

# Maximum-Likelihood Training of a Bipartite Acoustic Model for Speech Recognition

Florent Perronnin, Roland Kuhn, Patrick Nguyen, and Jean-Claude Junqua

Panasonic Speech Technology Laboratory, Panasonic Technologies Inc.  
Santa Barbara, California, USA

(florent, kuhn, nguyen, jcj@research.panasonic.com)

## Abstract

In a recent paper, we described a compact, context-dependent acoustic model incorporating strong *a priori* knowledge and designed to support extremely rapid speaker adaptation [9]. The two parts of this “bipartite” model are: **1.** A speaker-dependent, context-independent (SDCI) part with a small number of parameters called the “eigencentroid”. **2.** A speaker-independent, context-dependent (SICD) part with a large number of parameters called the “delta trees”. For the first time, we describe in the current paper the iterative maximum-likelihood procedure employed to train both parts of the model. This paper also gives the first unsupervised adaptation and self-adaptation results for the new model, showing that it outperforms standard techniques when small amounts of adaptation data (10 sec. or less of speech) are available. Relative error rate reduction (ERR) is 12.1% for supervised adaptation and 11.2% for unsupervised adaptation on three TIMIT sentences; it is 10.4% for self-adaptation on a single TIMIT sentence. Finally, the paper analyzes the correlation between sex and the SDCI part of the model, and shows how modeling of acoustic variability is affected by the explicit separation into SD and CD components.

## 1. Introduction

In [1], Acero and Huang introduced a bipartite model with separate, additive SDCI and SICD components; Hazen and Glass subsequently elaborated on the idea [6]. The SDCI “centroid” component represents the mean output vector for each phoneme, or each phoneme state, while the SICD component models allophones as offsets from the appropriate portion of the centroid. To model a new speaker, one only needs to estimate the centroid (which contains a small number of parameters), since the SICD portion obtained by pooling data from training speakers remains valid for new speakers.

For a mean  $m$  of Gaussian  $d$  in a mixture modeling phoneme  $p$  in a particular allophone and state context, when the speaker is  $S$ , we thus have  $m_{p,d}^S = \mu_p^S + \delta_{p,d}$  where  $\mu_p^S$  is the portion of the SDCI vector pertaining to phoneme  $p$ , and  $\delta_{p,d}$  is the SICD component. In the “Eigencentroid plus Delta Trees” (EDT) bipartite model described in [9], the SDCI centroid  $\mu^S$  lies in a constrained “eigenspace” obtained via a dimensionality reduction technique from training speaker data, and is referred to as the “eigencentroid”. The  $\delta_{p,d}$  SICD component is typically modeled by decision trees, and is thus referred to as the “delta trees” component. Some rapid speaker adaptation approaches similar to EDT

require huge amounts of memory (*e.g.*, [3]). By contrast, EDT requires about the same memory as is required for an SI system. For instance, the SI baseline and the EDT systems in the experiments below have about 500,000 parameters each (EDT requires extra parameters to store the eigenspace, but the SI trees have more Gaussians than the delta trees).

The obvious way to estimate the EDT model would be to train a large number of CI models, one per training speaker. Then, one would apply a dimensionality reduction technique to the vectors representing the CI models to obtain an eigenspace (as in [10,11]). Finally, one would estimate for each training speaker a  $\mu$  centroid lying in the eigenspace, “normalize” each speaker  $S$ ’s segmented data for phoneme  $p$  by subtracting  $\mu_p^S$ , and then pool normalized segments across speakers to grow the  $\delta$  trees, using the method of [12].

The procedure described in the previous paragraph is precisely how we currently initialize our EDT models. Note that it may not be completely successful in placing only SDCI information into the  $\mu$  component, and only SICD information into the  $\delta$  component. Faulty estimation is particularly likely to occur when training data have dramatically different allophone frequencies for different speakers, leading to a confusion between SD and CD effects. For instance, consider set S1 of training speakers whose data happen to contain only examples of phoneme “aa” preceding fricatives, and set S2 whose examples of “aa” always precede non-fricatives. Since the procedure for estimating the eigenspace only has information about the mean feature vectors for “aa” for each speaker, it may “learn” that S1 and S2 are two different speaker types, and yield an eigenvector that correlates strongly with membership in S1 or S2, thus wrongly putting CD information in the  $\mu$  component. Note that CD effects may be considerably more powerful than SD ones [8], increasing the risk that this kind of error will occur while estimating the eigenspace.

The maximum-likelihood re-estimation procedure described here compensates for such allophone sampling effects by providing feedback between the estimated  $\mu$  and estimated  $\delta$  components of the EDT model. After giving details of the procedure, the paper examines its results, showing that just as in earlier work by ourselves and others [3,10,11], the first dimension of the eigenspace correlates strongly with sex. Comparison of the distributions in leaves of corresponding SI and  $\delta$  acoustic trees also yields two interesting results: **1.** The procedure shrinks the entropies associated with static acoustic features more than it shrinks entropies associated with dynamic acoustic features. **2.** The

procedure has a more powerful effect on models for diphthongs and vowels than it has on models for consonants.

## 2. Re-Estimation Formulae for EDT

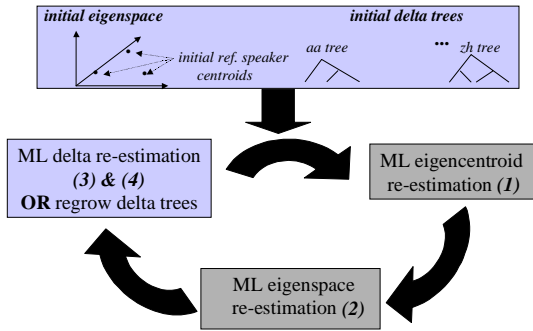


Figure 1 Re-estimation algorithm

Figure 1 shows the concept of the iterative maximum-likelihood (ML) re-estimation procedure. The iteration involves three mathematical objects: the training speaker eigencentroids (*i.e.*, the coordinates of each training speaker in the eigenspace), the eigenvectors defining the eigenspace, and the delta trees. Each delta tree leaf represents the output distribution of a phoneme for a given context as an offset from the SDCI eigencentroid. In our implementation, a typical leaf contains a mixture Gaussian of the form  $\sum_d c_d N_d(\delta_d, \sigma_d)$  where  $c_d$  is the weight for distribution  $d$ . In the ML framework, we maximize the likelihood of the observations given the data [2]. To re-estimate the Gaussian parameters, we must maximize the auxiliary function:

$$Q = \sum_{s,p,d,t} \gamma_{p,d}^s(t) \times \left\{ h(o_t, s, p, d) + \log | C_{p,d}^{-1} | \right\}$$

where  $h(o_t, s, p, d) = (o_t - \hat{m}_{p,d}^s)^T C_{p,d}^{-1} (o_t - \hat{m}_{p,d}^s)$  and  $o_t$  is the feature vector at time  $t$ ,  $C_{p,d}^{-1}$  is the precision matrix for distribution  $d$  of phoneme  $p$ ,  $\hat{m}_{p,d}^s$  is the adapted mean for distribution  $d$  of phoneme  $p$  for speaker  $S$ , and  $\gamma_{p,d}^s(t)$  is the occupation probability of distribution  $d$  of phoneme  $p$  for speaker  $S$  at time  $t$ . Recall that  $\hat{m}_{p,d}^s = \mu_p^s + \delta_{p,d}$  and that the centroid  $\mu$  is a linear combination of eigenvoice vectors:

$$\mu_p^s = e_p(0) + \sum_{j=1}^E w_s(j) e_p(j).$$

Here  $e(0)$  is a centre of mass vector for training speaker centroids. The dimensionality of the eigencentroid space is  $E$ .

### 2.1. Eigencentroid Re-Estimation

To re-estimate training speaker eigencentroids, assume fixed  $\delta$ 's and  $e$ 's. Set  $\frac{\partial Q}{\partial w_s(j)} = 0, j = 1, \dots, E$ . We derive the formula

$$\sum_{p,d,t} \gamma_{p,d}^s(t) e_p^T(j) C_{p,d}^{-1} (o_t - \delta_{p,d}) =$$

$$\sum_{p,d,t} \gamma_{p,d}^s(t) e_p^T(j) C_{p,d}^{-1} \sum_{k=1}^E w_s(k) e_p(k), \text{ for } j = 1, \dots, E. \quad (1)$$

This gives new coordinates  $w_s(1), \dots, w_s(E)$  for each  $S$  (and thus a new  $\hat{\mu}_p^s$  for each  $S$ ).

Note that precisely the same formula will be used to find the centroid for a new speaker during adaptation. For instance, for unsupervised adaptation, an SI recognizer would be used to find initial occupation probabilities  $\gamma$  for the speaker, leading to an initial estimate of the centroid  $\mu$ . In combination with the SI  $\delta$  trees, this would define an adapted CD model for the current speaker, yielding more accurate  $\gamma$ 's which could be re-estimated iteratively to give an increasingly accurate model for the speaker.

### 2.2. Eigenspace Re-Estimation

To re-estimate the eigenvectors spanning the eigenspace, assume fixed  $w$ 's and  $\delta$ 's. Set  $\frac{\partial Q}{\partial e_p(j)} = 0, j = 1, \dots, E$ .

We derive the formula  $[\sum_s (w_s(j))^2 \sum_{d,t} \gamma_{p,d}^s(t) C_{p,d}^{-1}] e_p(j) =$

$$\sum_{s,p,d} \gamma_{p,d}^s(t) w_s(j) C_{p,d}^{-1} (o_t - \hat{\mu}_p^s(j) - \delta_{p,d}), j=1, \dots, E \quad (2)$$

where  $\hat{\mu}_p^s(j) = \sum_{k \neq j} w_s(k) e_p(k)$ .

### 2.3. Delta-tree Re-Estimation

To re-estimate the  $\delta$ 's without changing the tree structure, assume that the  $w$ 's and  $e$ 's are fixed, and set  $\frac{\partial Q}{\partial \delta_{p,d}} = 0$ . We

obtain the formula

$$\delta_{p,d} = \frac{\sum_{s,t} \gamma_{p,d}^s(t) (o_t - \hat{\mu}_{p,d}^s)}{\sum_{s,t} \gamma_{p,d}^s(t)} \quad (3)$$

Let us assume in the following that the precision matrix  $C_{p,d}^{-1}$  is diagonal and that  $\sigma_{p,d}^2(i)$  is the  $i$ -th term on the diagonal of  $C_{p,d}^{-1}$ . If we want to re-estimate the variances

$\sigma_{p,d}^2(i)$ , we set  $\frac{\partial Q}{\partial \sigma_{p,d}^2(i)} = 0$ .

We derive the formula

$$\sigma_{p,d}^2(i) = \frac{\sum_{s,t} \gamma_{p,d}^s(t) (o_t(i) - \hat{m}_{p,d}^s(i))^2}{\sum_{s,t} \gamma_{p,d}^s(t)} \quad (4)$$

As Figure 1 shows, there is another way of computing new  $\delta$ 's, given new eigencentroids and a new eigenspace: regrow the trees. To do this, one resegments the training data, subtracts the appropriate portion of each speaker's eigencentroid from his or her training data, and pools the speaker-normalized phoneme segments to grow new  $\delta$  trees. Subsequently, one may apply the  $\delta$  re-estimation formula. Better speaker-normalization may lead to a better tree structure and thus to more accurate CD modeling [7].

### 3. Experiments

#### 3.1. Experimental Configuration

We carried out phoneme recognition experiments on the TIMIT database, with the standard partition of 462 training and 168 test speakers, and 8 sentences per speaker (about 3 sec. of speech per sentence). Speech was sampled at 16 kHz and parametrized using PLP cepstral features, with 9 static and 9 dynamic (first derivative) features. For training and EDT adaptation, there are 46 3-state phoneme HMMs and one 1-state HMM (for epenthetic silence); recognition employs a reduced 39-phoneme set. The language model was a bigram backoff. The baseline SICD system had 13,702 Gaussians. The EDT model with eigenspace dimension set to 1 had 11,451 Gaussians, while EDT (dim=5) had 11,553 Gaussians, EDT (dim=10) had 11,525 Gaussians, and EDT (dim=20) had 11,477 Gaussians. When the parameters for the eigenspace are taken into account, each of these systems including the SI requires roughly 500,000 floating-point parameters.

#### 3.2. Results

Tables 1 and 2 below show recognition rates when supervised and unsupervised adaptation were carried out on 1, 2, or 3 sentences of speech from the test speakers; testing was carried out on the remaining 7, 6, or 5 sentences. The MAP, MLLR, and MLLR=>MAP results shown here were obtained by adapting the SI baseline system (“MLLR=>MAP” denotes MLLR adaptation followed by MAP adaptation).

Figure 2 is derived from Table 2 and shows ERR for unsupervised adaptation. For both supervised and unsupervised adaptation on one adaptation sentence, MLLR and MLLR=>MAP lead to worse performance than SI. Overall performance is best for EDT (dim=20).

Table 3 shows self-adaptation results obtained over all 8 sentences by performing unsupervised EDT adaptation followed by recognition on each sentence (ignoring previous sentences from that speaker). Again, the best performance is obtained by EDT (dim=20).

Method	1 sent	2 sent	3 sent
SI	70.0%	70.1%	70.0%
EDT (dim=1)	71.2%	71.4%	71.3%
EDT (dim=5)	72.0%	72.3%	72.2%
EDT (dim=10)	72.3%	72.7%	73.1%
EDT (dim=20)	72.3%	73.2%	73.6%
MAP	70.2%	70.5%	70.5%
MLLR	69.7%	70.7%	71.1%
MLLR=>MAP	69.9%	70.9%	71.6%

Table 1 Supervised adaptation

Method	1 sent	2 sent	3 sent
SI	70.0%	70.1%	70.0%
EDT (dim=1)	71.2%	71.3%	71.3%
EDT (dim=5)	71.8%	72.1%	72.0%
EDT (dim=10)	71.9%	72.5%	72.6%
EDT (dim=20)	71.8%	72.8%	73.3%
MAP	69.9%	70.1%	70.1%
MLLR	69.0%	69.9%	70.5%
MLLR=>MAP	69.1%	70.0%	70.5%

Table 2 Unsupervised adaptation

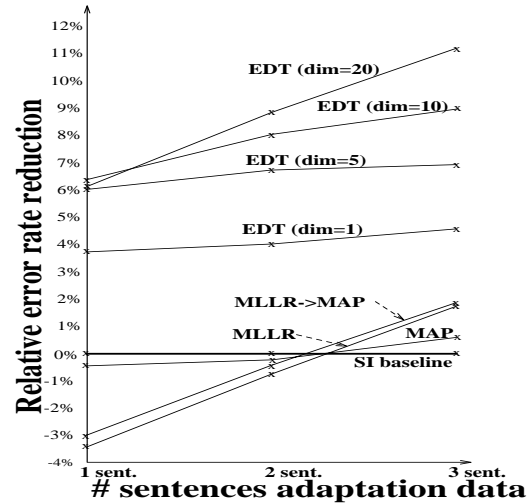


Figure 2 Uns. adaptation: rel. error rate reduction

Method	Results	Rel. Error Reduction
SI	70.1%	NA (0%)
EDT (dim=10)	72.9%	9.3%
EDT (dim=20)	73.2%	10.4%
MAP	71.0%	3.0%
MLLR	71.5%	4.7%
MLLR=>MAP	71.7%	5.3%

Table 3 Self-adaptation on single sentence

### 4. Analysis of Models

#### 4.1. Demographic Correlates of Eigenspace Location

Figure 4 shows the relationship between dimension 1 of the eigenspace for EDT (dim=20) and the training speaker’s sex. We employed the Lloyd algorithm [4] on the training speaker values of dimension 1 to fit a two-Gaussian model and find the optimal threshold between the Gaussians. On training speaker data, this classifier yields correct labels for 135 females (out of 136) and 324 males (out of 326). On test speaker data, the same threshold yields correct labels for all 56 females and 108 males (out of 112) – *i.e.*, 97.6% correct. We found only weak correlations between eigenspace location and other demographic factors such as age and dialect region.

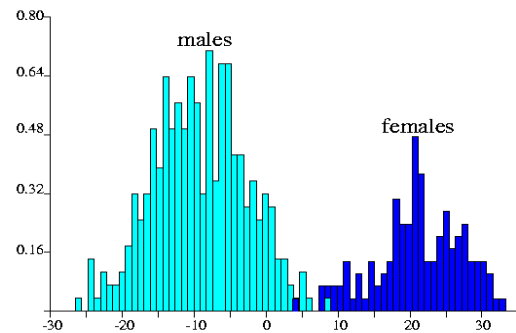


Figure 3 Value of dim. 1 and training speaker sex

## 4.2. Entropy Reduction Analysis

Why do the EDT models perform better than SI ones? We compared SICD and EDT (dim=20) single-Gaussian leaf distributions. First, we calculated average variances, weighting by the number of frames so well-populated leaves had more impact, and then the entropies [4]. The average, frame-weighted entropy of distributions is  $-18.5$  for the SI leaves and  $-20.0$  for the  $\delta$  leaves, an overall entropy decrease of 8%.

Table 4 shows this decrease as it affects each acoustic feature. Note that EDT has a much stronger impact on the PLP cepstral static coefficients than on the dynamic (first derivative) ones. The static coefficient entropies in the SI model ranged from  $-0.14$  to  $-0.60$ , while the dynamic entropies ranged from  $-1.35$  to  $-1.9$ . Evidently, the training procedure found it difficult to further lower the entropy of the dynamic coefficients. Also, perhaps static coefficients are more speaker-dependent than dynamic ones, and thus benefit more from EDT.

1. 17%	2. 28%	3. 50%	4. 54%	5. 34%	6. 37%
7. 29%	8. 27%	9. 90%	10. 1%	11. 1%	12. 2%
13. 2%	14. 2%	15. 2%	16. 2%	17. 2%	18. 1%

Table 4 Reduction in entropy due to EDT (by acoustic feature); features 1-9 static, 10-18 dynamic

Figure 4 shows SI and  $\delta$  tree frame-weighted average entropies by phoneme. Phonemes are ranked by the magnitude of entropy reduction caused by EDT. The entropy of “ey” shrinks most, while that of epenthetic silence “epi” actually increases. Note that the 11 phonemes whose entropy shrinks the most are all diphthongs and vowels, while the 18 whose entropy shrinks the least are all consonants. Thus, diphthongs and vowels are probably more speaker-dependent than consonants. In the SI model, consonants typically have higher entropy than vowels. This does **not** mean that consonants vary more than vowels; it means that they vary more than vowels after context-dependent effects have been taken into account for both.

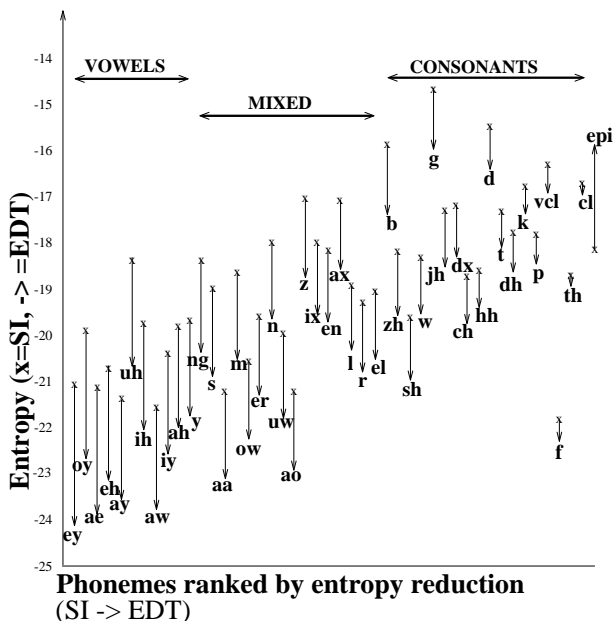


Figure 4 Entropy reduction due to EDT (by phoneme)

## 5. Discussion

EDT outperforms standard adaptation methods on small amounts of speech. Compared to an SI baseline, on three adaptation sentences it yields 12.1% ERR for supervised adaptation and 11.2% ERR for unsupervised adaptation; for self-adaptation it yields 10.4% ERR on a single sentence. Memory costs are comparable to those for SI (EDT models seem to need fewer Gaussians than SI models, compensating for the extra parameters needed for the eigenspace).

We have begun experiments on Wall Street Journal to see how well EDT works on a large-vocabulary task. We also hope to find out why EDT works at all. It is based on the very naïve assumption that an allophone can be modeled as the sum of a SDCI and a SICD component, implying that there is no interaction between speaker type and phoneme context. In informal TIMIT experiments, we relaxed this additive assumption by allowing the  $\delta$  trees to contain questions about the speaker’s location in eigenspace. To our surprise, trees containing such “eigenquestions” performed no better than the standard  $\delta$  trees. This means that the “eigenquestions” are too crude, that TIMIT does not have enough data per speaker for speaker-dependent allophone effects to be spotted, or – strangest of all – that the additive assumption is correct. We are planning Wall Street Journal experiments that will help to clarify this issue.

## 6. References

- [1] A. Acero and X. Huang. “Speaker and gender normalization for continuous-density HMMs”. *ICASSP-96*, V. 1, pp. 342-345, Atlanta, May 1996.
- [2] L. Baum. “An inequality and associated maximization technique”. *Inequalities*, V. 3, pp. 1-8, 1972.
- [3] H. Botterweck. “Very Fast Adaptation for Large Vocabulary Continuous Speech Recognition Using Eigenvoices”. *ICSLP-2000*, V. IV, pp. 354-357, Beijing, Oct. 2000.
- [4] T. Cover and J. Thomas. “Elements of Information Theory”. John Wiley & Sons, 1991.
- [5] M. Gales. “Cluster Adaptive Training of HMMs”. *IEEE Trans. SAP*, V. 8, no. 4, pp. 417-428, July 2000.
- [6] T. Hazen and J. Glass. “A Comparison of Novel Techniques for Instantaneous Speaker Adaptation”. *Eurospeech '97*, V. 4, pp 2047-2050, Rhodes, Sept. 1997.
- [7] Q. Huo and B. Ma. “Irrelevant variability normalization in learning HMM state tying from data”. *ICASSP-99*, V. 2, pp. 577-580, Phoenix, Mar. 1999.
- [8] S. Kajarekar, N. Malayalath, and H. Hermansky, “Analysis of Speaker and Channel Variability in Speech”. *ASRU-99*, Keystone, Colorado, Dec. 12-15, 1999.
- [9] R. Kuhn, F. Perronin, *et. al.*, “Very Fast Adaptation with a Compact Context-Dependent Eigenvoice Model”. *ICASSP-2001*, Salt Lake City, May 2001.
- [10] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. “Rapid Speaker Adaptation in Eigenvoice Space”. *IEEE Trans. SAP*, V. 8, no. 6, pp. 695-707, Nov. 2000.
- [11] P. Nguyen, C. Wellekens and J.-C. Junqua. “Maximum Likelihood Eigenspace and MLLR for Speech Recognition in Noisy Environments”. *Eurospeech-99*, V. 6, pp. 2519-2522, Budapest, Hungary, 1999.
- [12] S. Young, J. Odell, and P. Woodland. “Tree-Based State Tying for High Accuracy Acoustic Modeling”. *ARPA HLT Workshop*, pp. 286-291, New Jersey, March 1994.