HYPOTHESIS-DRIVEN ADAPTATION (HYDRA): A FLEXIBLE EIGENVOICE ARCHITECTURE

S. Douglas Peters

Nuance Communications, 111 Duke St., Montréal, CANADA, H3C 2M1, peters@nuance.com

ABSTRACT

In this article, a new architecture for speech recognition is introduced. As with many existing speech systems, this new approach involves multi-pass processing. In the present case, however, second-pass models are constructed on-line for each active hypothesis. Models for each hypothesized segment of the current utterance are constructed from linear combinations of "data cluster models" that have been trained on low-variability clusters of the training corpus. The data cluster weights are determined using an "eigenvoice" mechanism that is operative on low-complexity, low-definition models. Once determined, the same weights are used to construct high-complexity, high-definition second-pass models generated over the *same* data clusters. Results from a simple recognition task are reported to demonstrate the interesting properties of the new architecture. The limitations, trade-offs and some possible extensions of the proposed approach are discussed.

1. INTRODUCTION

State-of-the-art speaker-independent ASR systems perform considerably worse than corresponding speaker-dependent systems. The natural solution to this difficulty is adaptation: adjusting either models or features in a manner appropriate to the current speaker and environment. Unfortunately, while humans seem to be able to adapt to a new speech environment in just a few syllables [1], commercial ASR adaptation requires considerably more adaptation data.

There are two adaptation mechanisms that have received considerable attention in the last decade. The first, Maximum A Posteriori (MAP) adaptation perturbs the model parameters in the direction of the observed speech [2]. The second, Maximum Likelihood Linear Regression (MLLR) adaptation estimates a linear transformation that results in a maximum likelihood score using the given models [3, 4].

Two essential difficulties plague acoustic model adaptation. First, the appropriate transformation or perturbation has been shown to depend highly on the phonetic class of the observations. As a result, it is difficult to use data from a given phonetic class to adapt data from a significantly different phonetic class [5]. Second, both of these conventional techniques require a relatively large amount of data on the basis of the usual relationship between training data and parametric complexity. For commercial speech applications, however, we would wish to have reliable adaptation on the basis of minimal data. In fact, it would be highly desirable to have an adaptation process that provides significant gains on the very first word spoken by any given speaker.

Recently, Kuhn et al. introduced "eigenvoices," a mechanism that relies on considerably fewer parameters than either MLLR or MAP [6]. In consequence, it has been shown to provide successful adaptation on very little data indeed. A similar adaptation mechanism was recently proposed in [7].

There are three admitted weaknesses with the eigenvoices strategy. First, the mechanism seems to work best in a "diffuse" model space. Indeed, the observation probability density functions (pdfs) of the eigenvoice examples in the literature are Gaussians rather than the common, and typically much more effective in terms of recognition power, mixture-of-Gaussians. Second, it is not at all clear that a given speaker cannot have entirely different eigenvoice coefficients in different phonetic contexts. That is, it is possible that the phonetic dependence so critical to traditional adaptation mechanisms may also be active in the present case. In keeping with this consideration, recent work has demonstrated gains by maintaining speaker clusters at a subword-level [8]. Finally, eigenvoice gains have been most noticeable in the context of speakers for whom training data is available. This is in direct constrast to the more conventional adaptation mechanisms that are specially effective for speakers for whom some speaking characteristics do not exist in the training corpus [5]. Further to this observation, while the speaker-specific clustering inherent in the eigenvoice literature is natural and convenient, we observe that the intra-speaker variability of speech may well be of the same magnitude as the inter-speaker variability [9].

The present article attempts to address each of these issues. Indeed, the three significant departures of the present work from that due to Kuhn et al. are:

- Eigenvoice coefficients estimated from diffuse-model eigenvoices are exported to a complex model context for rescoring.
- Eigenvoices are constructed independently for phonetic classes (i.e., adaptation units).
- Data clustering rather than speaker identity is used as the basis for eigenvoice construction.

In the next section a new architecture for speech recognition that we will call **Hy**pothesis-**Dr**iven **A**daptation (HyDrA) will be introduced. Subsequent sections will report results of HyDrA processing on a simple recognition task and discuss their implications.

This article reports work that was done while the author was with Nortel Networks OpenSpeech Labs, Montréal. The helpful discussions and encouragement of Drs. D. Boies and B. Dumoulin are gratefully acknowledged.

2. HYDRA ARCHITECTURE

The eigenvoice concept depends on the observation that any amount of speech can be considered to reside in the space of "super-vectors" consisting of HMM parameters [6]. While it is possible to make use of any and all such parameters, the present work takes only the means of the Gaussians by which the HMM observation pdfs are parametrized. If HMM parameters are re-estimated over as little as a single utterance, that utterance is implicitly represented in the appropriate phonetic subspace. In the development that follows, this re-estimation procedure appears for both training sub-corpora and individual utterances. As a result, single utterances can be related directly to training sub-corpora.

In Figure 1, the proposed architecture for speech processing is illustrated. Data clusters for a given adaptation unit, represented by training sub-corpora and constructed to reduce the variability within each cluster, are labelled $DC_1\ DC_2\ DC_3 \cdots DC_L$. These clusters are identical above and below the dotted line which delimits the domain of diffuse models from that of complex models. The training of the unit-data-cluster models $\mathbf{m}_\ell$ (the eigenvoices) and $\mathbf{b}_\ell$ is accomplished using a re-estimation mechanism that maintains parameter-wise alignment across all of the these models. In effect, the segmentation that is active for all re-estimations is due to the same models $\mathbf{m}_0$ or $\mathbf{b}_0$ in the diffuse and complex domains, respectively.

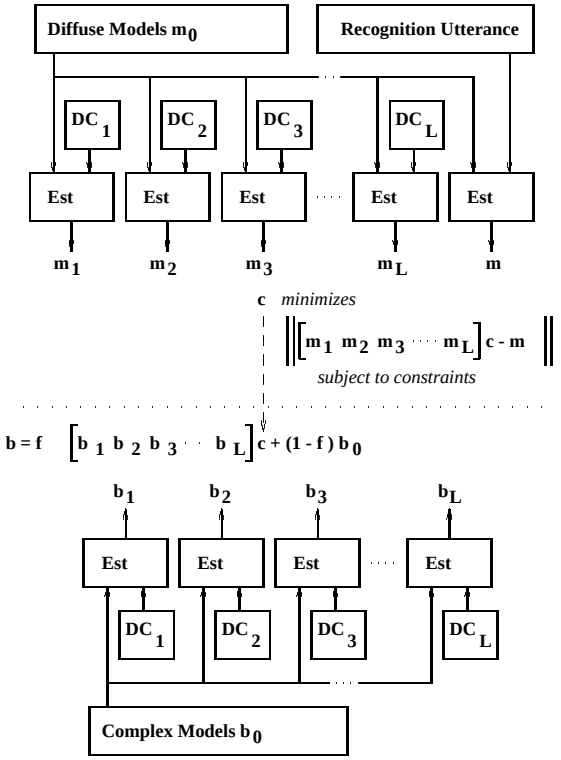

Fig. 1. Hypothesis-Driven Adaptation

For each utterance presented to the recognizer, a first-pass recognition reduces the search space down to a N-best list or word lattice. For each adaptation unit k represented in this rescoring domain, adaptation coefficients $\mathbf{c}_k$ are estimated. First, a super-vector is constructed by re-estimating the diffuse model parameters with the utterance at hand. A constrained projection $\hat{\mathbf{m}}_k$ of this supervector onto the space of appropriate eigenvoices $\mathbf{m}_{\ell,k}$ is now obtained. In effect,

$$\hat{\mathbf{m}}_k = [\mathbf{m}_{1,k}\ \mathbf{m}_{2,k} \cdots \mathbf{m}_{L,k}]\ \mathbf{c}_k$$

The resulting adaptation coefficients are then exported into the complex model space, and the adapted complex models $\mathbf{b}_k$ are constructed using the corresponding linear combination of the cluster-re-estimated complex models. Rescoring is now performed on the N-best list or word lattice using the constructed complex models.

The cluster-match mechanism at work in Figure 1 is neither a simple projection nor the MLED from [6], which is essentially a masked transformed projection. Rather, the adaptation coefficients are the solution to the "doubly-convex" optimization problem. That is, we minimize the quadratic (convex) function

$$J(\mathbf{c}) = \|[\mathbf{m}_1\mathbf{m}_2 \cdots \mathbf{m}_L]\,\mathbf{c} - \mathbf{m}\|^2$$

under the convex-set constraints

$$\begin{array}{ll} \text{a)} & \sum_{\ell=1}^{L} c_\ell = 1 \\ \text{b)} & c_\ell > 0; \quad 1 \le \ell \le L \end{array}$$

In this manner, we are certain to obtain a solution that lies "inside" the simplex defined by the data cluster models.

A final mechanism is necessary to ensure that the export of the adaptation coefficients from the diffuse model space to that of the complex models is meaningful. This is represented by a "fall-back" strategy for those utterance-choice pairs in which the observation cannot be represented accurately by the data clusters. A variability-capture coefficient f is introduced, and used to provide an interpolation between the base complex models ($\mathbf{b}_0$) and those suggested by the adaptation coefficients, as illustrated in Figure 1. For the present, we will use the simple relation

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

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 given adaptation unit, the objective function can be quite large, in which event the model adaptation falls back to the base models.

3. EXPERIMENTATION

HyDrA processing is now illustrated in the context of a small-vocabulary pseudo-isolated word task. A corpus consisting of around 15k noisy wireless "telephone number" utterances spoken in Quebec French were automatically segmented to create roughly 13k utterances of each of the ten French digits: *UN, DEUX, TROIS, QUATRE, CINQ, SIX, SEPT, HUIT, NEUF & ZERO*. Of these, one thousand of each digit were randomly selected and set aside as test sets, and the rest were used for training. Note that though the digits were spoken "continuously," the automatic segmentation results in artificially isolated words.

A simple mono-Gaussian-pdf CDHMM (i.e., the "diffuse" model) was created and trained independently with each training utterance. Word-specific models having between nine and twelve states were used to represent each digit. Feature vectors consisted of five MFCCs, their "deltas," delta-energy and delta-delta-energy.

Note that the use of a word as an adaptation unit was for concept exploration only. This issue will be discussed in a subsequent section.

The supervectors obtained by the re-estimation of the training utterances were clustered using the standard k-means algorithm. The training set partitions implicitly resulting from this clustering were then used to construct eigenvoices for each word for both the diffuse model and a complex model. The feature vectors for the complex models are similar in construction to those of the diffuse model but with nine MFCCs rather than five, for a feature vector size of twenty. Mixture pdfs were made up of eight Gaussians each, and each (word-dependent) phone had a single tied full covariance matrix.

While the standard HyDrA processing would involve an N-best rescoring, this aspect of the recognition process has been relaxed within the context of this artificially constrained task.

3.1. Protocol

Five different processings were applied to the 10^4 test utterances:

- HyDrA processing: "A_{18}": $L = 18$; "A_8": $L = 8$
- Benchmark HMM processing:
  - "B_8" $\mathbf{b}_0$ (8 Gaussians/pdf)
  - "B_{32}" 16 Gaussians/pdf and gender clusters
  - "B_{144}" pdf fusion of $\mathbf{b}_1\ \mathbf{b}_2 \cdots \mathbf{b}_{18}$ (144 G/pdf)

The reason that three benchmarks are considered is that it is difficult to know of what a "fair" benchmark model consists. On the one hand, B_8 represents the case in which the conventional processing shares the same parametric complexity as the "active" HyDrA model. On the other hand, the parametric complexity of B_{144} is the same as the *total* parametric complexity of the HyDrA processing. Obtained by the pdf-level fusion of the HyDrA cluster models, this benchmark makes each of the HyDrA parameters active in recognition. The remaining benchmark is another interesting case: an intermediate parametric complexity with a more traditional clustering strategy.

3.2. Results

Recognition results for the five processings under consideration are summarized in Table 1, below. The almost 50% ERR for the digit *SEPT* bears mention. Other results of interest include the surprising differences between the performance of B_8 and B_{32} for *DEUX* and *SIX*.

Table 1. Recognition summary

digit	A_{18}	A_8	B_8	B_{32}	B_{144}
1	96.3	96.3	94.9	95.1	96.0
2	95.8	95.9	94.3	93.8	95.2
3	98.6	98.5	98.5	99.2	99.0
4	95.5	96.0	95.4	96.1	96.4
5	97.3	97.2	96.7	97.2	97.7
6	95.4	95.3	91.7	95.1	96.1
7	95.2	95.1	91.3	91.8	91.7
8	96.9	96.6	96.6	95.4	95.9
9	96.8	96.3	96.4	97.1	97.0
0	99.1	98.9	98.9	98.9	99.2
all	96.7	96.6	95.5	96.0	96.4

3.3. Further analysis

In [6], an eigen-decomposition of speaker dependent models demonstrates that no fewer than fourteen dimensions (i.e., fifteen clusters) are required to capture half the variability in the spoken English alphabet. The same paper reports that the most significant dimension (eigenvoice) representing speaker sex is shown to capture less than 20% of all variability. Unfortunately, the reported observation is biased by the fact that the correlation matrix includes the "mean" supervector. Moreover, the variability over the speakers considered in [6] is also limited. Thus, the actual variation capture of the dominant dimension over all speakers is typically less than 10% for short words, and even less for longer words.

Variation dimensionality can similarly be examined by the construction of matrices whose columns are the super-vectors used in the clustering procedure used above. The typical matrix has between 108 and 144 rows (i.e., the number of states $\times$ the number of features for the diffuse models) and around 12k columns (i.e., the number of utterances).

The normalized cumulative squared singular values of the mean-removed matrices of this type were obtained to indicate the degree of variability present in a given number of dimensions. Figure 2 displays these curves for each of the digits. from this figure,

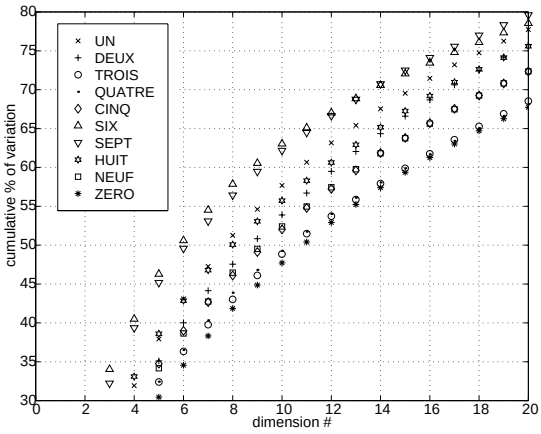

Fig. 2. Dimensional variation capture for French digits

we see, for example, that there is considerably greater variability in the two-syllable *ZERO* than in the one-syllable *SEPT*. Moreover, to capture 50% of the variability of the average digit, at least eight dimensions would be necessary. Note that this is less than reported in [6] due to the much smaller adaptation unit.

3.4. Discussion

The performance of HyDrA is clearly sensitive to the relative sizes of its components. For example, if the number of clusters is too large, the estimation difficulties inherent in traditional adaptation methods will arise. On the other hand, too few clusters will result in insufficient variability capture.

It is interesting to compare the recognition results of Table 1 (specifically the columns for A_8 and B_8) with Figure 2. The digits demonstrating the greatest recognition gains are also those with the greatest variation capture at a given dimensionality.

The interaction between HyDrA's inherent trade-offs is also interesting to consider. Figure 2 indicates gains for a large number

of clusters. Offsetting these gains, however, are estimation losses. First, there are the losses due to cluster training. For 18 clusters, for example, there were some clusters that were trained on less than 100 utterances. Second, there is the difficulty of the estimation of the adaptation coefficients themselves. Given our relatively short adaptation unit, it is clear that very few adaptation parameters can be reliably estimated. On the other hand, it bears emphasizing that the current architecture provides instantaneous adaptation on just a single word (often single syllable) of speech.

Another trade-off involves the relative parametric complexities of the diffuse and complex models. Here, the estimation of the adaptation coefficients is at odds with the validity of their export to the complex model domain. That is, the more similar the diffuse and complex models are, the more meaningful is the application in the complex domain of adaptation coefficients derived in the diffuse domain. On the other hand, the smoothing inherent in diffuse models provides for efficient estimation of the adaptation coefficients on a very small amount of data.

It bears mentioning that the HyDrA processing captures all variabilities without prejudice. As a result, it is likely that subtle phonetic variations present in tri- and penta-phone acoustic modeling could be captured in a HyDrA model using less defined (and fewer) decision-tree allophones, thus mitigating the parametric multiplication due to the HyDrA modeling.

4. FUTURE WORK

A final trade-off has been identified as critical for the future of this research. The size of the adaptation unit should not be so large that its inherent complexity requires too many clusters to capture. On the other hand, the present HyDrA architecture makes use of the long-term correlation implicit in a long adaptation unit.

In general, of course, one would like to be able to smooth adaptation parameters over a long duration of speech, having their estimation benefit from more data. To this end, we suggest a research trajectory below.

Let us choose a small adaptation unit, for example, an allophone state. This will result in a high degree of variability capture over a few dimensions. Further, let us construct a "cluster co-occurence matrix" that accumulates the times in which cross-adaptation-unit clusters co-exist in the same utterance over the training corpus. In effect, we abstract the long-term correlations that likely enhance HyDrA performance and re-apply them via an external mechanism. With suitable normalization, the cluster co-occurence matrix represents an estimate of the pair-wise conditional cluster probabilities. That is, we divide the co-occurence counts by the total number of times the associated adaptation unit is represented in each row.

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 adaptation coefficients for the hypothesis j, N_j is the total number of adaptation units for that hypothesis and $\mathbf{P}$ is the pair-wise conditional cluster probability matrix. We would expect that adaptation coefficients in keeping with the behaviour of the training corpus would score better with this criterion than those derived from misrecognition hypotheses. Of course, there are a number of possible ways in which to make use of this information. For example, the joint pair-wise adaptation coefficient probability could be added to the over-all hypothesis score or applied in a manner similar to the variability capture coefficient f.

Clearly, an external mechanism like the one outlined above is important to capture long-range correlations in the context of HyDrA processing. Moreover, it is interesting to consider such a mechanism to fill the role of a super-segmental processing apparently at work in human recognition [10]. We expect, in keeping with the behavior of traditional adaptation methods, that the averaging over a long duration of speech will enhance the adaptation performance considerably.

5. CONCLUSION

In this article, a novel method for instantaneous adaptation has been introduced. While cumbersome and computationally intensive, the present method provides gains over conventional processing of the same active parametric complexity. The behaviour of the proposed processing has been demonstrated at a state-of-the-art operating point on a pseudo-isolated word task in difficult circumstances (noisy wireless telephone speech) and a reasonably confusable lexicon of very short words.

6. REFERENCES

[1] Furui S. (1997) "Recent advances in robust speech recognition," *Proc. ESCA-NATO TRW '97*, Pont-à-Mousson, pp. 11–20.

[2] Gauvain J.-L. and Lee C.-H. (1994) "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains" *IEEE Trans. Speech Audio Process.* **2**, pp. 291–298.

[3] Leggetter C. and Woodland P. (1995) "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Comp. Speech Lang.* vol. 9, pp. 171–185.

[4] Sankar A. and Lee C.-H. (1996) "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Trans. on Audio and Speech Processing*, vol. 4, no.3, pp. 190–202.

[5] Zavaliagkos, G., Schwartz, R. and Makhoul J. (1995) "Batch, incremental and instantaneous adaptation techniques for speech recognition," *Proc. ICASSP '95*, vol. 1, Detroit, pp. 676–679.

[6] Kuhn R., Nguyen P., Junqua J.-C., Goldwasser L., Niedzielski N., Fincke S., Field K. and Contolini M. (1998) "Eigenvoices for speaker adaptation," *Proc. ICSLP '98*, vol. 5, Sydney, pp. 1771–1774.

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

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

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

[10] 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.