

A Review Of Eigenvoice Adaptation*

Peter Smit
Adaptive Informatics Research Centre
Aalto University
peter.smit@tkk.fi

November 29, 2010

Abstract

Eigenvoices is a principle of creating a low-dimensional space for efficiently describing different voices. This paper describes the most common techniques used like Projection and MLED, and extensions to these techniques like Kernel eigenvoices. Also combinations with other adaptations techniques, especially MAP and MLLR are discussed and comparison between these three is made.

The conclusion is made that eigenvoices are suitable in situations where only a small amount of data is available, or in combination with other adaptation techniques. Also eigenvoices can be used as basis for other applications like speech synthesis and speech separation.

1 Introduction

In Automatic Speech Recognition (ASR) there can be made a distinction between two different type of systems. Speaker Dependent (SD) systems are tailored for recognizing one specific speaker, and are therefore normally trained with only data of a specific speaker. Speaker Independent (SI) systems are made for giving a reasonable performance for every speaker, even if no data of that speaker is used in the training of the model. A SI system is trained with a huge variety of data from different speakers to achieve this general recognition capability.

In general, SI systems are giving a good performance for any speaker, but not as well as a SD system gives for a specific speaker. On the other hand, because a SD system needs to have a huge amount of data for every speaker it is going to recognize, it is often not practical and much more expensive in training and data collection than a single SI-system.

Adaptation techniques bring together the best of SI and SD models. An SI model is trained with data from different speakers and at recognition time,

*This review is written for the course T-61.6090 Special Course in Language Technology; Adaptation in speech and language processing

a small amount of speaker-specific data is used to transform the SI model into a SD model. Because most adaptation methods only need a little amount of data, sometimes only a few sentences are needed, they can be used with less effort and resources than SD models.

Generally there are three different classes of adaptation techniques as described in Woodland [2001]; linear regression (MLLR, Gales [1998]), maximum a posteriori adaptation (MAP) and clustering or eigenvoice techniques. This review will focus on the last class; the eigenvoice techniques. As shown, this technique is often not used by itself but in combination with adaptation techniques from the MLLR or MAP class.

In this paper first eigenvoices will be described and multiple different methods introduced. Also methods that combine eigenvoices with other adaptation techniques will be described and compared. At last other applications than speech recognition will be highlighted.

2 Eigenvoices

Eigenvoice adaptation was first proposed in Kuhn et al. [1998] and is inspired by the eigenfaces technique used in face recognition applications.

The general idea of eigenvoices is to create a low dimensional space that describes speaker variability. Ideally this space should have as few parameters as possible and still be able to cover all speakers, not only the speakers in the training data but also speakers that have not been observed before. Every point in space should represent to a speaker and a mapping should exist to the parameters of a speaker-dependent model.

A dimensionality reduction technique (DRT) is used to reduce the amount of parameters of the speech model. For example Principal Component Analyses can be used and a small number of the first eigenvectors can be used to define the eigenspace. Any DRT would work, but only results on PCA are reported.

When the eigenspace is defined, a method should be found to map a small amount of new speech to a point in eigenspace which can be used to build an adapted model. This process is called *Eigenvoice Decomposition*.

A lot of different methods of defining an eigenspace and calculating eigenvoices have been proposed. The following sections discuss different techniques and their differences.

2.1 PCA on the parameters of Speaker Dependent models

The original paper describing eigenvoices [Kuhn et al., 1998] uses the parameters of Speaker-Dependent models to define an eigenspace with Principal Component Analysis (PCA).

The training procedure starts with the training of T full SD models, all with same structure and an equal number of parameters. For every model

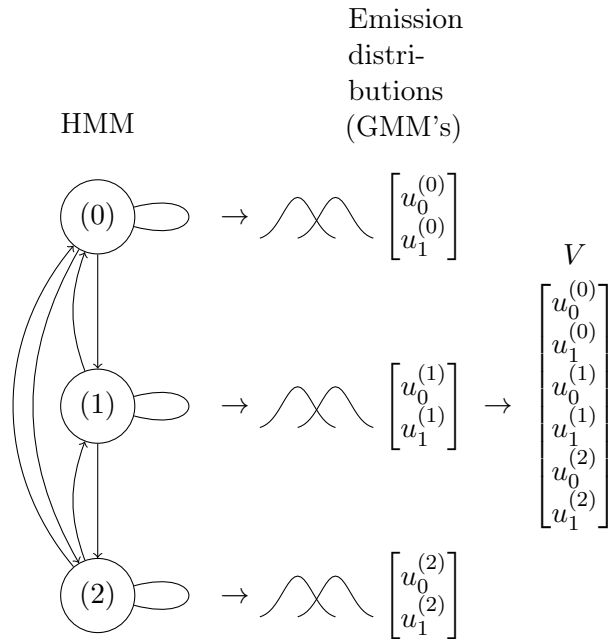


Figure 1: Construction of vector V from a SD model

a supervector V is constructed from its mean parameters as also is shown in Figure 1. The order of the elements of this vector is not important, however it should be equal for all created vectors.

The resulting T vectors of size D (the number of mean parameters in a SD model) are reduced with Principal component Analyses (PCA). The first K principal components are selected (with $K \ll D$) resulting in K orthogonal base vectors (eigenvectors) of size D . Every model can now be described by K parameters, a linear combination of the K base vectors.

In formula's, when the data matrix is \mathbf{T} , than it can the eigenvoice parameters \mathbf{Y} can be written as:

$$\mathbf{Y} = \mathbf{W}^T * \mathbf{T}$$

where \mathbf{W} are the eigenvectors (or eigenvoices).

PCA can be both applied on the covariance or the correlation matrix. Important is that when the data is mean-centered, the mean is stored. This is necessary for reconstruction of a SD-model.

For the recognition procedure there are two basic options, projection and Maximum Likelihood Eigenspace Decomposition. Both are described in Kuhn et al. [1998].

Note that because only mean parameters are used here, only these parameters will be adapted. Other parameters like covariance matrices and transition probabilities are copied from the Speaker-Independent models.

2.1.1 Projection

Projection is the most basic technique for using eigenvoices. For a new speaker, a SD model is trained with the available adaptation data. From the model, the vector V is created, in the same as was done in the training procedure. Now a more robust SD model is obtained by projecting this vector with PCA. The resulting parameter vector P is obtained by the linear equation: $P = E \times E^T \times V$ where E is the matrix with the eigenvoices. Other parameters like the transition probabilities of the HMM and variances are taken from a separately trained SI model.

The disadvantage of this approach is that a SD model must be trained of the adaptation data, which means that all states of the model must be present in the adaptation data. Especially with little adaptation data or big models this often poses a problem. Therefore this technique is not commonly used.

To overcome the problems of projection, Westwood [1999] proposes weighted projection. Instead of operating on the mean supervectors, it uses a linear transformed supervector. This transform is designed in such way that a parameters influence in the vector is weighted by it's variance and occurrence count. To be exactly, V is replaced by $V' = \Omega DV$ where Ω is a diagonal matrix with weights (or occupancy count) and D a block-diagonal matrices for the covariances within the Gaussian Mixture Distributions.

Because the new vector V' will not contain any unestimated parameters, projections will always work. The new supervector for an adapted model can be obtained by calculating the inverse transformation. Westwood [1999] discusses multiple options for the exact weight matrix and functions to obtain the best results.

2.1.2 Maximum Likelihood Eigen-Decomposition

Maximum Likelihood Eigen-Decomposition (MLED) [Kuhn et al., 1998, 2000] uses the Maximum Likelihood to estimate the weight for the new SD model.

The auxiliary function $Q(\lambda, \hat{\lambda})$ is defined by

$$Q(\lambda, \hat{\lambda}) = -\frac{1}{2}P(O | \lambda) \sum_s \sum_t \gamma_s(t) \mathbf{f}(\mathbf{o}_t, s)$$

with

$$\mathbf{f}(\mathbf{o}_t, s) = (n \log(2\pi) + \log |C_s| + \mathbf{h}(\mathbf{o}_t, s))$$

and

$$\mathbf{h}(\mathbf{o}_t, s) = (\mathbf{o}_t - \hat{\mu}_s)^T C_s^{-1} (\mathbf{o}_t - \hat{\mu}_s)$$

where s is a state, n is the number of features, \mathbf{o}_t the observation vector at time t , $C_m^{(s)-1}$ the inverse covariance in state s , $\hat{\mu}_s$ the new adapted mean

of state s and $\gamma_s(t)$ the current likelihood ($L(s|\lambda, \mathbf{o}_t)$). This likelihood is in fact the state alignment, resulting in a 1/0 output.

The K weights can be solved with $\frac{\partial Q}{\partial w(j)} = 0$. The resulting equation which has to be solved is

$$\sum_s \sum_t \gamma_s(t) (e_s(j))^T C_s^{-1} \mathbf{o}_t = \sum_s \sum_t \gamma_s(t) \left(\sum_{k=1}^K w(k) (e_s(k))^T C_s^{-1} e_s(j) \right), j = 1 \dots K$$

Solving this equation is not a computationally hard problem, therefore this method is feasible to use in recognition. It even requires less computation than CMLLR. (This method requires one matrix inverse for solving the weights, CMLLR requires n inverses where n is the dimension of the feature vector). It becomes more clear when it is written in vector formulation:

$$\sum_t \sum_s \gamma_s(t) \mathbf{e}_s^T C_s^{-1} \mathbf{o}_t = \mathbf{w}^T \sum_t \sum_s \gamma_s(t) \mathbf{e}_s^T C_s^{-1} \mathbf{e}_s$$

Because $\gamma(t)$ is a zero/one function with only 1 for one state, only statistics have to be collected for all frames and the function $A = \mathbf{w}^T B$ has to be solved for the unknown \mathbf{w} .

The results can be further defined by applying this method iteratively. After the first iteration, the new weights can be used to construct a SD model that will give a new state-alignment. A new \mathbf{w} can now be calculated. This procedure can be repeated until \mathbf{w} converges.

In contrast to Projection, MLED does not need to fully train a SD model from the adaptation data. Hence the adaptation data does not have to cover every Gaussian present in the model.

MLED is often used as baseline for comparison with other eigenvoice techniques.

2.2 Maximum Likelihood Eigenspace (MLES)

Instead of PCA as used in the LSES procedure of creating an eigenspace, also other methods can be used. Nguyen et al. [1999] proposes the Maximum Likelihood Eigenspace (MLES) which is estimated with a Baum-Welch like procedure, utilizing prior information like sex or dialect.

MLES integrates nicely in to the standard Baum-Welch model training procedure. In the maximum likelihood formula, a hidden speaker-dependent parameter is added for the weight of the eigenvoices. The training procedure itself only requires to keep an accumulator for each speaker instead of one global accumulator. The formula in the training procedure becomes

$$\hat{M} = \arg \max_M \sum_{q=1}^T \int \log L(O, w|M) P_0(w, q) dw$$

where $P_0(w, q)$ contains the prior information.

By giving a prior for the eigenvoice parameter in the ML formula, the results could be improved even more, a method called MAPES.

Except for the smaller computational requirements, MLES does not seem to outperform the standard PCA method. The results published by Nguyen et al. [1999] indicated a close, but worse performance (60% vs 62% error rate).

2.3 Maximum A Posteriori Linear Regression (MAPLR)

Chen and Wang [2001] proposes a method that uses eigenvoices as a prior for Maximum A Posteriori (MAP) adaptation.

MAPLR differs in two ways of the standard eigenvoices technique. Firstly, the eigenspace is not created with normal PCA but Probabilistic PCA, an extension to the normal PCA framework which also include an explicit model for noise. Secondly, the eigenvoices are not used directly for making a Speaker Dependent model, but as prior for MAP adaptation.

The results are similar to other methods, giving an improvement in accuracy from 50% to 52% compared with normal MLED adaptation.

A related approach is the Bayesian speaker adaptation [Kim and Kim, 2000] which uses also PPCA and the eigenvoices as a prior for the transformation estimation.

2.4 Properties of eigenvoices

Because eigenvoices take into account a priori information about the space of models, it requires less parameters, and therefore less adaptation data, than other adaptation methods like MLLR and MAP. As shown in Kuhn et al. [2000] this makes them perform much better when only a small amount of adaptation data is available.

There is also a downside. For creating the eigenspace, full SD models must be completely trained. This is possible for digit or small vocabulary recognizers, but for large vocabulary recognizers there is often not enough training data available for training such models. Also for bigger models, the computational requirements of PCA can become a bottleneck, even though this only has to be done at model training.

Besides the need of training SD models and computational costs, eigenvoice adaptation on itself seems not to outperform MLLR and MAP techniques for bigger amounts of adaptation data. To reduce the need for training SD models, MLLR eigenvoices can be used as described in the next section.

3 MLLR eigenvoices

Eigenvoices is a technique that is easily pooled with different adaptation methods. Especially Maximum Likelihood Linear Regression [Gales, 1998] (MLLR) can be utilized in both the estimation of eigenvoices as it can utilize eigenvoices for performing a more accurate transformation.

3.1 MLLR based eigenvoice training

Instead of training speaker dependent models for each speaker, eigenvoices can also be estimated by using models adapted by another method. For bigger models this prevents a lot of problems as parameters can be still mapped easily to a supervector. With gaussian mixture models for example, it is hard to define an order. When MLLR transforms are used, the mixture parts can be all matched back to the independent model and therefore easily ordered in a supervector.

MLLR eigenvoices removes the need for training any Speaker Dependent models. Also, less speaker-specific data is needed as a MLLR transform needs less data than the training of a SD model.

Results for applying MLLR eigenvoices for large vocabulary Speech Recognition can be found in Botterweck [2000], Chen and Wang [2001]. Up to 15% relative improvement could be achieved, depending on the amount of adaptation data.

Wang et al. [2001] also shows that the eigenvoice technique can be applied directly on the MLLR parameters instead of the with MLLR adapted model parameters. This technique makes the eigenvoice approach independent from the model, even enabling to use the eigenvoices cross-model. Also the computational requirements are less, because the number of MLLR parameters is lower than the number of model parameters.

3.2 Using eigenvoices for improving MLLR adaptation

Also it is possible to use eigenvoices to refine other adaptation techniques. Chen et al. [2000] proposes a method which uses eigenvoices to establish boundaries for the full MLLR transform by reducing the free parameters and so requiring less adaptation data. A parameter smoothing technique was used. The experiments showed that for amounts as little as 10 seconds the method outperformed conventional MLLR.

4 Kernel Eigenvoices

Commonly, normal linear PCA is used for the estimation of the eigenvoice space. More recent studies like Mak et al. [2004], Mak and Hsiao [2004],

Mak et al. [2005] have shown that non-linear method can give significant improvements.

Mak and Hsiao [2004] proposes to use Kernel PCA instead of PCA as Dimensionality Reduction technique. The difference between normal PCA and Kernel PCA is that instead of a direct linear mapping, first a non-linear kernel function is applied to transform the vectors into a “kernel space”. Of course there is a huge choice of kernel functions that could be used and the right one needs to be selected.

With the linearity of PCA is not a problem to use new weights for construction a SD model. However, with kernel functions, care must be applied that the used kernel function can be inverted. To meet this requirement composite kernels can be used.

Mathematically, every element in the mean vectors are mapped to a kernel-space \mathcal{F} . So vector \mathbf{y} is transformed to $\{\varphi(\mathbf{y}_1) \dots \varphi(\mathbf{y}_N)\}$. Now instead of on the covariance or correlation matrix, eigenvoice decomposition is applied to the Kernel Matrix \mathbf{K} with $\mathbf{K}_{ij} = \varphi(\mathbf{Y}_i)^T \varphi(\mathbf{Y}_j)$

For estimating eigenvoice coefficients, a Maximum Likelihood function is defined and the evaluation happens with Gradient Descent. Obviously it costs more computational resources than techniques with an exact solution like MLED for normal eigenvoices.

As mentioned, composite kernels can be used, as they are invertible, hence making it possible to construct mean parameters when the coefficients are found. Mak et al. [2004] describes two options, the direct sum kernel and the tensor product kernel. Both were shown to have similar performance.

Kernel Eigenvoices can be very effective, giving up to 27.5 % accuracy improvement for a digit recognition task [Mak et al., 2005], compared to a Speaker Independent model.

5 Comparison with other adaptation methods

From the three classes of adaptation methods mentioned in Woodland [2001], eigenvoice adaptation is the method that has the least free parameters and performs the best for very small amounts of adaptation data. In these respect, MLLR needs more and MAP much more adaptation data to estimate robust adaptations.

The main difference causing this is the utilization of prior knowledge by the eigenvoices technique. Because the training dataset is used for establishing boundaries on the speaker space enables eigenvoices to only require a very small number of parameters.

MLLR has the power of not needing any external or prior knowledge in order for it to be effective. Transformations are only based on the difference between the Speaker Independent model and the adaptation data itself. The method doesn't have even parameters that need to be tuned, except for the

number of different Regression Classes.

Even though the other techniques quickly outperform eigenvoices when more data comes available, eigenvoices have the right to exist by their ability to be combined with the other adaptation techniques. For example it can be very well used to estimate a prior for MAP as was already done in the original paper [Kuhn et al., 1998].

6 Other applications of eigenvoices

Besides for improving basic Automatic Speech recognition, eigenvoices are also used in other situations. This section describes two of such applications

6.1 Speech Separation

In Weiss and Ellis [2010], eigenvoices are used for creating Speaker Dependent models. However, instead of recognition, these models are used for separating multiple speech sources.

Because a speaker can be expressed in a small number of parameters, eigenvoices were found useful even when a speech signal was not purely from one speaker, allowing for iteratively enhancing the separation procedure until the speech signals were well enough separated.

The reason for using eigenvoices was the small amount of data needed for estimating the adaptation and the robustness of the method, due to the small number of parameters. The method was extended with a gain parameter to account for different signal-to-noise ratio's in the multi-speaker samples.

6.2 Speech Synthesis

Eigenvoices have also proven useful in Text to Speech (TTS) applications, or speech synthesis. In Toda et al. [2006] eigenvoices are used to adapt a gaussian mixture model that is used in the synthesis procedure.

Again, the advantages are quite similar to the ones for ASR, especially the ability to make a reasonable transform with only a small number of utterances. On the other hand, like with ASR there is no advantage over other adaptation methods when more utterances are used. Toda et al. [2006] showed that after more than 20 utterances other techniques gave a better performance.

Even though eigenvoices seem to be able to give a better transform for a small number of utterances, there use in TTS is limited as TTS systems need more detailed transformations to give a better performance.

7 Conclusions

Eigenvoices is a class of methods that is powerful for Speaker Adaptation with small amounts of adaptation data. Because of its simplicity and low requirements it is an easy to use adaptation technique.

The real power of eigenvoices is when combined with other methods. Combinations with both MLLR and MAP are giving an improvement over using these methods by itself.

One thing that this paper has shown is that there is a huge variety of eigenvoice methods, which is not beneficial for the usage and adaption of eigenvoice methods. If one would want to use eigenvoices, a lot of possible implementation paths present themselves, without having some ‘best’ methods.

Even though there is a huge variety of methods, almost all methods are compared with the standard PCA technique and Maximum Likelihood Eigenvoice Decomposition from Kuhn et al. [1998], making them the de facto standard. On the other hands, most other and newer methods are giving a better performance.

Interesting are alternative applications of eigenvoices like speech separation and synthesis. Especially the latter could become actual with the rise of HMM-based synthesis solutions in the context of speaker personalization.

References

- H. Botterweck. Very Fast Adaptation for Large Vocabulary Continuous Speech Recognition Using Eigenvoices. In *Sixth International Conference on Spoken Language Processing*. ISCA, 2000.
- K. Chen, W. Liao, H. Wang, and L. Lee. Fast speaker adaptation using eigenspace-based maximum likelihood linear regression. In *Sixth International Conference on Spoken Language Processing*. Citeseer, 2000.
- K.T. Chen and H.M. Wang. Eigenspace-based maximum a posteriori linear regression for rapid speaker adaptation. In *IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING*, volume 1. Citeseer, 2001.
- M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98, 1998. doi: 10.1006/csla.1998.0043. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.37.8252&rep=rep1&type=pdf>.
- Dong Kook Kim and Nam Soo Kim. Bayesian speaker adaptation based on probabilistic principal component analysis. In *ICSLP-2000*, volume 3, pages 734–737, 2000.

- R. Kuhn, P. Nguyen, J.C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Eigenvoices for speaker adaptation. In *Fifth International Conference on Spoken Language Processing*, 1998. URL http://www.isca-speech.org/archive/icslp_1998/i98_0303.html.
- R. Kuhn, J.C. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6):695–707, 2000. doi: 10.1109/89.876308. URL <http://dx.doi.org/10.1109/89.876308>.
- B. Mak and R. Hsiao. Improving eigenspace-based MLLR adaptation by kernel PCA. In *Eighth International Conference on Spoken Language Processing*. Citeseer, 2004. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.2333&rep=rep1&type=pdf>.
- B. Mak, J.T. Kwok, and S. Ho. A study of various composite kernels for kernel eigenvoice speaker adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP'04)*, volume 1, 2004.
- B. Mak, J.T. Kwok, and S. Ho. Kernel eigenvoice speaker adaptation. *IEEE Transactions on Speech and Audio Processing*, 13(5):984–992, 2005.
- Patrick Nguyen, Christian Wellekens, and Jean-Claude Junqua. Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments. In *Sixth European Conference on Speech Communication and Technology*, pages 2519–2522. Citeseer, 1999.
- T. Toda, Y. Ohtani, and K. Shikano. Eigenvoice conversion based on Gaussian mixture model. In *Ninth International Conference on Spoken Language Processing*. Citeseer, 2006.
- N.J.-C. Wang, S.S.-M. Lee, F. Seide, and Lin-Shan Lee. Rapid speaker adaptation using a priori knowledge by eigenspace analysis of mllr parameters. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 1, pages 345–348 vol.1, 2001. doi: 10.1109/ICASSP.2001.940838.
- R.J. Weiss and D.P.W. Ellis. Speech separation using speaker-adapted eigenvoice speech models. *Computer Speech & Language*, 24(1):16–29, 2010.
- Robert Westwood. Speaker Adaptation Using Eigenvoices. Master’s thesis, University of Cambridge, 1999.
- P.C. Woodland. Speaker adaptation for continuous density HMMs: A review. In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Meth-*

ods for Speech Recognition, 2001. URL http://www.isca-speech.org/archive/adaptation/adap_011.html.