

VERY FAST ADAPTATION WITH A COMPACT CONTEXT-DEPENDENT EIGENVOICE MODEL

R. Kuhn, F. Perronnin, P. Nguyen, J.-C. Junqua, and L. Rigazio

Panasonic Speech Technology Laboratory, Panasonic Technologies Inc.
Santa Barbara, California, USA
(kuhn, florent, nguyen, jcj, rigazio@research.panasonic.com)

1. ABSTRACT

The “eigenvoice” technique achieves rapid speaker adaptation by employing prior knowledge of speaker space obtained from reference speakers to place strong constraints on the initial model for each new speaker [9, 10]. It has recently been shown to yield very fast adaptation for a large-vocabulary system [3] ([5] modifies the technique in an interesting way). In this paper, we describe a new way of applying the eigenvoice technique to context-dependent acoustic modeling, called the “Eigencentroid plus Delta Trees” (EDT) model. Here, the context-dependent model is defined so that it consists of a speaker-dependent component with a small number of parameters linked to a speaker-independent component with far more parameters. The eigenvoice technique can then be applied to the speaker-dependent component alone to attain very fast adaptation of the entire context-dependent model (*e.g.*, 10% relative reduction in error rate after 3 sentences). EDT requires only a small number of parameters to represent speaker space and works even if only a small amount of data is available per reference speaker (in contrast to the system described in [3]).

2. BACKGROUND

2.1. Work on Bipartite Acoustic Models

To our knowledge, Acero and Huang were the first to propose what could be called a “bipartite acoustic model” for speech recognition [1]. Such models have two components, one (with a small number of parameters) which models the speaker-dependent and environment-dependent part of the acoustic model, and the other (with a larger number of parameters) which deals with the residual speaker-independent, environment-independent, part of the model. In the Acero-Huang scheme for continuous-density HMMs, each mixture Gaussian mean was modeled as the sum of a speaker-cluster-dependent, context-independent (CI) mean vector μ and a speaker-independent (SI), context-dependent (CD) offset δ . Acero and Huang showed that this scheme could support gender normalization and batch-mode speaker adaptation, yielding good results in a test on Wall Street Journal (WSJ). For batch-mode speaker adaptation, an SI recognizer was used in a first pass over the 41 test utterances from a particular speaker to train a speaker-dependent (SD) μ . In the second pass, this μ was combined with SI δ s estimated from the training speakers to carry out recognition. Compared to the SI baseline, the second pass yielded a relative error rate reduction of 30%.

Recently, Bocchieri proposed that allophones be modeled as CD linear transforms of speaker- and environment-dependent CI models [2]. This model can be viewed as a generalization of Acero and Huang’s, and yields good results when applied to environment adaptation. In his experiments, Bocchieri used a minimum of 300 sentences to estimate the CI component of the model.

Bipartite acoustic models such as those just described are potentially very powerful. Only one element is missing: a fast method for estimating CI models for each new speaker and environment. We show here that this can be achieved by the eigenvoice approach.

2.2. Eigenvoices

In the eigenvoice approach, one uses a dimensionality reduction technique to infer strong *a priori* knowledge about speaker space from a set of reference speakers. For instance, by applying Principal Component Analysis (PCA) to a set of SD vectors, each representing the concatenated Gaussian means for a given speaker’s model, one can obtain a low-dimensional eigenspace to which the model for each new speaker is confined. This constraint is so powerful that the model for the new speaker can be estimated accurately on very small amounts of adaptation data by means of a maximum likelihood technique called MLED (see [9], which describes isolated-word experiments). Nguyen *et al.* outline a maximum-likelihood method called MLES for re-estimating the eigenspace given by PCA (or by some other method), and show that this yields significantly better performance than PCA alone for CI recognition [10]. Techniques related to the eigenvoice approach include M. Gales’s “soft clustering” [6] and especially T. Hazen’s and J. Glass’s “reference speaker weighting” [7].

Recently, Botterweck applied the eigenvoice approach to large-vocabulary recognition for the 34K WSJ task, showing that it yielded 14.8% relative error rate reduction for only 3 sec. of adaptation data [3]. He also showed that the eigenvoice approach performs better than MLLR for less than 165 sec. of adaptation data. To obtain these results, Botterweck carried out PCA and MLES on complete CD models derived from 300 reference speakers (200 of whom supplied 15 min. of speech each, while the remaining 100 supplied an hour of speech each). Each eigenvoice vector had approximately one million parameters, so that the eigenspace of (*e.g.*) dimensionality 100 involved 100 million parameters.

The “Eigencentroid plus Delta Trees” (EDT) model described here yields rapid adaptation for a CD system, requires few parameters to model eigenspace, and works reasonably well even if the amount of speech data per reference speaker is small.

3. EIGENCENTROID PLUS DELTA TREES (EDT)

3.1. The structure of EDT

EDT is a variant of the Acero-Huang model in which the CI component is located in eigenspace. Consider the mean m_i of Gaussian i in the mixture for state s of phoneme p in a particular allophone context a , when the speaker is S and the environment E . Let $m_i(S, E, p, a, s, i) = \mu(S, E, p, s) + \delta(p, a, s, i)$, where $\mu(S, E, p, s)$ is the portion of a CI “eigencentroid” vector pertaining to state s of phoneme p . By definition, μ is located in an eigenspace trained on reference speakers; each point in the eigenspace represents a possible CI model. Given adaptation data from speaker S in environment E , one can use MLED to estimate μ [9]. To obtain the full CD model, one adds the appropriate part of μ to the means of the relevant CD distributions in the SI δ portion of the model. In our implementation, the δ distributions are found in the leaves of a set of decision trees, one per phoneme (note that the symbol δ here has nothing to do with delta acoustic features). The questions in the trees pertain to phonetic context and to state.

Eigencentroid

Delta Trees

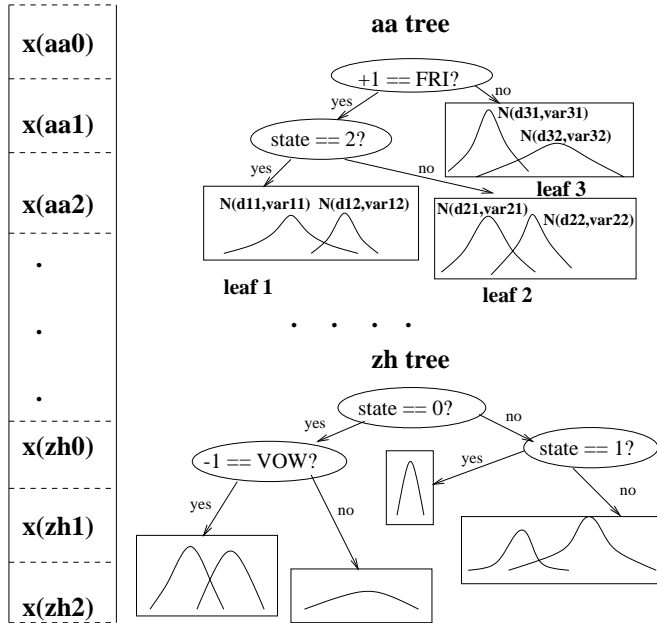


Fig. 1. Eigencentroid and δ trees

Figure 1 shows an eigencentroid vector and δ trees (the trees in the figure are invented). The δ trees are trained offline, as is the eigenspace in which the eigencentroid lies; the eigencentroid itself is estimated for each new speaker on adaptation data. For each phoneme from aa to zh , each state is associated with a subvector $\mathbf{x}()$ of the eigencentroid that has the same dimensionality as the number of acoustic features. For instance, $\mathbf{x}(aa0)$ is a rough estimate of the mean feature vector for state 0 of phoneme aa . This model is very economical of parameters compared to the case where each eigenvoice vector represents a full CD model.

For instance, our experiments involved 139 phoneme states and an 18-dimensional acoustic feature vector, implying a total of 2502 Gaussian mean parameters per eigenvoice or eigencentroid (as compared with about a million in [3]).

Suppose that we wish to find the CD model for state 0 of aa before a fricative. The first question in the aa δ tree means “does the aa precede a fricative?” The answer is ‘yes’, so we proceed to the question “is the state equal to 2?” Since the answer is ‘no’, we end up in leaf 2 of the aa δ tree. There are two Gaussians here - one with mean $d21$ and variance $var21$, the other with mean $d22$ and variance $var22$ (the mixture weights are also stored here). The estimated distributions for state 0 thus have mean $x(aa0) + d21$ with variance $var21$, and mean $x(aa0) + d22$ with variance $var22$ respectively (and the same mixture weights). Although leaf 2 is shared by states 0 and 1 of aa , the Gaussians for the same allophone of the two states are **not** the same; the Gaussians for state 1 of aa preceding a fricative have mean $x(aa1) + d21$ with variance $var21$ and mean $x(aa1) + d22$ with variance $var22$.

3.2. Training and Using EDT models

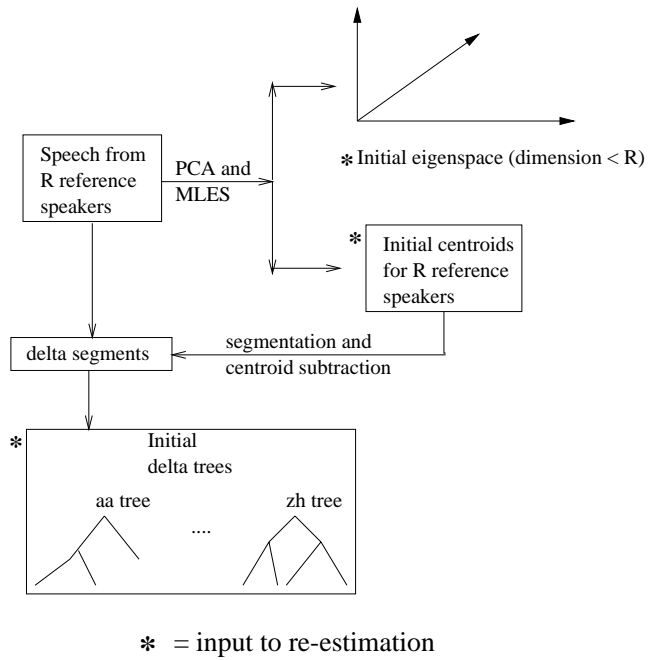


Fig. 2. EDT Initialization

Figure 2 summarizes the initialization step for EDT. There are three outputs from this step: **1.** An initial eigenspace; **2.** Initial centroid locations for each of the R reference speakers in the eigenspace; **3.** Initial δ trees.

The first two outputs are obtained by applying the techniques described in [9, 10]. To produce the δ trees, the data for each of the reference speakers are segmented and the relevant portion of each speaker’s centroid is subtracted from each segment. For instance, a segment labeled as belonging to state 0 of aa for speaker i will have $x(aa0)$ from speaker i ’s estimated eigencentroid subtracted from it (*i.e.*, from each of its frames). The δ segments from the

same phoneme are pooled across speakers and used to grow the δ trees (see [8] for a similar approach).

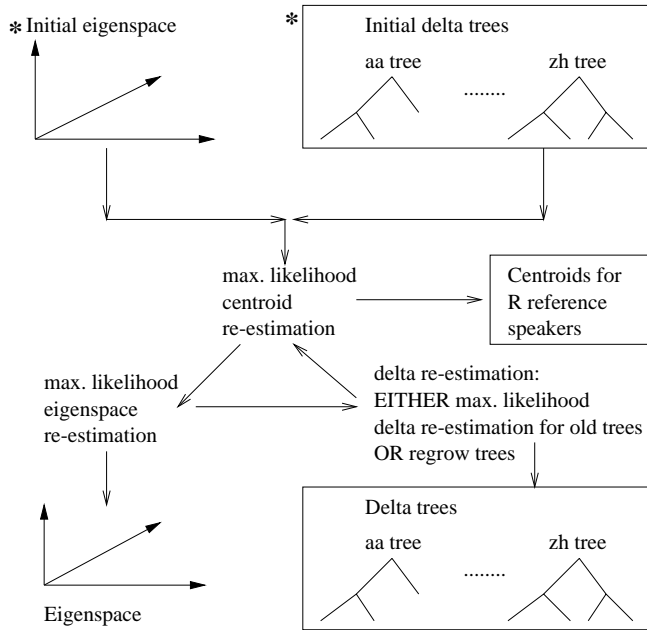


Fig. 3. EDT Re-estimation

Figure 3 shows the iterative re-estimation procedure. We have derived three re-estimation formulas (to be given in a later paper): one re-estimates reference speaker eigencentroids, one re-estimates the eigenspace, and one re-estimates the δ s in a set of pre-existing phoneme trees. Instead of using the δ re-estimation formula, one may choose to regrow the trees completely on δ segments obtained by subtracting each speaker’s eigencentroid from segmented data (as for initialization).

At runtime, the initial EDT model for the new speaker has all eigencoordinates set to zero; two or three iterations of maximum-likelihood re-estimation are carried out on the adaptation data to calculate the eigencoordinates for the new speaker.

4. EXPERIMENTS

4.1. Configuration

Phoneme recognition experiments were conducted on the TIMIT database, using the standard train/test partition with 462 speakers in the training set and 168 in the test set. Each speaker pronounces 8 sentences, with an average sentence length of 4 sec. (including silence). Speech was sampled at 16 kHz and parametrized using PLP cepstral features. There are 18 acoustic features: 9 static coefficients, including residual energy, and 9 delta features. For training, there are 46 phonemes with 3-state HMMs, plus ‘closure’ with a 1-state HMM, all of which were adapted using EDT; scoring was done using a reduced 39-phoneme set. The 1-state silence model with 28 Gaussians was not adapted. The recognizer employed a bigram backoff language model.

We carried out preliminary experiments in which the eigenvoice approach was applied to full CD models (as in [3]). Initializing with the SI model with one Gaussian per leaf, we carried out maximum likelihood (ML) re-estimation on the 8 sentences from each training speaker to obtain 462 CD models with one Gaussian per leaf. Six MLES iterations were carried out to estimate each eigenspace.

To initialize the EDT system, we grew one-Gaussian-per-leaf δ trees. We carried out three iterations of re-estimation with one-Gaussian trees with the same structure (each iteration re-estimated the eigencentroids, eigenspace, and δ s) and then grew a completely new set of trees with the desired final maximum number of Gaussians per leaf (using a splitting procedure). There are many ways of applying the EDT re-estimation procedure; this way may not be optimal.

4.2. Supervised Adaptation Results

The SI and CD baseline with one Gaussian per leaf yielded 65.1% unit accuracy. Preliminary experiments with eigenspaces trained on full CD models yielded no significant improvement over the baseline: the results were 65.2% for 5 eigenvoices, 65.3% for 10 eigenvoices, and 65.2% for 20 eigenvoices. Although we might have obtained slightly better results by using a different method for training the CD models, we decided not to continue experiments along these lines; estimating good CD models with only 8 sentences per speaker is problematic.

Figures 4 and 5 show recognition results for EDT adaptation on one and three TIMIT sentences, with testing being done on the remaining seven and five sentences for that speaker respectively. The maximum number of Gaussians per leaf for each set of δ trees was set to 1, 2, 4, 8, 16, and 32; the actual number is data-dependent.

The experimental results with 20 eigenvoices for one adaptation sentence are not shown in Figure 4 because they are roughly the same as or slightly worse than those for 10 eigenvoices. For three sentences, Figure 5 seems to indicate that 20 eigenvoices yield the best results. The relative error rate reduction (ERR) for 10 eigenvoices on one adaptation sentence varies from 10.0% to 6.2%; the ERR for 20 eigenvoices on three adaptation sentences varies from 11.9% to 10.0%. All eigenvoice results in these figures were obtained with re-estimation (Figure 3). When the initial eigencentroid and δ trees were used for experiments, results were slightly worse (from 0.5% to 1.0% higher absolute error rate).

For a fixed number of Gaussians, the EDT systems always perform better than the SI-CD baseline. Furthermore, the best SI system is always outperformed by an EDT system with far fewer Gaussians - for instance, in Figure 5 an SI system with 13,702 Gaussians yields 70.0% accuracy, but the 10-eigenvoice system with 2281 Gaussians yields 70.7% accuracy. It might be argued that this comparison is unfair, since it does not include the parameters needed for eigencentroid estimation (*i.e.*, the eigenvoices). Each eigenvoice corresponds to a CI model with 139 Gaussians, so one could argue that the 10-eigenvoice system corresponds to a model with 3671 Gaussians. Even in this pessimistic accounting, however, EDT comes out ahead.

In Figures 4 and 5, “MLLR=>MAP” denotes MLLR adaptation followed by MAP adaptation of the SI-CD baseline. This technique performs worse than the baseline for one adaptation sentence. For three adaptation sentences, results for “MLLR=>MAP” lie between those of 1-eigenvoice EDT and 5-eigenvoice EDT.

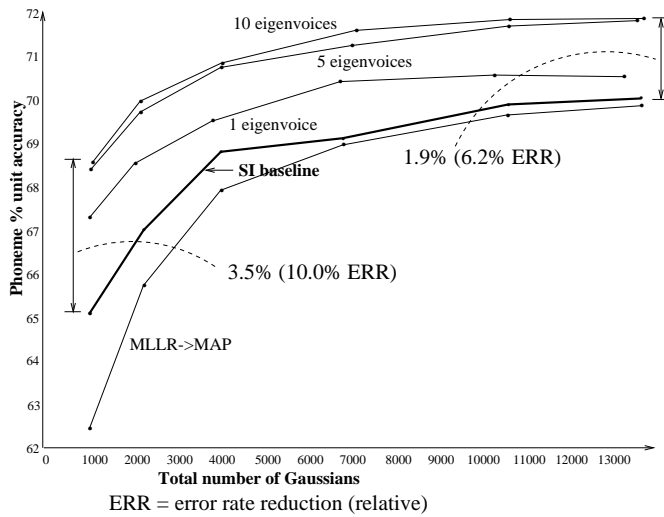


Fig. 4. EDT Adaptation on 1 TIMIT sentence

Interestingly, EDT seems to enhance pooling of data across states. For instance, only 1% of the 1147 SI leaves for the above experiments were shared by two or three states, but 15% of the 1157 leaves in the 10-eigenvoice δ tree were shared.

5. DISCUSSION AND FUTURE WORK

We presented the EDT context-dependent model and showed that it yields rapid adaptation. As compared with the original eigenvoice approach [9, 10] applied directly to context-dependent modeling [3], the advantage of EDT is that only a few parameters are allocated to the eigenvoice portion of the model; thus, the eigenspace is also small, and can be estimated from a small amount of data per reference speaker. The experiments described above were a tough test of EDT, since the eigenspace was estimated on only about 30 sec. of speech per reference speaker (vs. at least 15 minutes per speaker in [3]). Nevertheless, performance was good.

Further improvements might be obtained by considering two sources of error in the EDT model. First, the best possible estimate for a given speaker's CI (centroid) model may lie outside rather than within the eigenspace obtained from reference speakers. This problem could be handled by applying MLLR or MAP to the eigencentroid (once enough adaptation data are available).

Second, some CD offsets from SD phoneme means may **not** be independent of speaker type, contrary to the Acero-Huang assumption. To handle this, we have implemented δ trees that contain questions about the current speaker's eigenspace coordinates - i.e., questions of the form "is dimension $i < k$?" where k is some empirically derived constant. Interesting values of k for each dimension are determined from histograms of reference speaker eigencoordinates. Results so far show little or no improvement over δ trees with no knowledge of speaker coordinates, perhaps because interaction effects between allophones and speaker type are poorly predicted by centroid information. "The tree of knowledge is not that of Life" [4].

Future work will focus on experiments with more data per reference speaker and on incorporation into an EDT-like framework

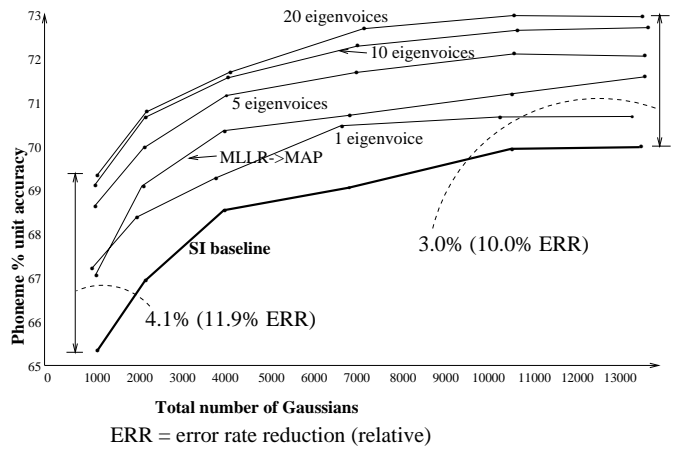


Fig. 5. EDT Adaptation on 3 TIMIT sentences

of a bipartite model that makes more reasonable assumptions (such as Bocchieri's).

6. REFERENCES

1. A. Acero and X. Huang. "Speaker and gender normalization for continuous-density Hidden Markov Models". *ICASSP-96*, V. 1, pp. 342-345, Atlanta, USA, May 1996.
2. E. Bocchieri. "Phonetic Context Dependency Modeling by Transform". *ICSLP-2000*, V. IV, pp. 179-182, Beijing, China, Oct. 2000.
3. H. Botterweck. "Very Fast Adaptation for Large Vocabulary Continuous Speech Recognition Using Eigenvoices". *ICSLP-2000*, V. IV, pp. 354-357, Beijing, China, Oct. 2000.
4. Lord Byron. "Manfred", act 2, scene 2.
5. K.-T. Chen, W.-W. Liao, H.-M. Wang, and L.-S. Lee. "Fast Speaker Adaptation Using Eigenspace-Based Maximum Likelihood Linear Regression". *ICSLP-2000*, V. III, pp. 742-745, Beijing, China, Oct. 2000.
6. M. J. F. Gales. "Cluster Adaptive Training of Hidden Markov Models". *IEEE Trans. SAP*, V. 8, no. 4, pp. 417-428, July 2000.
7. T. Hazen and J. Glass. "A Comparison of Novel Techniques for Instantaneous Speaker Adaptation", *Eurospeech '97*, V. 4, pp 2047-2050, Rhodes, Greece, Sept. 1997.
8. Q. Huo and B. Ma. "Irrelevant variability normalization in learning HMM state tying from data based on phonetic decision tree", *ICASSP-99*, V. 2, pp. 577-580, Phoenix, Arizona, March 1999.
9. R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski. "Rapid Speaker Adaptation in Eigenvoice Space". *IEEE Trans. Speech Audio Proc.*, V. 8, no. 6, pp. 695-707, Nov. 2000.
10. 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.