0	1	4.22	5.03	4.44
1	10	1	4.27	5.13	4.51
0	10	1	(5.27)	6.69	(5.70)
1	1	10	(5.33)	6.75	(5.76)
1	10	10	(5.38)	6.75	(5.76)
0	0	1	(5.35)	6.81	(5.80)
0	1	10	(5.32)	6.82	(5.80)
0	1	1	(5.38)	6.79	(5.83)
baseline			4.61	7.58	5.39

() indicates greater than baseline EER

The entries in Table 1 fall into three statistically indistinguishable groups. Comparable EERs are achieved with either channel compensation alone or same-speaker compensation alone. This may result from a rich background corpus that includes many utterances per speaker, as well as multiple channels. The DET curves for two of the top scoring systems are presented in Figure 2.

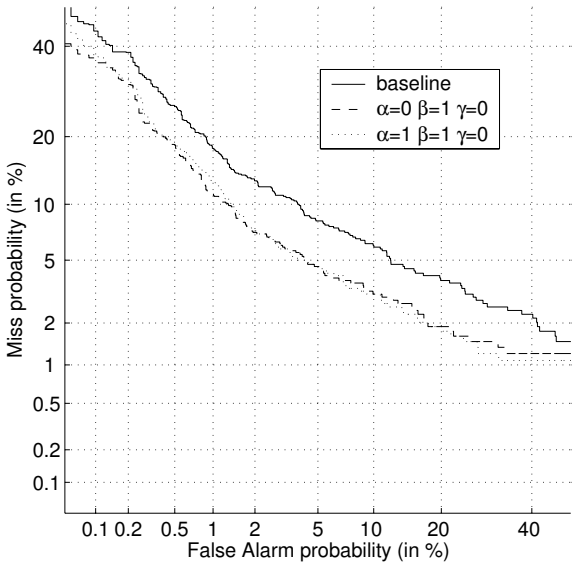

Fig. 2. Two DET curves for male speakers at projection corank = 128 using different weight matrix parameters on dev2.

Table 2. EER (%) for same and different channel conditions in training and testing. Projection corank = 128. CB: carbon button, EL: electret.

	channel		test set	
	train	test	dev. 1	dev. 2
baseline	CB and EL		4.64	7.15
projected			3.69	4.66
baseline	CB	CB	4.38	3.7*
projected			4.44	1.7*
baseline	EL	EL	3.40	4.90
projected			2.83	4.12
baseline	CB	EL	4.63	0.2*
projected			3.97	3.0*
baseline	EL	CB	4.77	7.58
projected			5.15	5.15

* statistically insignificant number of true trials for CB enrollments

4.3. Channel Effects

Table 2 gives a break-out of the scores for same vs. cross channel conditions on train and test, with channel compensation only, $(\alpha, \beta, \gamma) = (1, 0, 0)$. This analysis indicates that overall error rate on these tests is reduced by channel compensation but that this improvement is not attributable exclusively to cross-condition trials.

5. DISCUSSION AND CONCLUSIONS

The weight matrix terms W_{channel} and W_{ss} appear to be about equally valuable in reducing EER, and in fact don't seem to compliment each other. But, they work from different label types, which means a user can get good channel compensation when either label type is available. In addition, our channel compensation training corpora all had many utterances per speaker, typically about 10. If a training corpus had few or no repeated speakers, then W_{ss} probably would be of little value.

W_{channel} and W_{ss} have very different numbers of nonzero elements, which makes us suspect that appropriate magnitudes for α and β might be very different. Normalization of α, β and γ should be an interesting area of further work.

Tuning α, β and γ has been rather laborious and possibly prone to overtuning – the difference in EER between two different values of (α, β, γ) is often much less than our confidence intervals, in spite of using quite large test sets. A better understanding of appropriate values of (α, β, γ) would be valuable.

Because of the labor of searching the parameter space, some of the results we have shown do not use the best possible parameter values. For example, in Figure 1, $(\alpha, \beta, \gamma) = (1, 8, 1)$ is used, which is in the "mediocre" group of parameter values.

In conclusion, this group of techniques gives dramatic gains in SVM speaker ID performance.

Acknowledgments: We would like to acknowledge Carl B. Quillen for many useful discussions about this work.

6. REFERENCES

[1] Qi Li, S. Parthasarathy, and Aaron E. Rosenberg, "A fast algorithm for stochastic matching with applications to robust speaker verification," in *Proc. ICASSP*, 1997, pp. 1543–1546.

[2] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, 2003, pp. II–53–6.

[3] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[4] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *International Conference on Spoken Language Processing*, 2000.

[5] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. ICASSP*, 2002, pp. 161–164.

[6] A. Solomonoff, W. Campbell, and C. Quillen, "Channel compensation for SVM speaker recognition," in *Proc. Odyssey04*, 2004, pp. 57–62.

[7] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Proc. Odyssey04*, 2004, pp. 219–226.

[8] Nello Cristianini and John Shawe-Taylor, *Support Vector Machines*, Cambridge University Press, Cambridge, 2000.

[9] M. Przybocki and A. Martin, "The NIST year 2003 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2003/index.htm, 2003.