Discriminative Classifiers with Generative Kernels for Noise Robust ASR

M.J.F. Gales and C. Longworth

Cambridge University Engineering Department
Trumpington St., Cambridge CB2 1PZ, U.K.
{mjfg,cl336}@eng.cam.ac.uk

Abstract

Discriminative classifiers are a popular approach to solving classification problems. However one of the problems with these approaches, in particular kernel based classifiers such as Support Vector Machines (SVMs), is that they are hard to adapt to mismatches between the training and test data. This paper describes a scheme for overcoming this problem for speech recognition in noise. Generative kernels, defined using generative models, allow SVMs to handle sequence data. By compensating the generative models for the noise conditions noise-specific generative kernels can be obtained. These can be used to train a noise-independent SVM on a range of noise conditions, which can then be used with a test-set noise kernel for classification. Initial experiments using an idealised version of model-based compensation were run on the AURORA 2.0 continuous digit task. The proposed scheme yielded large gains in performance over the compensated models.

Index Terms: speech recognition, noise robustness, support vector machines.

1. Introduction

Speech recognition is normally based on generative models, in the form of Hidden Markov Models (HMMs), and class priors, the language model. These are then combined using Bayes' decision rule. An alternative approach is to use discriminative models, or discriminative functions such as Support Vector Machines (SVMs) [1]. One of the problems with using these discriminative models and functions is that it is normally hard to adapt them to changing speakers or acoustic environments. This is particularly true of kernel based approaches, such as SVMs, where individual training examples are used to determine the decision boundaries. One approach to handling SVM-based adaptation is described in [2]. This involves using the support vectors from the original, unadapted, model in combination with the adaptation data.

An obvious application area where there are large mismatches between the training and test sets is speech recognition in noise. Handling changing acoustic conditions has been an active area of research for many years. Model-based compensation schemes [4, 3, 5] are a powerful approach to handling mismatches between training and test conditions. Well implemented model-based compensation schemes tend to outperform feature-based compensation schemes as it is possible to more accurately model situations where speech is, for example, masked by the noise. This paper examines an approach that allows discriminative classifiers to be combined with model-based compensation schemes to improve the noise-robustness.

In this work rather than attempting to modify the SVM itself the form of the kernel is altered to reflect the changing acoustic conditions. For the class of kernels that make use of generative models [6, 7], such as HMMs, this involves performing model-based compensation to adapt the generative model. This noise-specific generative kernel is then used by the SVM. Provided the form of kernel compensates for the effects of the environment changes, it should be possible to train (and classify with) a noise-independent SVM on a range of noise conditions with the appropriate noise dependent kernels. As this work is primarily interested in the feasibility of using noise-independent SVMs, an idealised version of model-based compensation is used, single-pass retraining (SPR) [4].

Rather than using discriminative classifiers, discriminative training of generative models can be implemented. These discriminatively trained generative models can be adapted using, for example, model-based compensation schemes. However, the models are still generative so could be used to define an improved generative kernel for use with noise-independent SVMs as above.

This paper is organised as follows. The next section briefly reviews model-based compensation schemes and the idealised form examined in this work. This is followed by a brief discussion of SVMs and the forms of dynamic kernel that can be used with generative models. Section 4 then describes the complete scheme for using noise-independent SVMs. Results on the AURORA 2.0 database are then given in section 5.

2. Model-Based Noise Compensation

The first stage in producing a noise compensation scheme is to define the impact of the acoustic environment and channel on the clean speech data, the *mismatch function*. In the mel-cepstral domain used in this work the following approximation between the static clean speech, noise and noise corrupted speech observations is used ($\log(.)$ and $\exp(.)$ indicate element-wise logarithm or exponential functions)

$$\boldsymbol{y}_t^{\mathsf{s}} = \boldsymbol{x}_t^{\mathsf{s}} + \boldsymbol{h} + \mathbf{C}\log\Big(1 + \exp(\mathbf{C}^{-1}(\boldsymbol{n}_t^{\mathsf{s}} - \boldsymbol{x}_t^{\mathsf{s}} - \boldsymbol{h})\Big) \quad (1)$$

where $\mathbf{C}$ is the DCT matrix. For a given set of noise conditions, the observed (static) speech vector $\boldsymbol{y}_t^{\mathsf{s}}$ is a highly non-linear function of the underlying clean (static) speech signal $\boldsymbol{x}_t^{\mathsf{s}}$, noise $\boldsymbol{n}_t^{\mathsf{s}}$ and convolutional noise $\boldsymbol{h}$. Noise compensation schemes are further complicated by the addition of dynamic parameters. The observation vector $\boldsymbol{y}$ is often formed of the static parameters appended by the delta and delta-delta parameters. Thus $\boldsymbol{y}_t^{\mathsf{T}} = \left[\begin{array}{ccc} \boldsymbol{y}_t^{\mathsf{sT}} & \Delta\boldsymbol{y}_t^{\mathsf{sT}} & \Delta^2\boldsymbol{y}_t^{\mathsf{sT}} \end{array}\right]$. Mismatch functions for all the parameters can be obtained [4, 8].

The aim of model-based compensation schemes is to obtain the parameters of the noise-corrupted speech model from

Chris Longworth was supported by a Schiff studentship.

the clean speech and noise models. Most model-based compensation methods assume that if the speech and noise models are Gaussian then the combined noisy model will also be Gaussian. Thus to compute the expected value of the observation for each clean speech component (assuming a single noise component) the following must be computed

$$\boldsymbol{\mu}_m = \mathcal{E}\left\{\boldsymbol{y}\right\}; \quad \boldsymbol{\Sigma}_m = \mathrm{diag}\left(\mathcal{E}\left\{\boldsymbol{y}\boldsymbol{y}^{\mathsf{T}}\right\} - \boldsymbol{\mu}_m\boldsymbol{\mu}_m^{\mathsf{T}}\right) \tag{2}$$

where the expectation is over the clean speech "observations" from component m and noise "observations" combined using equation 1. There is no simple closed-form solution to these equations so various approximations have been proposed. These include Parallel Model Combination [4] and Vector Taylor Series [3]. An additional problem that must be solved is that noise models are not normally available. Thus these must be estimated from the observed data. Schemes that allow all the model parameters to be estimated have been proposed [3, 9, 5].

In this work an idealised version of these model-based compensation schemes is used, Single-Pass retraining (SPR) [4]. Here it is assumed that stereo data is available of the form $\{\mathbf{Y}, \mathbf{X}\}$ where $\mathbf{Y} = \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T, \mathbf{X} = \boldsymbol{x}_1, \ldots, \boldsymbol{x}_T$. The expectations in equation 2 are then approximated as

$$\boldsymbol{\mu}_m \approx \frac{\sum_{t=1}^{T} \gamma_m^{\mathsf{x}}(t)\boldsymbol{y}_t}{\sum_{t=1}^{T} \gamma_m^{\mathsf{x}}(t)} \tag{3}$$

where $\gamma_m^{\mathsf{x}}(t)$ is the posterior probability of component m generating the observation at time t using the clean-speech model and clean-speech data, $\mathbf{X}$.

3. SVMs and Generative Kernels

Support Vector Machines (SVMs) [1] are an approximate implementation of structural risk minimisation. They have been found to yield good performance on a wide range of tasks. The theory behind SVMs has been extensively described in many papers and is not discussed here. This section concentrates on how SVMs can be applied to tasks where there is sequence data, for example speech recognition.

One of the issues with applying SVMs to sequence data, such as speech, is that the SVM is inherently static in nature; "observations" (or sequences) are all required to be of the same dimension. A range of *dynamic kernels* have been proposed that handle this problem. Of particular interest in this work are those kernels that are based on generative models, such as Fisher kernels [6] and generative kernels [7]. In these approaches a generative model is used to determine the feature-space for the kernel. An example first-order feature-space for a generative kernel with observation sequence $\mathbf{Y}$ may be written as

$$\boldsymbol{\phi}(\mathbf{Y};\boldsymbol{\lambda}) = \frac{1}{T}\left[\begin{array}{c} \log\left(p(\mathbf{Y};\boldsymbol{\lambda}^{(\omega_1)})\right) - \log\left(p(\mathbf{Y};\boldsymbol{\lambda}^{(\omega_2)})\right) \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}^{(\omega_1)}} \log p(\mathbf{Y};\boldsymbol{\lambda}^{(\omega_1)}) \\ \boldsymbol{\nabla}_{\boldsymbol{\lambda}^{(\omega_2)}} \log p(\mathbf{Y};\boldsymbol{\lambda}^{(\omega_2)}) \end{array}\right] \tag{4}$$

where $p(\mathbf{Y};\boldsymbol{\lambda}^{(\omega_1)})$ and $p(\boldsymbol{Y};\boldsymbol{\lambda}^{(\omega_2)})$ are the likelihood of the data using generative models associated with classes ω_1 and ω_2 respectively. HMMs are used as the generative model in this paper. Considering only the derivative with respect to the means, the feature-space will have the form

$$\frac{\partial}{\partial\boldsymbol{\mu}_m^{(\omega_1)}} \log p(\mathbf{Y};\boldsymbol{\lambda}^{(\omega_1)}) = \sum_{t=1}^{T} \gamma_m(t)\boldsymbol{\Sigma}_m^{(\omega_1)\text{-}1}\left(\boldsymbol{y}_t - \boldsymbol{\mu}_m^{(\omega_1)}\right) \tag{5}$$

where $\gamma_m(t)$ is the posterior probability that component m generated the observation at time t given the complete observation sequence $\mathbf{Y}$. Only the derivatives with respect to the means are used in this work, though it is possible to use other, and higher-order, derivatives.

As SVM training is a distance based learning scheme it is necessary to define an appropriate metric for the distance between two points. The simplest approach is to use a *Euclidean* metric. However, in the same fashion as using the *Mahalanobis*, rather than Euclidean, distances for nearest-neighbour training, an appropriately weighted distance measure may be better. One such metric which is maximally non-committal is given by

$$K(\mathbf{Y}_i, \mathbf{Y}_j; \boldsymbol{\lambda}) = \boldsymbol{\phi}(\mathbf{Y}_i;\boldsymbol{\lambda})^{\mathsf{T}}\mathbf{G}^{\text{-}1}\boldsymbol{\phi}(\mathbf{Y}_j;\boldsymbol{\lambda}) \tag{6}$$

where $\mathbf{Y}_i$ and $\mathbf{Y}_j$ are two observation sequences and $\mathbf{G}$ is related to the Fisher Information matrix (the log-likelihood ratio is also included). In common with other work in this area [7, 10], $\mathbf{G}$ is approximated by the diagonalised empirical covariance matrix of the training data.

Classification with this form of generative kernel with observation sequence $\mathbf{Y}$ and training data $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is then based on the SVM score $\mathcal{S}(\mathbf{Y})$

$$\mathcal{S}(\mathbf{Y}) = \sum_{i=1}^{n} \alpha_i^{\mathsf{svm}} z_i K(\mathbf{Y}_i, \mathbf{Y}; \boldsymbol{\lambda}) + b \tag{7}$$

$$\hat{\omega} = \begin{cases} \omega_1, & \mathcal{S}(\mathbf{Y}) \geq 0 \\ \omega_2, & \mathcal{S}(\mathbf{Y}) < 0 \end{cases} \tag{8}$$

where α_i^{svm} is the Lagrange multiplier for observation sequence $\mathbf{Y}_i$ obtained from the SVM maximum margin training, b is the bias and $z_i \in \{1, -1\}$ indicates whether the sequence was a positive (ω_1) or negative (ω_2) example.

4. SVMs for Noise Robustness

The previous two sections have described model-based compensation and support vector machines with generative kernels. This section describes how these schemes can be combined together to allow noise-specific generative kernels to be used with a noise-independent SVM for speech recognition.

One of the problems with using SVMs for speech recognition is that standard SVMs are binary classifiers whereas speech is a multi-class task; for large vocabulary systems there are a vast number of classes. One approach to handling this problem is acoustic code-breaking [11]. Here confusable pairs in the training data are obtained by finding the most confusable word with each of the reference words. This provides a set of training examples for a binary classifier. In this work a slightly modified version of acoustic code-breaking is used. During training rather than just selecting the data from the specific confusable pairs all the data from each of the words is used during training. This yields a far larger number of training examples.

The procedure for training the noise-independent SVMs is:

1. For each training noise condition perform model-based compensation
2. Align all the training utterances $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ using reference, $\mathbf{r} = r_1, \ldots, r_K$ to give the word-segmented data sequence $\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_K$
3. For each confusable pair (ω_l, ω_j) set $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{(\omega_l)}, \boldsymbol{\lambda}^{(\omega_j)}\}$
   (a) obtain $\boldsymbol{\phi}(\tilde{\mathbf{Y}}_i;\boldsymbol{\lambda})$ for all training examples of ω_l using the appropriate noise compensated $\boldsymbol{\lambda}$

(b) obtain $\boldsymbol{\phi}(\tilde{\mathbf{Y}}_i;\boldsymbol{\lambda})$ for all training examples of ω_j using the appropriate noise compensated $\boldsymbol{\lambda}$

(c) train a noise-independent SVM for pair (ω_l, ω_j) using all positive (b) and negative (c) examples.

In this work only the log-likelihood ratio and derivatives with respect to the means are used. There is an issue with using equation 5. Model-based compensation schemes normally modify the variances of the acoustic models. To keep the dynamic ranges of each set of features consistent standard-deviation normalisation, rather than the variance normalisation in equation 5, is used. Note this is not usually a problem as the same covariance matrices are used for all sequences and the dynamic-range effects handled by the metric $\mathbf{G}$.

During recognition the following procedure is used:

1. Compensate the acoustic models for the test noise condition
2. Recognise the test utterance $\mathbf{Y}$ to obtain 1-best hypothesis, $\mathbf{h} = h_1, \ldots, h_K$ and align to give the word-segmented data sequence $\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_K$
3. For each segment $\tilde{\mathbf{Y}}_i$ and for each confusable pair (ω_l, ω_j) set $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}^{(\omega_l)}, \boldsymbol{\lambda}^{(\omega_j)}\}$ and

   (a) If $(h_i = \omega_l)$ or $(h_i = \omega_j)$ then obtain $\mathcal{S}(\tilde{\mathbf{Y}}_i)$

$$\hat{\omega} = \begin{cases} h_i, & \text{if } \left|\log\left(\frac{p(\tilde{\mathbf{Y}}_i;\boldsymbol{\lambda}^{(\omega_l)})}{p(\tilde{\mathbf{Y}}_i;\boldsymbol{\lambda}^{(\omega_j)})}\right)\right| \geq \beta \\ \omega_l, & \text{if } \mathcal{S}(\tilde{\mathbf{Y}}_i) + \epsilon\log\left(\frac{p(\tilde{\mathbf{Y}}_i;\boldsymbol{\lambda}^{(\omega_l)})}{p(\tilde{\mathbf{Y}}_i,\boldsymbol{\lambda}^{(\omega_j)})}\right) \geq 0 \\ \omega_j; & \text{otherwise} \end{cases}$$

   set $h_i = \hat{\omega}$

β and ϵ are empirically set values. They control two different aspects of the recognition stage:

- β determines the number of pairs that are rescored. If the magnitude of the log-likelihood ratio is above the threshold β the HMM classification is considered to be sufficiently "confident" that rescoring is unnecessary. As $\beta \to 0$ the performance will become the same as the standard HMM classification.
- ϵ is used to scale the contribution of the log-likelihood ratio to the SVM score. The log-likelihood ratio is the most discriminatory of the dimensions of the score-space is the log-likelihood ratio. However using a maximally non-committal metric, $\mathbf{G}$, all dimensions are treated equally. Thus ϵ is used to reflect the usefulness of the log-likelihood ratio. As $\epsilon \to \infty$ the performance of the system will tend to the HMM performance.

For all the experiments in this paper they were roughly set using a single confusable-pair at a single SNR. They were then fixed for all pairs. Thus there is minimal bias as no extensive tuning on the test data was performed.

This routine is known to be suboptimal in a number of ways. A simple scheme is used to combine classifier outputs so the results will depend on the order that the confusable pairs are applied in. Thus the performance is expected to be a slight underestimate of the possible gains. This issue will be investigated in future work. The alignment associated with each word-segment is not updated if the hypothesis sequence changes. It is possible to repeat the alignment if the hypothesis changes. The computational load associated with this scheme increases approximately linearly with the number of confusable pairs. However the number of confusable pairs will increase roughly as the square of the vocabulary. Thus the scheme is suited for small vocabulary tasks, such as digit string recognition.

5. Results

The performance of the proposed scheme was evaluated on the AURORA 2.0 task. AURORA 2.0 is a small vocabulary digit string recognition task. As the vocabulary size (excluding silence) is only eleven (one to nine, plus zero and oh) the number of possibly confusable pairs is limited making it suitable for the proposed scheme. The utterances in this task are one to seven digits long based on the TIDIGITS database with noise artificially added. The clean training data comprises 8440 utterances from 55 male and 55 female speakers. For the idealised model-based compensation scheme, SPR, stereo data from 422 sentences (a subset of all the training data) are provided for each of 16 conditions: 4 different SNRs ranging from 20 to 5 dB, combined with the 4 different additive noise sources N1 to N4, subway, babble, car and exhibition hall. Each of the 16 conditions also has a test set of a 1001 sentences with 52 male and 52 female speakers. Only these 16 noise conditions are examined in this work as stereo data is not provided for any of the other noise conditions. A 39 dimensional feature vector consisting of 12 MFCCs appended with the zeroth cepstrum, delta and delta-delta coefficients was used. This differs from the standard parameterisation and performs slightly worse. However it is the form of frontend that can be easily used with practical model-based compensation schemes such as VTS. The acoustic models are 16 emitting state whole word digit models, with 3 mixtures per state and silence and inter-word pause models.

SNR	Noise				Avg
(dB)	N1	N2	N3	N4	
20	1.60	1.81	1.76	2.01	1.79
15	2.67	3.17	2.48	2.90	2.80
10	5.22	6.77	4.86	4.75	5.40
05	12.28	18.83	10.62	9.81	12.90
Avg	5.44	7.65	4.93	4.87	5.72

Table 1: SPR WER (%) on AURORA 2.0 test set A.

Table 1 shows the performance of the single-pass retrained system on each of the noise conditions. As expected, as the SNR decreases the word error rate (WER) increases. Note the word error rate for the clean, uncompensated, model set on the 5dB SNR system was 66.75%. There is an 80% relative reduction in error rate by using this idealised model-based compensation at 5dB. It is possible to achieve greater performance gains than this "idealised" set-up. The noise models can be estimated at a file-by-file level where a level of speaker-adaptation is effectively performed, see for example [5] where an error rate of 10.22% was obtained using VTS at 5dB SNR. It is expected that improvements in the model-based compensation scheme will be reflected in gains after the SVM rescoring stage.

A set of 20 confusable digit-pairs were selected based on the overall confusion matrix for the 16 noise conditions. In addition all insertion/deletion confusable pairs were selected (i.e. silence against each of the vocabulary words). A total of 31 confusable pairs were thus trained and used for rescoring. These covered approximately 80% of the total number of substitutions and all insertions/deletions. In order to check that the SVM can operate in a noise-independent fashion the N1 noise condition and the 5dB SNR condition were removed from the SVM training configuration. This means that there are 9 noise conditions to train the SVM. Note the model-based compensation is still run in an SPR fashion for the N1 and 5dB conditions to get the generative models for the noise-specific kernels. For these experiments SVMs were built using the top 1500 dimensions of $\boldsymbol{\phi}(\tilde{\mathbf{Y}}_i;\boldsymbol{\lambda})$ ranked using the Fisher ratio.

SNR (dB)	Noise N1	N2	N3	N4	Avg
20	1.38	1.51	1.55	1.79	1.56
15	2.00	2.42	2.18	2.68	2.32
10	3.56	4.41	4.21	4.13	4.08
05	7.22	11.09	8.35	8.52	8.80
Avg	3.54	4.86	4.07	4.28	4.19

Table 2: SVM rescoring WER (%) on AURORA 2.0 test set A, SVM trained on N2-N4 10-20dB SNR.

Table 2 shows the performance of the SVM rescoring for the 16 noise conditions. For the average of all the noise conditions the SVM rescoring has reduced the WER rate by 1.53% absolute, a 27% relative reduction in WER. If only the noise condition (N1) that the SVMs were not trained in are considered then the reduction is 1.90% absolute, a 35% relative, reduction. For the unseen SNR condition (SNR05) a 4.10% absolute, 32% relative, reduction in WER was obtained. From these preliminary results the noise-dependent kernels appear to allow a noise-independent SVM to be reliably used. It is interesting to note that the SVM rescoring reduced the error rate for all of the 16 noise conditions. Only two of the individual SVMs made the WER worse overall, and in these cases only by one-or-two errors (individual SVMs did increase the WER for particular noise conditions). The single fixed values for β and ϵ that were only crudely tuned on a single pairing appear to be fairly robust.

Test Condition	System	Errors (%) Sub	Del	Ins	Tot (%)
N1	SPR	2.56	0.45	2.43	5.44
	+SVM	2.15	0.58	0.81	3.54
SNR05	SPR	5.79	1.09	6.01	12.90
	+SVM	4.90	1.67	2.23	8.80
All	SPR	2.74	0.60	2.39	5.72
	+SVM	2.30	0.76	1.13	4.19

Table 3: SPR and SVM rescoring WER (%) on AURORA 2.0 test set A, SVM trained on N2-N4 10-20dB SNR.

Table 3 shows the breakdown of the errors for the held out conditions N1 noise and 5dB SNR as well as the overall performance. The most interesting condition is the 5dB SNR one. This has more insertions than substitutions for the SPR HMM system. On the insertion/deletion errors, SVM rescoring reduced the insertions from 6.01% to 2.23% whilst only increasing the deletions by 0.58%. This illustrates one of the aspects of the AURORA 2.0 task, for the lower SNR conditions handling insertions yields larger gains than substitutions.

6. Conclusions

This paper has described a new approach to noise-robust speech recognition. The scheme combines model-based noise compensation schemes with a discriminative classifier, in this case an SVM. Rather than adapting the discriminative classifier, changing noise conditions are handled by the modifying the kernel. As generative kernels are used, model-based compensation can be used to adapt the generative models used to obtain the kernel features to the specific noise environment. To handle the multi-class issue (the SVM is inherently binary) a modified version of acoustic code-breaking is used. Thus the scheme allows a noise-independent SVM with noise-dependent generative kernels to be used to rescore the recognition output from a standard HMM-based speech recognition systems. Initial experiments on the AURORA 2.0 task using an "ideal" model-based compensation scheme, single-pass retraining, are presented. To ensure that the SVMs could handle unseen noise conditions and SNRs, no data from the N1 noise condition and 5 dB SNR were used to train the SVMs. Compared to the SPR trained system large reductions in WER were observed with SVM rescoring for all noise conditions, including the ones on which the SVMs were not trained.

The results presented in this paper are preliminary for a number of reasons. Single-pass retraining has been used as the model-based noise compensation scheme. However, well implemented schemes yield both performance and parameters that are very close to these SPR systems. Discriminatively trained, or more complex, acoustic models can be used for this task. The proposed scheme can be applied using these improved generative models. Finally SVMs were used as the discriminative classifier. For larger vocabulary tasks other discriminative classifiers, such as conditional augmented models [10], may be more appropriate.

7. References

[1] VN Vapnik, *Statistical learning theory*, John Wiley & Sons, 1998.

[2] X Li and J Bilmes, "Regularized adaptation of discriminative classifiers," in *Proc. ICASSP*, Toulouse, France, 2006.

[3] PJ Moreno, *Speech Recognition in Noisy Environments*, Ph.D. thesis, Carnegie Mellon University, 1996.

[4] MJF Gales, *Model-based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, Cambridge University, 1995.

[5] J Li, L Deng, Y Gong, and A Acero, "High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector taylor series," in *ASRU 2007*, Kyoto, Japan, 2007.

[6] T Jaakkola and D Hausser, "Exploiting generative models in disciminative classifiers," in *Advances in Neural Information Processing Systems 11*, S.A. Solla and D.A. Cohn, Eds. 1999, pp. 487–493, MIT Press.

[7] Smith, ND and Gales, MJF, "Speech recognition using SVMs," in *Advances in Neural Information Processing Systems*, 2001.

[8] RA Gopinath, MJF Gales, PS Gopalakrishnan, S Balakrishnan-Aiyer, and MA Picheny, "Robust speech recognition in noise - performance of the IBM continuous speech recognizer on the ARPA noise spoke task," in *Proc. ARPA Workshop on Spoken Language System Technology*, Austin, Texas, 1995.

[9] H Liao and MJF Gales, "Adaptive Training with Joint Uncertainty Decoding for Robust Recognition of Noise Data," in *Proc. ICASSP*, Honolulu, USA, 2007.

[10] MI Layton and MJF Gales, "Augmented statistical models for speech recognition," in *Proc. ICASSP*, Toulouse, 2006.

[11] V Venkataramani, S Chakrabartty, and W Byrne, "Support vector machines for segmental minimum Bayes risk decoding of continuous speech," in *ASRU 2003*, 2003.