APPLICATION OF SVM-BASED CORRECTNESS PREDICTIONS TO UNSUPERVISED DISCRIMINATIVE SPEAKER ADAPTATION

*Matthew Gibson & Thomas Hain**

Department of Computer Science, Sheffield University, Regent Court, 211 Portobello, Sheffield S1 4DP

ABSTRACT

The effectiveness of unsupervised speaker adaptation is typically limited by errors in the estimated transcription of the adaptation data. Previous work has mitigated this negative effect by using only those sections of the adaptation data which are transcribed with relatively high confidence. In this work, phoneme correctness predictions are integrated into a discriminative unsupervised speaker adaptation procedure. Significant accuracy improvements (over the equivalent likelihood-based technique) are observed when using discriminative unsupervised speaker adaptation in combination with support vector machines to predict phoneme correctness.

***Index Terms*—** Discriminative speaker adaptation, confidence measures, SVM, minimum phone error.

1. INTRODUCTION

Speaker adaptation techniques based upon optimisation of model likelihood, e.g. maximum likelihood linear regression (MLLR, [1]) have been successful in supervised and unsupervised scenarios. Discriminative speaker adaptation methods differ from likelihood-based techniques in that they alter the acoustic model to optimise a discriminative measure, e.g. the minimum phone error criterion (MPE, [2]). This paper extends previous work on unsupervised linear regression speaker adaptation using the MPE criterion [3]. More specifically, this work demonstrates how estimation of the correctness of each phoneme in the estimated transcription can be used to mitigate the negative impact of incorrectly transcribed phonemes in the cases of both MLLR and MPE-based linear regression (MPELR). Further, it is shown that, with this mitigation technique in place, significant accuracy improvements over MLLR are obtained using MPELR.

The paper is structured as follows. Section 2 introduces the theory and implementation of MPELR. Section 3 explains how correctness estimates may be integrated into the MPE criterion to yield a correctness-adjusted MPE criterion suitable for unsupervised speaker adaptation. The correctness estimation methods used in this work are introduced in Section 4. Section 5 describes the large vocabulary recognition system used to evaluate the proposed technique. An evaluation of the correctness prediction and correctness-adjusted MPELR methods is presented in Section 6. A concluding summary and proposals for future research are found in Section 7.

2. MPE-BASED LINEAR REGRESSION

The linear regression speaker adaptation framework [1] uses adaptation data from a speaker to estimate one or more affine transforms of the SI acoustic model parameters. MPE-based linear regression adaptation deploys the same adaptation framework as MLLR, but the affine transforms are chosen to optimise the MPE criterion $R_{\mathrm{MPE}}(\theta)$ (Equation 1) instead of the model likelihood function.

$$R_{\mathrm{MPE}}(\theta) = \sum_{r=1}^{R} \sum_{w_1^N \in \mathcal{W}} p(w_1^N|\mathbf{o}_r, \theta) L(w_1^N, \hat{w}_1^{M(r)}) \qquad (1)$$

The set $\mathcal{W}$ comprises all possible phoneme transcriptions of the acoustic data $\mathbf{o}_r$, $\hat{w}_1^{M(r)}$ is the correct phoneme transcription of $\mathbf{o}_r$ and $L(w_1^N, \hat{w}_1^{M(r)})$ is the Levenshtein distance between the correct transcription and hypothesis w_1^N. The symbol θ represents the model parameters and R is the number of training set examples. Re-estimation equations for the transform parameters are derived in [3] and reused in this work.

The statistics necessary for MPELR transform estimation are calculated via a lattice-based implementation described in [2]. This work follows that implementation, with the exception that the symmetrically normalised frame error (SNFE, [4]) approximation is used since it has been shown to yield improved estimation of errors when compared to the error approximation method introduced in [2].

To address the issue of overfitting the adaptation data, a version of the I-smoothing regularisation technique, previously introduced in the context of MPE acoustic model estimation has also been applied to MPE-based acoustic model adaptation in [5], a technique adopted in this work. Additionally, in this work the MPELR transform estimation procedure adopts the same complexity control mechanism as MLLR to control the amount and type of MPELR transforms estimated. This ensures a fair comparison between MLLR and MPELR since the number of transforms are identical in each case.

3. CORRECTNESS-ADJUSTED MPELR

The idea behind correctness-adjusted MPELR is to disregard errors assigned with respect to labels of the reference transcription which are predicted to be incorrect. The correctness-adjusted MPE criterion is defined by Equation 2.

$$R^C_{\mathrm{MPE}}(\theta) = \sum_{r=1}^{R} \sum_{w_1^N \in \mathcal{W}} p(w_1^N|\mathbf{o}_r, \theta) L^C(w_1^N, \hat{w}_1^{M(r)}) \qquad (2)$$

Here $L^C(w_1^N, \hat{w}_1^{M(r)})$ is a correctness-adjusted error function which incorporates knowledge of the predicted correctness associated with the estimated reference transcription labels $\hat{w}_1^{M(r)}$. Since only a refinement of the error function is involved, correctness-adjusted MPELR transforms are estimated using the same theory as standard MPELR. Figure 1 illustrates the correctness-adjusted modification to the Levenshtein error approximation. Section A of Figure 1 shows an alignment of an estimated reference transcription and a hypothesis alignment. The overall SNFE of the hypothesis is 2, the sum

*This research was funded by CISCO systems.

of the SNFE for each aligned hypothesis label. Section B of Figure 1 shows the predicted correctness associated with each label of the estimated reference transcription, where *1* indicates a prediction of 'correct' and *0* denotes a prediction of 'incorrect'.

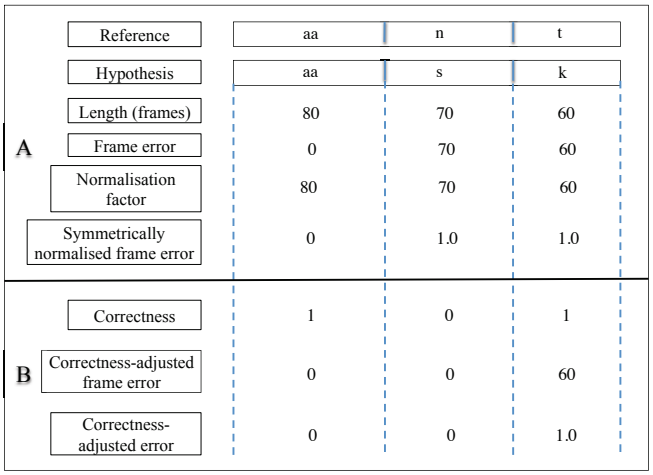

Fig. 1. *(A) Standard and (B) correctness-adjusted error approximations.*

The correctness-adjusted frame error is a modified version of the frame error which assigns errors only with respect to labels of the reference alignment which are predicted to be correct. For each segment of the hypothesis alignment, the correctness-adjusted frame error is zero if the segment overlaps with a label of the reference alignment which has an 'incorrect' prediction and equal to the standard frame error otherwise. The correctness-adjusted error for each hypothesis segment is then the normalised correctness-adjusted frame error, where the normalisation factor is the length of the shorter of the overlapping labels. The overall correctness-adjusted error for the hypothesis is then the sum of the correctness-adjusted error over each segment. Modifying the error in this way reduces the impact of errors associated with labels of the estimated reference transcription which are predicted as incorrect.

The question arises of how to formulate the I-smoothing prior distribution over the transform $\mathbf{W}$ in the case of correctness-adjusted MPELR. In this work, a correctness-adjusted prior $p^C(\mathbf{W})$, of the form described by Equation 3, is used. This prior is similar to the standard MPELR prior with the exception that correctness-adjusted occupancies replace standard occupancies.

$$\log p^C(\mathbf{W}) = \frac{\tau^I}{2} \sum_{m \in \mathcal{R}} \sum_{r=1}^{R} \sum_{t=1}^{T(r)} \gamma_m^C(t,r)(\mathbf{o}_t^r - \mathbf{W}\boldsymbol{\xi}_m)^{\mathsf{T}} \mathbf{C}_m^{-1}(\mathbf{o}_t(r) - \mathbf{W}\boldsymbol{\xi}_m) + k \quad (3)$$

Correctness-adjusted occupancies are given by Equation 4, where $C(t)$ is the correctness prediction associated with the label overlapping frame t in the most likely alignment of the reference transcription $\hat{w}_1^{M(r)}$.

$$\gamma_m^C(t,r) = \left\{ \begin{array}{cc} 0 & \text{if } C(t) \text{ incorrect} \\ \gamma_m(t,r) & \text{otherwise} \end{array} \right\} \quad (4)$$

The quantity k is a normalisation term used to ensure the prior probability distribution sums to one and $\mathbf{C}_m$ is the covariance of component m. Note that the prior is formulated for each regression class $\mathcal{R}$.The t-th frame of the r-th training utterance is denoted by $o_t(r)$ and $\gamma_m(t,r)$ is the posterior probability that state m is the t-th element of the hidden state sequence. The I-smoothed correctness-adjusted MPE criterion is formulated by subtracting the prior defined above (Equation 3) from the correctness-adjusted MPE criterion (Equation 2). The parameter τ^I determines the influence of the prior term.

With regard to complexity control for correctness-adjusted MPELR, errors corresponding to much of the data may be disregarded. To introduce sensitivity to the disregarded data, the correctness-adjusted occupancies of Equation 4 replace the standard occupancies in the complexity control framework.

4. CORRECTNESS PREDICTION

Much research has been done to identify suitable features (often referred to as confidence measures) and classifiers for correctness prediction in speech recognition. Word and phoneme posterior probabilities are used for correctness prediction in this work. These posterior probabilities are derived from phoneme-marked word lattices as described in [6][1].

Using a single feature, e.g. phoneme posterior, it is possible to construct a simple phoneme correctness classifier by defining a threshold. Such a classifier will be used to provide a useful comparison in the experimental work of this paper. Since correctness is a binary classification task it is possible to deploy support vector machines (SVMs). In the experimental work reported here, SVMs are constructed for phoneme correctness classification using subsets of the following features: phoneme posterior, posterior of containing word, posterior of adjacent (previous and following) phonemes, posterior of adjacent words, phoneme acoustic likelihood, phoneme duration, containing word duration, number of phonemes in containing word, phoneme identity, adjacent phoneme identities, and identities of all phonemes in containing word. The phoneme identity is represented by a sparse vector whose dimensions correspond to phonemes in the phoneme inventory. The vector has value 1 for the dimension corresponding to the associated phoneme, and 0 elsewhere. The adjacent phoneme identities and identities of all phonemes in the containing word are represented by similar sparse vectors. All non-sparse dimensions of the feature vector are scaled and biased to have zero mean and unit variance on the training dataset. The same scales and biases are applied to test set features.

In [7], where simlar SVMs are constructed for word correctness prediction, the use of the non-linear radial basis function (RBF) kernel provides improvements in classification accuracy when compared with a linear kernel. Motivated by this observation, the RBF kernel (parametrised by the variable γ) will be used here. Due to the size of the datsets used in this work (of the order of a hundred thousand datapoints), an efficient algorithm called 'cutting-plane subspace pursuit' [8] is used to train the SVMs.

5. EXPERIMENTAL SYSTEM

A conversational speech transcription (CTS) system is used to evaluate the techniques described above. Training and development test datasets are sourced from the Fisher corpus. The evaluation dataset is the NIST RT03 English conversational speech dataset. Table 1

[1] In the terminology of the cited report, the *maximal frame posterior* measure is used.

summarises these datasets, detailing the number of speakers and their volume.

Dataset	Speakers	Volume (hours)
fisher_train	8259	1007.2
fisher_devtest	91	5.7
rt03_eval	144	6.2

Table 1. Training, development and evaluation datasets.

The system uses 13 Mel-frequency-based perceptual linear prediction coefficients (and their first and second-order time derivatives) to represent the speech signal. Speaker and channel-specific cepstral mean and variance normalisation is applied to these features. The acoustic models are state-clustered triphone HMMs with three emitting states and left-to-right topology, trained using the fisher_train dataset and the maximum likelihood (ML) criterion. A total of $14,368$ tied states are used and state output distibutions are 16-component Gaussian mixture models.

The recognition dictionary contains $48,850$ words. An interpolated trigram language model is used which is based upon component LMs derived from the HUB4 LM96 broadcast news text source and the fisher_train dataset.

5.1. System operation

The evaluation system uses three stages for recognition. The first stage uses the unadapted acoustic models to provide a transcription of the test data. This transcription is then used as input to the second stage, where the acoutic models are adapted using several passes of MLLR adaptation. A total of five passes of MLLR adaptation are used, where the adapted acoustic models are used both as a starting point in the subsequent adaptation pass, and to estimate the transcription used in the subsequent adaptation pass. Six iterations of expectation-maximisation are used in each MLLR adaptation pass. In the third stage, MPELR is used to further adapt the acoustic models. The fifth-pass MLLR-adapted models are used as a starting point, and to estimate the transcription used here.

5.2. Adaptation configuration

The adaptation procedure alters only the mean of each component of the acoustic model. A regression tree comprising 64 base classes, block diagonal transforms and a node occupancy threshold of 1600 are used.In the case of MPELR, model-marked lattices ('denominator' lattices) are generated using a bigram language model (derived from the trigram used in recognition) and the fifth-pass MLLR-adapted acoustic models. The lattices are subsequently pruned to a maximum density of 500 arcs per second, retaining only the lattice paths of highest posterior probability. 'Numerator' lattices are generated using the fifth-pass MLLR-adapted acoustic models and their estimated transcription. The numerator lattices are merged into the pruned denominator lattices to give the 'consolidated' lattices used in MPELR. An acoustic model scaling factor of $\frac{1}{52}$ and a language model scaling factor of $\frac{1}{4}$ (maintaining the ratio of $\frac{1}{13}$ used in recognition) are used in transform estimation to broaden the distribution of posterior probabilities. An I-smoothing factor τ^I of 0.2, a learning rate E (see [5]) of 1.0 and forty iterations of MPELR transform estimation are used.

When using correctness-adjusted MPELR, the consolidated lattices and the fifth-pass MLLR estimated transcription are used to calculate the word and phoneme posterior probabilities described in Section 4.

6. EVALUATION

6.1. Correctness estimation

To evaluate the correctness estimation procedures used in this work, each phoneme in the estimated transcription is firstly marked as correct or incorrect. This reference is then used to measure the error rate (correctness error rate, CER) of a particular correctness classifier.

SVM correctness classifiers are trained using a subset of the fisher_devtest dataset comprising $20,000$ phonemes and tested on a separate held-out subset comprising $141,876$ phonemes (referred to as fisher_devsub1). Table 2 displays the performance (on fisher_devsub1 and rt03_eval) of SVM classifiers which use different input feature spaces. As the rows of the table are descended, more features are added to the input feature space. So e.g. the features corresponding to the second row include the word duration in addition to the word and phone posteriors of the current, previous and next word and phoneme. The parameter γ is optimised on the set fisher_devsub1 using a grid search procedure.

Features	γ	CER(%)	
		fisher_devsub1	rt03_eval
Posteriors (current and adjacent words and current and adjacent phonemes)	0.1	21.7	21.7
+ duration of containing word	0.2	21.6	21.8
+ number of phonemes in containing word	0.2	21.2	21.3
+ identity of phonemes in containing word	0.1	20.8	21.3
+ phoneme duration	0.1	20.8	21.3
+ phoneme identity	0.05	20.6	21.3
+ adjacent phoneme identities	0.03	20.6	21.0
+ phoneme acoustic likelihood	0.03	20.5	21.1

Table 2. Performance of SVM classifier with varying imput features.

To investigate the impact of larger training datsets, SVM correctness classifiers are trained using several different subsets of the fisher_devtest dataset (comprising $20,000$, $60,000$, $100,000$ and $150,000$ phonemes) and tested on a separate held-out subset comprising $50,000$ phonemes (referred to as fisher_devsub2). These classifiers use all the features mentioned above and a γ factor of 0.03. Table 3 displays their performance on the datsets fisher_devsub2 and rt03_eval.

Table 4 compares the performance (on rt03_eval) of the classifier which uses a posterior threshold (0.3, optimised on fisher_devtest) with the SVM trained on $150,000$ phonemes detailed in Table 3. Classification of all phonemes as correct yields the 'dumb baseline' of Table 4.

6.1.1. Discussion

The results of Table 4 indicate the superiority of the SVM correctness classifier over the posterior-thresholded classifier. The analysis of Table 2 indicates that this is due, in part at least, to the use of additional features to the phoneme posterior. The use of an RBF kernel to

Volume of training dataset (phonemes)	CER(%)	
	fisher_ devsub2	rt03_ eval
20,000	20.7	21.1
60,000	20.6	21.0
100,000	20.4	20.7
150,000	20.4	20.8

Table 3. Performance of SVM classifier as training data varies.

Classifier	CER(%)
Dumb baseline	24.0
Posterior (threshold=0.3)	21.8
SVM ($\gamma = 0.03$)	20.8

Table 4. Performance of correctness classifiers (rt03_eval).

induce a non-linear decision boundary may also contribute to the improved correctness classification over the posterior-thresholded classifier (which uses a linear decision boundary).

6.2. Unsupervised adaptation

Correctness-adjusted MPELR is evaluated when used in the third stage of recognition. For comparison, the performance of correctness-adjusted MLLR technique is also measured. Correctness-adjusted MLLR is implemented by by replacing standard occupancies with the correctness-adjusted occupancies during transform estimation. The techniques are evaluated using the posterior threshold and SVM correctness predictors evaluated above. The ideal predictor (which outputs the reference correctness for each phoneme in the estimated transcription) and no predictor (i.e. standard MLLR and MPELR) scenarios are also included for comparative purposes.

The WER (on `rt03_eval`) of the unadapted system and the fifth-pass MLLR system are 36.6% and 31.8% respectively. Table 5 records the performance (on `rt03_eval`) of the adaptation methods described above when used in the third stage of recognition. Each pair of systems with the same correctness predictor is compared. Where a significant difference is found the performance of the better system is indicated in bold font. Significance is set at the 95% confidence level using the matched pairs sentence segment word error test [9].

Correctness predictor	WER (%)	
	MLLR	MPELR
none	31.8	31.6
posterior threshold	31.5	31.4
SVM	31.6	**31.2**
ideal	30.4	**29.3**

Table 5. Performance of third stage adaptation techniques.

6.2.1. Discussion

The results of Table 5 indicate that, without use of any correctness predictions, MPELR yields a small, but not significant, performance improvement over MLLR. Inspection of the remaining rows of Table 5 reveals that the incorporation of correctness predictions leads to improved performance in the case of both MLLR and MPELR. In the case of ideal correctness predictions, a relatively large (1.1% absolute WER) performance improvement is observed when comparing MPELR with MLLR. In this case, where the effect of errorful transcriptions has been nullified, the benefit of using a discriminative criterion over the likelihood criterion is observed.

In the realistic scenarios of imperfect correctness predictions, and comparing the systems which use the same predictor, MPELR yields accuracy improvements over MLLR in all cases. This improvement is significant in the case of the SVM correctness predictor. Further, use of the SVM predictor yields improved accuracy (over the posterior threshold predictor) in the case of MPELR but slightly degraded accuracy in the case of MLLR. This evidence indicates that use of improved correctness prediction is of greater benefit in the case of MPELR than in the case of MLLR. The small degradation observed in the case of MLLR merits further investigation.

7. CONCLUSIONS

This paper has demonstrated how correctness predictions may be incorporated into unsupervised discriminative speaker adaptation to deliver significantly improved accuracy over the equivalent likelihood-based procedure. The relationship between the performance of correctness-adjusted MPELR and the nature of the correctness classifier (in terms of performance, and false negative and positive tradeoffs) should be investigated in future work.

8. REFERENCES

[1] C. J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[2] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, Cambridge University, 2003.

[3] L. Wang and P.C. Woodland, "MPE-based discriminative linear transform for speaker adaptation," in *Procedings ICASSP*, Montreal, Canada, 2004, pp. 321–324.

[4] M. Gibson and T. Hain, "Error approximation and minimum phone error acoustic model estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 18, no. 2, pp. 1269–1279, 2010.

[5] L. Wang and P.C. Woodland, "MPE-based discriminative linear transforms for speaker adaptation," *Computer Speech and Language*, vol. 22, no. 3, pp. 256–272, 2008.

[6] M. Gibson and T. Hain, "Confidence-informed unsupervised minimum Bayes risk acoustic model adaptation," Tech. Rep. CS-11-01, Sheffield University Department of Computer Science, 2011.

[7] R. Zhang and A. Rudnicky, "Word level confidence annotation using combinations of features," in *Proceedings European Conference on Speech Communication and Technology*, 2001, pp. 2105–2108.

[8] T. Joachims and C.-N. Yu, "Sparse Kernel SVMs via Cutting-Plane Training," *European Conference on Machine Learning (ECML), Machine Learning Journal, Special ECML Issue*, vol. 76, no. 2-3, pp. 179–193, 2009.

[9] L. Gillick and S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings ICASSP*, Philadelphia, 1989, pp. 532–535.