Proceedings of the Institute of Acoustics

DATA-DRIVEN CLUSTERING AND INTEGRATION FOR SPEECH RECOGNITION

S. Douglas Peters Nuance Communications, 111 Duke St., Montréal, CANADA, H3C 2M1

1 Introduction

It has been demonstrated throughout the literature that there are recognition gains to be had by clustering available training data according to speaker sex, channel, noise conditions or speaking rate. Essentially, these results demonstrate that models conditioned according to categories of variability are more effective than unconditioned models. Recently, there have been a number of papers extending this concept in systematic ways. Padmanabhan et al. move from two gender clusters to many speaker clusters in 1996 [1]. In 1998, Kuhn et al. introduced eigenvoices, a method that provides both a novel mechanism of clustering and an efficient adaptation [2].

Both of these works promise rapid adaptation due to the conditioning of the pre-trained cluster models. Looking at the matter another way, the adaptation mechanism has been partially effected in the training procedure. Given that training corpora are continually expanding, it makes sense to exploit their richness as much as possible. By providing for macro-adaptation in this manner, the on-line adaptation procedure now reduces to micro-adaptation. As a result, it now requires much less adaptation data to achieve successfully adapted models.

We propose an eigenvoice-like architecture in which adapted models are constructed as a linear combination of pre-trained cluster models, deviating from the eigenvoice approach in four important ways:

1. phonetic class-based cluster models,
2. data-driven cluster model definition,
3. non-linear programming for cluster coefficient calculation, &
4. cluster coefficient export to high-definition models.

Recently, eigenvoices and speaker clustering have taken turns toward class-dependent modeling. Kenny, Boulianne and Dumouchel consider cluster modeling at the Gaussian level [3]. Jiang and Huang observe gains by adding subword dependence to speaker clustering [4]. Considering that the modeling-unit dependence of traditional adaptation methods has been considered in great detail, this research trajectory is not surprising. Unfortunately, the smaller the granularity of unit-dependence,

the higher the parametric complexity of the adaptation mechanism will be. In consequence, the adaptation time is once again increased.

Given that the trend is toward clusters over phonetically-derived classes, it is clear that the value of speaker conditioning is somewhat lost. In effect, we cannot expect a single construction of training speakers to match a test speaker for every phonetic class. As a result, we consider the case of data-driven clustering over such classes, hoping to capture all variabilities within a consistent mechanism. That is, inter-speaker variability, intra-speaker variability, channel variability and noise variability are all handled without bias in the clustering process.

In the matter of cluster coefficient estimation, we follow Huo and Ma who constrain cluster coefficients so that the resulting adaptation models are sure to lie inside the simplex defined by the cluster models [5]. We have found that unconstrained projection methods result in far too many inappropriate adaptations.

Finally, cluster coefficients are estimated in the domain of simple models but applied in the domain of complex models. In this manner, the coefficients can be estimated smoothly and efficiently while the resulting models can be at the state of the art. By applying adaptation independently to all units in all rescoring hypotheses, the proposed method gives significant same-syllable recognition gains on a difficult wireless pseudo-isolated digit task.

In the following section, a brief literature review is undertaken. Subsequent sections will outline the present strategy for extending eigenvoices with data-driven clustering and class-based coefficient estimation. A simple experiment is then reported on which the concepts outlined here are shown to be effective. Finally, the work is discussed and conclusions are made.

2 Previous work

In this section, three research efforts are compared. Presented in chronological order, we consider speaker clustering [1], eigenvoices [2] and adaptive prior fusion [5]. Each of these strategies maintains training-data clusters on the basis of speaker. That is, the critical source of variability in each case is considered to be due to inter-speaker differences.

2.1 Speaker Clustering [1]

The following Figure stylizes (for presentational purposes) the speaker clustering methodology of Padmanabhan et al [1]. Offline, many models $\{\mathbf{m}_1 \mathbf{m}_2 \cdots \mathbf{m}_M\}$ are trained on speech data from M training speakers. In effect, each speaker's speech data represents a separate data cluster (DC).

Adaptation data is used first to identify $N < M$ training speakers that will be considered "similar" to the source of the adaptation data. Then, an adapted model $\mathbf{b}$ is derived using the adaptation data from those N similar speaker models using MLLR techniques.

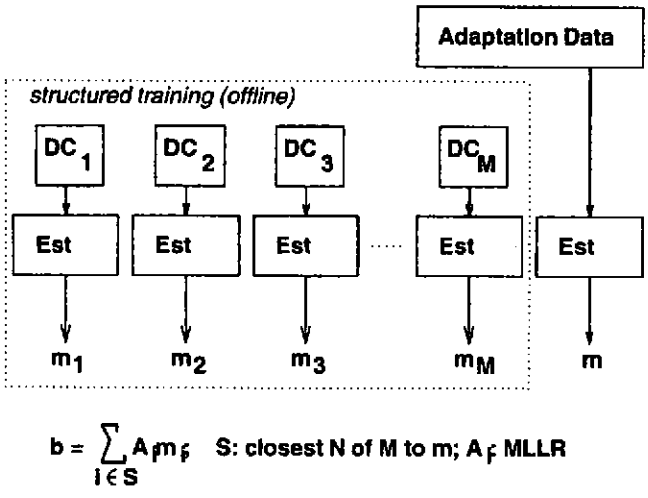

Figure 1: Stylized Speaker Clustering

This approach leads to significantly improved adaptation over an MLLR baseline. Unfortunately, gains were not reported as a function of adaptation time, as it is likely that the speed of adaptation was enhanced by this approach considerably.

One of the significant drawbacks of the technique is the fact that each of the speaker models must be used to decode the adaptation data, resulting in a computationally intensive adaptation phase. This difficulty is largely eliminated in the strategies considered below.

2.2 Eigenvoices [2]

The eigenvoice approach due to Kuhn et al. provides an optimal compression of M single-speaker models into L "eigenvoices" that span the most significant dimensions of speaker space [2]. The key to the eigenvoice development is the appreciation that any amount of speech can be considered to reside in the space of "super-vectors" consisting of HMM parameters. While it is possible to make use of any and all such parameters, it is typical to take only the means of the Gaussians by which the HMM observation pdfs are parameterized. If HMM parameters are re-estimated over as little as a single utterance, that utterance is implicitly represented in the appropriate phonetic subspace. As a result, a single utterance (adaptation or testing) can be compared directly to pre-trained models.

On the basis of super-vector mathematics, the relatively costly MLLR adaptation can be replaced with the very efficient MLED projection mechanism. A considerable reduction in the parametric complexity of the adaptation itself results. Indeed, only the vector $\mathbf{c} = \{c_1 c_2 \cdots c_L\}$ is estimated. Because very few parameters need to be estimated, very little adaptation data is now necessary. Figure 2 sketches

the eigenvoice strategy.

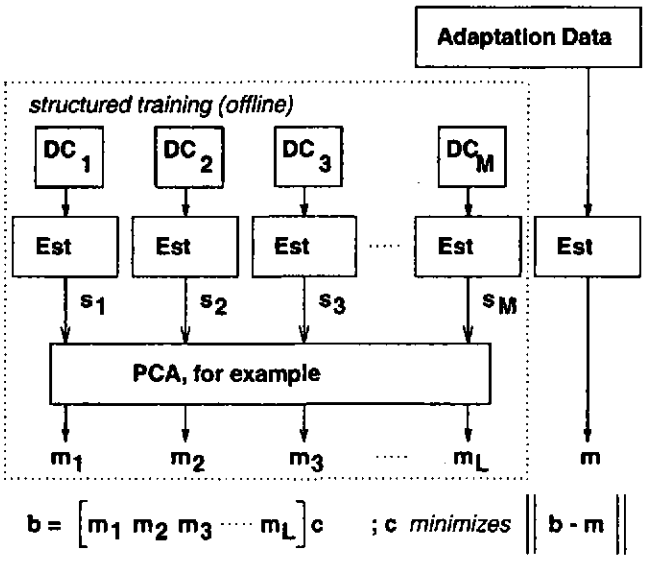

Figure 2: Eigenvoice Architecture

Here, the single-speaker models as used in speaker clustering (represented by their super-vectors $\mathbf{s}_m$ $m \in \{1, 2, \cdots, M\}$) are replaced by eigenvoices (represented by $\mathbf{m}_\ell$ $\ell \in \{1, 2, \cdots, L\}$) determined via a dimensionality reducing mechanism such as Principal Component Analysis (PCA). In effect, only the most significant $L \ll M$ dimensions of the speaker variability are retained.

While the strengths of the eigenvoice strategy are considerable, one of its inherent weaknesses is the fact that reliable estimation of the cluster coefficients depends, to some degree, on the smoothness of the eigenvoices themselves. Indeed, the examples demonstrating both eigenvoices and speaker clustering use single-Gaussian pdf parametrizations [1, 2, 6]. Clearly this represents a drawback if the final adapted model $\mathbf{b}$ is derived directly from the relatively simplistic cluster models $\mathbf{m}_\ell$. In the sequel, an architecture that solves this difficulty will be presented.

2.3 Adaptive Prior Fusion [5]

Huo and Ma use super-vectors to construct cluster-trained models in a manner analogous to the eigenvoice method [5]. It is intuitive that successful clustering should yield clusters that span a space similar to that derived from Principal Component Analysis. Indeed, our experiments have shown that the resulting subspaces of the two approaches are substantially the same [9]. In effect, the super-vector clustering done by Huo and Ma has a dimensionality reduction effect essentially equivalent to that of PCA.

The contribution of Huo and Ma is twofold. First, the cluster coefficients are constrained so that the adapted model $\mathbf{b}$ lies within the simplex defined by the cluster models $\mathbf{m}_\ell$. In effect, convex-set constraints are imposed:

$$\text{a)} \quad \sum_{\ell=1}^{L} c_\ell = 1$$
$$\text{b)} \quad c_\ell > 0; \quad 1 \leq \ell \leq L$$

Second, the need for adaptation data has been eliminated altogether: the recognition data itself is used for what in previous approaches was called adaptation. In our proposal to follow, we adopt both of these processing aspects. Figure 3 illustrates the adaptive prior fusion architecture.

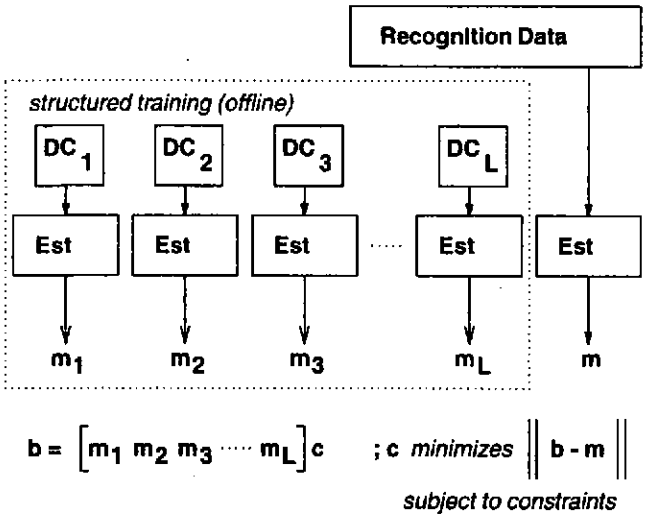

Figure 3: Adaptive Prior Fusion

In contrast with the preceeding methods, the data clusters (DC) in the adaptive prior fusion technique consist of the data from a collection of speakers, rather than from single speakers.

3 Hypothesis-Driven Adaptation

As previously mentioned, each of the strategies surveyed above maintains training-data clusters on the basis of speaker. While it is clear that inter-speaker variability is a considerable challenge, it is not clear that this challenge renders insignificant other sources of variability. Indeed, observations have shown that intra-speaker variability can be as significant as inter-speaker variability [7]. Moreover, variabilities due to environmental factors can also be considerable, and just as critical to the commercial success of network-oriented speech recognition. In Section 3 we introduce an extension to the works considered here that handles all manners of variability consistently.

Via Figure 4, we introduce an architecture that we have called Hypothesis-Driven Adaptation (HyDrA) [9]. There are two principal extensions of previous work here. First, there is an explicit dependence on phonetic context (denoted by the subscript p in Figure 4). In effect, for every phonetic unit p, an independent clustering takes place. These units should be relatively small: between HMM states and syllables in length. Phonetic dependence multiplies the number of adaptation units. On the other hand, the observed variability on any one phonetic unit in the training set will certainly be more contained than the variability across all phonetic units. As a result, we need many fewer per-unit clusters as the phonetic domain over which we apply those clusters reduces. As an example of this, we compare the fourteen dimensions reported to be necessary to capture half the variability in the spoken English alphabet [2] to the eight dimensions require to capture half the variability in a single digit [9].

The second principal extension to previous work is that we engage the cluster coefficients derived from diffuse (smooth) models $\mathbf{m}_{ip}$ and use them to construct adapted models from complex models $\mathbf{m}_{ip}$. In this manner, we can obtain high quality estimates of cluster coefficients based on smooth models, while using those same coefficients as the basis for constructing high-quality models for recognition. Required for this to be effective is the fact that the data clusters active in the two domains are identical. One can consider that the cluster coefficients under the convex constraints mentioned above represent the probability of membership in each of the clusters, i.e.,

$$c_{\ell p} \sim p(\mathbf{RD}_p \in \mathbf{DC}_{\ell p}).$$

Under this interpretation, the export of the cluster coefficients from one domain to the other is rendered permissible by the maintenance of the data clusters across domains. Also necessary is the fact that parameter re-estimation is explicitly dependent on segmenting models $\mathbf{m}_{0p}$ and $\mathbf{b}_{0p}$. Using these segmenting models as priors in MAP processing is appropriate. While this is not so critical in the context of the complex models, it is important to maintain a consistent relationship between the cluster models $\mathbf{m}_{\ell p}$ and the parameters estimated from the recognition data, $\mathbf{m}_p$.

In practice, HyDrA processing will be performed in a second pass. A first pass processing will provide an N-best list or lexical graph over which HyDrA models will be derived for every constituent phonetic unit. The convex optimization problem of determining the cluster coefficients will also yield a "success factor" f, which is given by

$$f_p = 1 - \min\left[\frac{J(\mathbf{c}_p)}{\|\mathbf{m}_p\|^2}, 1\right].$$

where the quadratic objective function J is given by

$$J(\mathbf{c}_p) = \|[\mathbf{m}_{1p}\mathbf{m}_{2p}\cdots\mathbf{m}_{Lp}]\,\mathbf{c}_p - \mathbf{m}_p\|^2.$$

When the variability capture is high, the objective function J is small relative to the norm of $\mathbf{m}$ and f is consequently close to unity. On the other hand, if the cluster models cannot represent the observed phonetic unit hypothesis, the objective function can be quite large, in which event the model adaptation falls back to the base models. For hypothesized phonetic units that are consistent with variability observed in training, models of much better quality than the base models will be derived.

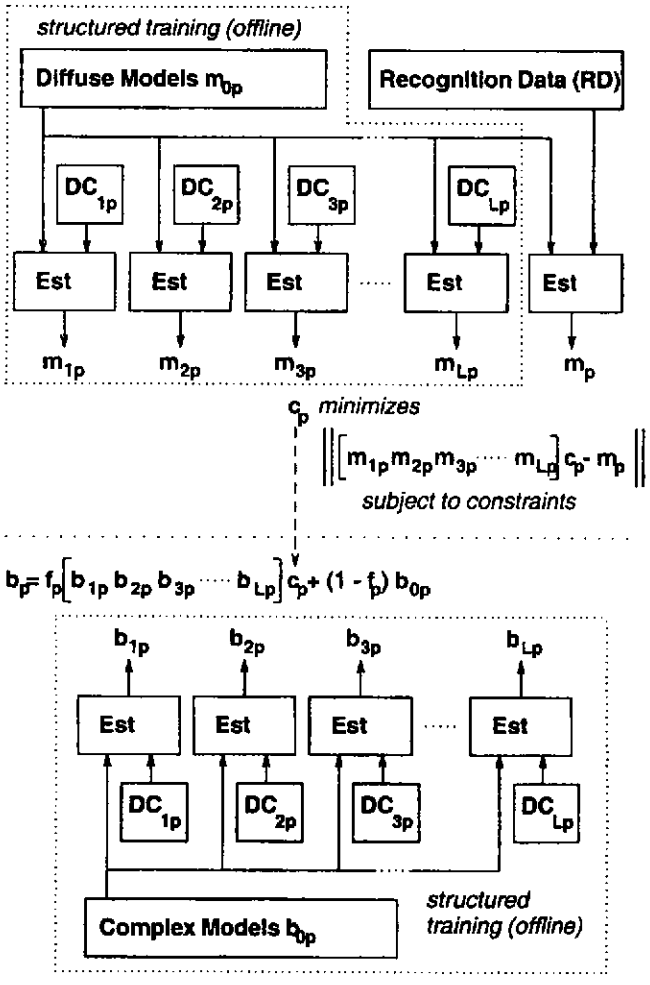

Figure 4: Hypothesis-Driven Adaptation

Unfortunately, the independent clustering over different phonetic units breaks the advantageous correlation inherent in speaker-oriented clustering. However, the approach enables every source of variability to be managed within a consistent mechanism. The correlatedness of different clusters over different different phonetic units will now be recovered via a separate level of processing.

For each of the P phonetic units we now have L data clusters capturing the significant variabilities observed for that unit in the training set. Moreover, each data cluster represents a subset of that training set. As a result, it is possible to construct a cluster co-occurrence matrix (somewhat analogous to that proposed in [4]).

For every utterance in the training set, we accumulate counts of pair-wise data cluster membership. This gives us a large $PL \times PL$ matrix capturing the correlations lost in the elimination of speaker-oriented clusters. Moreover, other significant correlations are also captured via this mechanism. With suitable normalization, the cluster co-occurrence matrix represents an estimate of the pair-wise conditional cluster probabilities.

For each hypothesis j in a recognition first pass of a given utterance, the average joint pair-wise probability of the set of estimated adaptation coefficients will be estimated by the quadratic form $\mathbf{C}_j^T \mathbf{P} \mathbf{C}_j / N_j$ where $\mathbf{C}_j$ is the concatenation of cluster coefficients for the hypothesis j, N_j is the total number of phonetic units for that hypothesis and $\mathbf{P}$ is the pair-wise conditional cluster probability matrix. We would expect that cluster coefficients in keeping with the behaviour of the training corpus would score better with this criterion than those derived from misrecognition hypotheses.

This explicit evaluation of the cluster coefficient probabilities represents a super-phonetic but sub-lexical level of processing. Research has suggested that human modeling of speech includes processing at an intermediate level such as this [8]. As a result, it is possible that a super-phonetic layer, made available by the proposed mechanism, for example, may represent a missing element for most modern speech recognition systems.

4 Experiments

In this section, results of a simple proof-of-concept experiment are reported. Hands-free, cellular, noisy "telephone number" Quebec French utterances were segmented by an off-the-shelf recognizer into pseudo-isolated single-digit utterances that provided both training and test material. Approximately thirteen thousand of each of ten digits were obtained in this manner. Of these, one thousand of each digit were set aside for testing and the remainder were used for training. Note that the fact that the test set matches the training set in terms of all its variabilities is quite artificial. Our benchmark processing benefits equally from this match. Moreover, the assumption implicit in the HyDrA development, however optimistic, is that the training set is sufficiently rich to observe any testing variability.

A "base" model, to be used as a benchmark as well as $\mathbf{b}_0$, was trained having a mixture of eight Gaussians for state observation pdfs. Features were extracted according to the standard mel-frequency

cepstrum processing. Nine such features, their deltas, and the first and second time-derivative of energy were concatenated together to provide twenty features at intervals of 12ms. Cepstral mean subtraction, with estimates generated over an entire digit, was employed as a feature debiasing mechanism. Another model, $\mathbf{m}_0$ was trained having single-Gaussian pdf parameterizations and twelve-component feature vectors.

As a first investigation, the words themselves were taken to be the phonetic unit on which HyDrA depends. Unfortunately, this has the effect of eliminating any contribution to recognition that may have been obtained by the super-phonetic processing. It will, however, provide a first look at a more primitive HyDrA processing.

Given roughly ten states per digit and twelve parameters per state mean, k-means clustering on around thirteen thousand vectors in one hundred twenty dimensions yielded eight cluster models, $\mathbf{m}_{\ell p}$, for each digit. The training sub-corpora corresponding to each of these cluster models were then used to re-train the complex model $\mathbf{b}_0$ to obtain the complex cluster models $\mathbf{b}_{\ell p}$.

While HyDrA is designed to be performed in a second pass, the present experiment applied HyDrA processing to all ten digits. Each test utterance was decoded as each possible digit, and parameters were re-estimated to obtain $\mathbf{m}_p$; $p \in \{0, 1, \cdots, 9\}$. Note that this process is very fast due to the small feature vector and the low complexity of the diffuse models $\mathbf{m}_{0p}$.

Cluster coefficients for each digit were now obtained using $\mathbf{c}_p = \operatorname{argmin} J(\mathbf{c}_p)$ under the constraints described above. Now, adapted models $\mathbf{b}_p$; $p \in \{0, 1, \cdots, 9\}$ were constructed in accordance with:

$$\mathbf{b}_p = f_p [\mathbf{b}_{0p}\ \mathbf{b}_{1p} \cdots \mathbf{b}_{8p}]\, \mathbf{c}_p + (1 - f_p)\mathbf{b}_{0p}.$$

Once again, this procedure is an efficient one. The relatively few constraints render the solution of the convex programming problem manageable, and the complex model construction is very simple by comparison to the subsequent decoding on its basis.

We remark that while the HyDrA construct requires a large number of parameters, most of these are abstracted from the recognition procedure. In fact, the number of "active" parameters, i.e. those used explicitly for recognition purposes, are only increased by a small proportion from the benchmark decoding via $\mathbf{b}_0$. Further, this proportion would be even smaller should the complexity of $\mathbf{b}_0$ be increased in keeping with standard practice. In the present case, however, there was insufficient training data to support large Guassian mixtures in high-dimensional spaces.

The results of the experiment are tabulated below. The 24% reduction in error rate is not uniformly distributed across the digits. Rather, the digits for which the benchmark model was least effective benefit the most from HyDrA processing.

5 Conclusions

It is clear that a systematic capture of variability in training enables efficient adaptation in recognition. A number of earlier strategies toward this end have been surveyed and a new processing architecture

Table 1: Recognition summary

digit:	1	2	3	4	5	6	7	8	9	0	all
HyDrA ($L = 8$)	96.3	95.9	98.5	96.0	97.2	95.3	95.1	96.6	96.3	98.9	96.6
benchmark (b_o)	94.9	94.3	98.5	95.4	96.7	91.7	91.3	96.6	96.4	98.9	95.5

has been introduced. Based on developments in eigenvoices [2], the present approach follows [5] in its use of recognition material for adaptation. By imposing phonetic-unit dependence on this processing, we need to adapt for each first-pass recognition hypothesis. Hypothesis-driven adaptation performs this adaptation in an inexpensive modeling domain and then exports the cluster (adaptation) coefficients into a more complex modeling domain by way of common training sub-corpora. Testing on a simple but very challenging recognition task demonstrates that HyDrA processing yields significant recognition gains. Further results and analysis are forthcoming in [9].

6 Acknowledgements

This article reports work that was done while the author was with Nortel Networks OpenSpeech Labs, Montréal. The co-operation and encouragement of Daniel Boies, Serge Robillard, Peter Stubley, Matthieu Hébert and Benoit Dumoulin are gratefully acknowledged. Nortel Networks closed its speech recognition research operations in the fall of 1999. The author also wishes to thank the staff at Nuance Communications for their continued support of this fascinating research.

References

[1] Padmanabhan M., Bahl L. R., Nahamoo D. and Picheny M. A., (1996) "Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems," *Proc. ICASSP '96*. Atlanta, pp. 701–704.

[2] Kuhn R., Nguyen P., Junqua J.-C., Goldwasser L., Niedzielski N. A., Fincke S. C., Field K. L. and Contolini M., (1998) "Eigenvoices for speaker adaptation," *Proc. ICSLP '98*, Sydney, pp. 1771–1774.

[3] Kenny P., Boulianne G. and Dumouchel P., (2000) "Bayesian adaptation revisited," *Proc. ASR2000*, Paris, pp. 112–119.

[4] Jiang L. and Huang X., (2000) "Subword-dependent speaker clustering for improved speech recognition," *Proc. ICSLP2000*, Beijing, vol. IV, pp. 137–140.

[5] Huo Q. and Ma B., (2000) "Robust speech recognition based on off-line elicitation of multiple priors and on-line adaptive prior fusion," *Proc. ICSLP2000*, Beijing, vol. IV, pp. 480–483.

[6] Kuhn R., Nguyen P., Junqua J.-C., Boman R. C., Niedzielski N. A., Fincke S. C., Field K. L. and Contolini M., (1999) "Fast speaker adaptation using a priori knowledge," *Proc. ICASSP '99*, Phoenix, pp. 749–752.

[7] Peters S. D. and Stubley P. (1998) "Visualizing speech trajectories," *Proc. ESCA TRW '98*, Rolduc, pp. 97–101.

[8] Peters S. D., Stubley P. and Valin J.-M., (1999) "On the limits of speech recognition in noise," *Proc. ICASSP '99*, Phoenix, pp. 365–368.

[9] Peters S. D. (2001) "Hypothesis-driven adaptation (HyDrA): a flexible eigenvoice architecture," *to appear in Proc. ICASSP '01*, Salt Lake City.