A Category-Dependent Feature Selection Method for Speech Signals

Woojay Jeon and Biing-Hwang Juang

School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, U.S.A.
{wjjeon,juang}@ece.gatech.edu

Abstract

We present a novel method of dimension reduction and feature selection that makes use of category-dependent regions in high-dimensional data. Our method is inspired by phoneme-dependent, noise-robust low-variance regions observed in the cortical response, and introduces the notion of category-dependence in a two-step dimension reduction process that draws on the fundamental principles of Fisher Linear Discriminant Analysis. As a method of applying these features in an actual pattern classification task, we construct a system of multiple speech recognizers that are combined by a Bayesian decision rule under some simplifying assumptions. The results show a significant increase in recognition rate for low signal-to-noise ratios compared with previous methods, providing motivation for further study on hierarchical, category-dependent recognition and detection.

1. Introduction

In this study, we introduce a method of extracting category-dependent features from high-dimensional data. Specifically, we use a physiological model of the mammalian primary auditory cortex(A1)[1], discussed in previous studies[2], which offers a rich, dimension-expanded set of auditory features that may lead to computational models of cognition that better approximate the robust cognitive abilities of humans.

As a method of studying how the A1 response maps audio signals to the dimension-expanded space, we previously analyzed the variance[2] of the cortical response to different classes of phonemes. The A1 model exhibits distinct regions of low variance for each class of English phonemes. Drawing an analogy with existing physiological evidence of location-wise specialized processing in the brain, we postulated that these low-variance regions(LVR) contain information that is more relevant to identifying each phoneme compared to high variance regions. Furthermore, if these low-variance regions are indeed consistent with the noise-robust cognitive capabilities of human physiology, it is possible that the data in these dimensions are less susceptible to noise than others.

Although we used an LVR common to all phoneme classes as a means of reducing the dimensions of the cortical response in [2], it was already observed that distinct LVR's exist for certain *categories* of phoneme classes. Hence, this motivates us to conceive of *category-dependent* low-variance regions from which we select *category-dependent features* that contain better discriminative information compared with the single, category-independent LVR used in [2]. We propose a two-step method of feature selection where we introduce this notion of category-dependence inspired from experimental observation of the auditory model. However, this method of category-dependent feature selection can also be presented in a more general context of classification-constrained dimension reduction techniques, in particular, the well-known Fisher Linear Discriminant Analysis(LDA)[3]. Our method does not assume a specific mathematical model for the observed features. Rather, it is based more on heuristic, intuitive principles similar to those of LDA that result in a practical way to discard dimensions in high-dimensional data.

Finally, we quantitatively assess the viability of this method of feature selection by applying it in a conventional speech recognition experiment under some simplifying assumptions. Our results show a substantial increase in recognition accuracy than the previous study [2], and motivates us to develop recognition and detection methods using hierarchical category-dependency. We also note that the auditory model used in this work is only one of many possible models. What we wish to emphasize in this study is our methodology drawing on the notion of dimension expansion and category-dependent feature selection, which could be potentially applied to other auditory models or other automatic pattern classification tasks in general.

2. The dimension reduction problem

Pattern classification tasks are often based on Bayesian decision theory, where one tries to find the class that maximizes the *a posteriori probability* of a given observation. We also know that the probability of error is non-increasing as the number of dimensions in the observations increases. However, this framework is only applicable when we have an accurate estimate of the underlying distributions. False density estimates can lead to classification errors that are especially a problem for high dimensional data. The "curse of dimensionality"[3] is a well-known phenomenon where the amount of data required for density estimation increases exponentially with the number of dimensions. When training data is limited, the classification error can rapidly increase for an increasing number of dimensions[4]. Even if enough training data were available, the added computation cost in processing a larger number of features may be too high to justify the improvement in accuracy. Furthermore, attempting to estimate the densities accurately does not necessarily improve the classification error because estimation error and classification error do not always follow the same trends[5].

Hence, preprocessing the observations to reduce the number of dimensions is a conventional practice in pattern recognition, and can be performed based on two primary guidelines: 1. allow more reliable estimates of the underlying distributions of the data, and/or 2. retain as much discriminative information as possible. One common method that follows the second approach is LDA[3], which serves as a practical way to reduce the dimensions with minimal loss (in some sense) of discriminating power between the data classes. Although it is known to have optimality properties for Gaussian distributions with equal co-

variances, it is usually applied without assuming a specific underlying model, and has been found to generally give reasonable performance in many applications in part due to the stability of the between-class and within-class statistics[6].

3. Category-dependent feature selection

Using the total scatter matrix S_T and within-class scatter matrix S_W defined in [3], the goal of LDA is to find the transformation W that maximizes the quantity (note that "Tr" means "Trace"):

$$J(W) = \mathrm{Tr}\left\{ \left(W^T S_W W\right)^{-1} \left(W^T S_T W\right) \right\} \quad (1)$$

This is equivalent[7] to maximizing the determinant form in [3]. Also note that one can replace S_T with the between-class scatter matrix S_B in [3] and still have the exact same results [7].

One problem is that this requires the calculation of S_W and S_T followed by solving a generalized eigenvalue problem, which can be impossible when the number of dimensions are extremely high and computation and storage resources are limited. The dimension reduction method we propose invokes the two basic principles behind LDA, i.e., "reducing" S_W while "amplifying" S_T, as a natural way of introducing multiple, category-dependent features inspired by the notion of place-coding in the cortical response. By use of individual variances rather than covariances, and by decoupling the two elements of LDA into a two-step process, we avoid computing scatter matrices and solving the generalized eigenvalue problem for the entire observation space.

3.1. Low variance and categorization

First, assume a set C of the d-dimensional observation vector $\mathbf{x}$ that can be partitioned into N classes, i.e.,

$$C = w_1 \cup w_2 \cup \cdots \cup w_N, \;\; w_i \cap w_j = \emptyset \;\; (i \neq j) \quad (2)$$

Define the set $D = \{1, 2, \cdots, d\}$ as the set of indices of the d dimensions in $\mathbf{x}$. For each class, the within-class scatter is[3]

$$S_i = \sum_{\mathbf{x} \in w_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T, \; \mathbf{m}_i = \frac{1}{|w_i|} \sum_{\mathbf{x} \in w_i} \mathbf{x} \quad (3)$$

where $|w_i|$ denotes the cardinality of the set w_i. We define the variance of the k'th dimension of vector $\mathbf{x}$ in class w_i as:

$$\sigma_i^2(k) = \frac{1}{|w_i|} \sum_{\mathbf{x} \in w_i} \{x(k) - m_i(k)\}^2 \quad (4)$$

where $x(k)$ denotes the k'th element in $\mathbf{x}$. In this study, we are also interested in the "normalized variance,"

$$v_i(k) = \sigma_i^2(k) \Big/ \max_{k' \in D} \sigma_i^2\left(k'\right) \quad (5)$$

In our method, we first invoke LDA's notion of minimizing within-class scatter as a separate process of dimension reduction. In a sense, we separate the $\left(W^T S_W W\right)^{-1}$ term away from (1). We also constrain all columns of W to be unit vectors, which results in simply finding those raw dimensions in $\mathbf{x}$ with low variance. This greatly reduces the number of variables – with d dimensions in $\mathbf{x}$, $d(d+1)/2$ variables must be computed for the full scatter matrix, whereas d variables are needed when dealing with only its diagonal entries.

Furthermore, in the case where certain groups of classes share a more or less common set of dimensions with low variance, we group these classes as a *category*. Define a mapping of the set of all classes onto a set of categories:

$$g(w_i) : \{w_1, w_2, \cdots, w_N\} \rightarrow \{C_1, C_2, \cdots, C_M\} \quad (6)$$

This results in M non-overlapping categories, where each category C_j is the union of one of more classes:

$$C_j = \bigcup_{g(w_i) = C_j} w_i \quad (7)$$

where $\cup_{j=1}^{M} C_j = C$ and $C_l \cap C_j = \emptyset \; (l \neq j)$. For each C_j, we define the normalized and summed within-category scatter as follows:

$$T_j = \sum_{w_i \in C_j} \frac{S_i}{|w_i| \max_{k' \in D} \sigma_i^2\left(k'\right)} \quad (8)$$

Defining the summed normalized variance $u_j(k)$, we have:

$$\mathrm{Tr}\{T_j\} = \sum_{k=1}^{d} u_j(k) \;\text{ where }\; u_j(k) = \sum_{w_i \in C_j} v_i(k) \quad (9)$$

We then try to find a set D_j that contains the indices of the lowest values in $\mathrm{Tr}\{T_j\}$, under the constraint that the size of D_j must be some constant e_j. Note that this represents the first stage of dimensionality reduction, from d to e_j.

$$D_j = \arg\min_{D_j} \sum_{k \in D_j} u_j(k) \text{ where } |D_j| = e_j \quad (10)$$

This can be written in matrix form if we consider the *low-variance filter* matrix E_j consisting of unit vectors of the dimensions indexed by D_j:

$$E_j = \arg\min_{E_j} \mathrm{Tr}\left\{ E_j^T T_j E_j \right\}, \;\; E_j \text{ has } e_j \text{ columns} \quad (11)$$

The dimension-reduced vector resulting from this low-variance filtering is:

$$\mathbf{y}_j = E_j^T \mathbf{x} \quad (12)$$

The value e_j probably depends on a variety of factors such as the amount of available training data. Because of the simplicity of the statistic $v_i(k)$, e_j should be sufficiently large to avoid compromising too much data. In practice, it is possible that feasible values of e_j may be still too high for the given computational and storage resources, and other heuristic constraints may have to be added to allow a lower value of e_j. In the case of the cortical response, we determined e_j by gain-normalizing $u_j(k)$, thresholding it under some fixed constant, and constraining D_j such that it can only contain one point on each ϕ-axis in the (x, s, ϕ) space, motivated by the fact that the cortical response is usually smooth along the ϕ-axis.

3.2. Principal component analysis

Once we have obtained $\mathbf{y}_j$ for each C_j, we invoke the second notion of LDA, which is to transform the data such that there is maximum between-class scatter, and retain a reduced number of dimensions where the scatter is highest. When this is viewed separately as a second step in dimension reduction, it becomes equivalent to Principal Component Analysis(PCA)[7]. While PCA is originally by minimizing a squared-error function for data approximation, it effectively results in finding an orthogonal transformation W that maximizes $\mathrm{Tr}\left\{W^T S_T W\right\}$, which, in some sense, is like separating out the second multiplicative term inside the trace function in (1).

Since we now have different LVR's that are specific to each category of observations, it only make sense to apply PCA to each of these category-specific LVR's. Hence, for each category C_j, we compute the scatter matrix:

$$K_j = \sum_{\mathbf{y}_j \in C_j} (\mathbf{y}_j - \mathbf{u}_j)(\mathbf{y}_j - \mathbf{u}_j)^T, \quad \mathbf{u}_j = \frac{1}{|C_j|} \sum_{\mathbf{y}_j \in C_j} \mathbf{y}_j \quad (13)$$

To find the transformation P_j that maximizes $\mathrm{Tr}\left\{P_j^T K_j P_j\right\}$, we solve the eigenvalue problem $K_j \mathbf{p} = \lambda \mathbf{p}$ and let the columns of P_j contain a full set of orthonormal eigenvectors.

Table 1: *Categorization of 50 phonemes.*

No.	Phonemes	No.	Phonemes
1	b, p, hh, ax-h	7	y, iy, ux, ih, ey
2	d, t, g, k	8	m, n, en, nx, ix
3	dx, hv, f, dh, th	9	l, el, w
4	v, em, ng	10	aa, ao, ow, oy, uw
5	jh, s, sh, ch	11	ay, aw, eh, ae, ax, ah, uh
6	z, zh	12	r, axr, er

The data vector $\mathbf{y}_j$ is now transformed into a decorrelated(in the ideal case) vector $\mathbf{z}_j$ by the operation:

$$\mathbf{z}_j = P_j^T \left(\mathbf{y}_j - \mathbf{u}_j\right) \tag{14}$$

Finally, another matrix F_j of unit vectors is applied to select f dimensions from $\mathbf{y}_j$ corresponding to the largest eigenvalues of K_j, to form the category-dependent feature vector $\mathbf{x}_j$:

$$\mathbf{x}_j = F_j^T \mathbf{z}_j = F_j^T P_j^T \left(\mathbf{y}_j - \mathbf{u}_j\right) \tag{15}$$

4. Phoneme recognition

4.1. Phoneme categorization

To find the mapping function in (6) for a set of English phoneme classes, we perform an iterative, binary grouping of phoneme classes. We first define the following distance measure between two classes w_m and w_n using $v_i(k)$ defined in (5):

$$d_{m,n} = \sum_{k \in D} \left\{\log v_m\left(k\right) - \log v_n\left(k\right)\right\}^2 \tag{16}$$

The log function is applied as a practical way of emphasizing the similarity in low-variance rather than high-variance dimensions. For a category C_j, we define the intra-category distance δ_j as the sum of distances between all possible pairs (combined, not permuted) of classes contained in the category.

$$\delta_j = \sum_{w_m, w_n \subset C_j} d_{m,n} \tag{17}$$

Categorization is performed by first denoting each phoneme class as a category. At each iteration, we search for the two categories that, when combined, have the least δ_j, fusing them into a single category. The procedure is repeated until an arbitrary number of categories is obtained. Table 1 shows the categorization of 50 phoneme classes obtained using this iterative method. It is interesting to note that much of the grouping also makes intuitive sense, such as the tendency of vowels and consonants to be separated. Figure 1 shows the summed normalized variance $u_j(k)$ for these 12 categories $C_j (1 \leq j \leq 12)$ where k are those dimensions in D where $\phi = 0$.

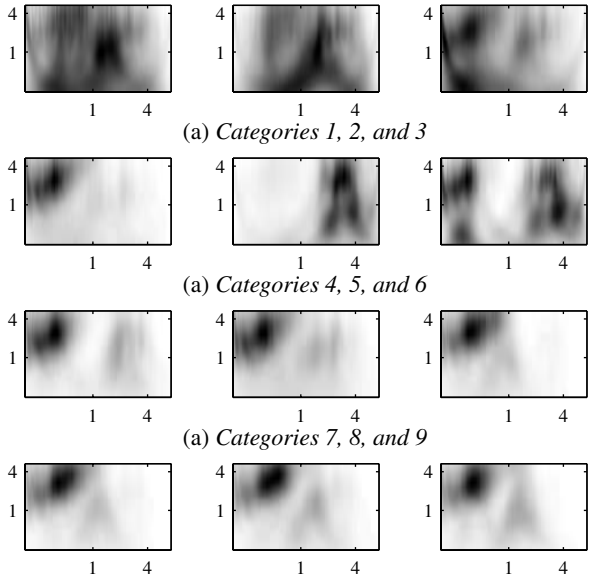
4.2. A composite recognizer

We now need a way to classify the original observation vector $\mathbf{x}$ into the N classes. Under a *maximum a posteriori*(MAP) decision rule, our ultimate goal is to find:

$$\arg\max_{w_i} P\left(w_i \middle| \mathbf{x}\right) = \arg\max_{w_i} p\left(\mathbf{x} \middle| w_i\right) P\left(w_i\right) \tag{18}$$

Assuming uniform priors, we need only the likelihood $p\left(\mathbf{x} \middle| w_i\right)$, which we decompose as follows:

$$\begin{aligned} p\left(\mathbf{x} \middle| w_i\right) &= p\left(\mathbf{y}_j, \mathbf{y}_j^c \middle| w_i\right) = p\left(\mathbf{y}_j \middle| w_i\right) p\left(\mathbf{y}_j^c \middle| w_i\right) && (19)\\ &= \frac{1}{\left|\det P_j\right|} p\left(\mathbf{z}_j \middle| w_i\right) p\left(\mathbf{y}_j^c \middle| w_i\right) && (20)\\ &= p\left(\mathbf{x}_j \middle| w_i\right) p\left(\mathbf{x}_j' \middle| w_i\right) p\left(\mathbf{y}_j^c \middle| w_i\right) && (21)\\ &= p\left(\mathbf{x}_j \middle| w_i\right) p\left(\mathbf{x}_j', \mathbf{y}_j^c \middle| w_i\right) && (22)\\ &= p\left(\mathbf{x}_j \middle| w_i\right) p\left(\mathbf{x}_j^c \middle| w_i\right) && (23) \end{aligned}$$



(a) *Categories 1, 2, and 3*

(a) *Categories 4, 5, and 6*

(a) *Categories 7, 8, and 9*

(a) *Categories 10, 11, and 12*

Figure 1: *Summed normalized variance (dark is high)* $u_j(k)$ *of cortical response at* $\phi = 0$ *for the categories in Table 1. The horizontal axis is frequency* (kHz), *while the vertical axis is scale* (cyc/oct).

Our first assumption in (19) is that the vector random variable $\mathbf{y}_j$ consisting of the low-variance dimensions and the vector random variable $\mathbf{y}_j^c$ consisting of the remaining dimensions are independent given the class w_i (for convenience's sake, we will use the same lower-case notation for both observations and random variables). An intuitive explanation for this assumption is that the low variance regions contain information on the identity of the phoneme class that is fairly robust to noise or other distorting factors, while the high-variance regions are more susceptible to noise. We conjecture that as long as the phoneme class is given, the probability density function(pdf) of the stable region can be described without dependence on the unstable region, i.e., $p\left(\mathbf{y}_j \middle| \mathbf{y}_j^c, w_i\right) = p\left(\mathbf{y}_j \middle| w_i\right)$.

In (20), we have applied the relation in (14) to transform the pdf's. Since P_j is an orthogonal matrix, $\left|\det P_j\right| = 1$. In (21), we have applied a second assumption, that the vector $\mathbf{x}_j$, which contains features corresponding to high eigenvalues, is independent of $\mathbf{x}_j'$, which contains the remaining features in $\mathbf{z}_j$. Although the transformation P_j (ideally) decorrelates the two variables, we go further to assume they are independent.

Since we already assumed in (19) that $\mathbf{y}_j^c$ and $\mathbf{y}_j$ are independent, $\mathbf{y}_j^c$ and $\mathbf{x}_j'$ are also independent. Combining $\mathbf{x}_j'$ and $\mathbf{y}_j^c$ as the single vector $\mathbf{x}_j^c$, we obtain the final decomposition in (23). In effect, $\mathbf{x}_j^c$ represents all those dimensions that are discarded in the entire process of feature selection. However, since it is hard to estimate its distribution, we make the following assumption that it can be approximated by the product of likelihoods of all the other category-dependent features.

$$p\left(\mathbf{x}_j^c \middle| w_i\right) = \prod_{m \neq j}^{M} p\left(\mathbf{x}_m \middle| w_i\right) \tag{24}$$

This results in the following decision rule:

$$\arg\max_{w_i} p\left(\mathbf{x} \middle| w_i\right) = \arg\max_{w_i} \prod_{j=1}^{M} p\left(\mathbf{x}_j \middle| w_i\right) \tag{25}$$

Figure 2 shows a schematic overview of the feature selection process followed by recognition. Hidden Markov

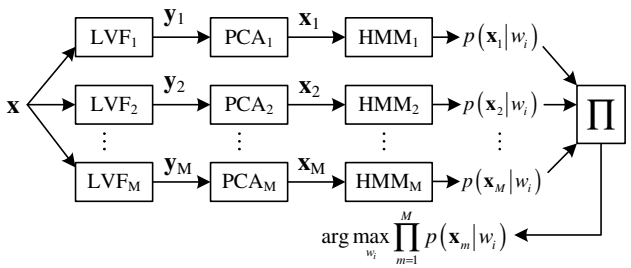


Figure 2: *Schematic overview of category dependent feature selection followed by a composite recognizer.* $\mathbf{x}$ *has* d *elements,* $\mathbf{y}_j$ *has* e_j *elements, and* $\mathbf{x}_j$ *has* f *elements. The low variance filter* LVF_i *symbolizes the transformation in* (12) *while the* PCA_i *block indicates the transformation in* (15).

Model(HMM)'s are used to compute the individual likelihoods of each feature vector for all classes, then combined at the final stage to produce the final MAP decision.

5. Experimental results

We used the experiment in [8] as a reference to test the features in an actual speech recognition task. For the category-independent features, we took 50 phonemes from [8] excluding the silence and closures and added the "ax-h" phoneme to obtain a total number of 51 phonemes. We performed the folding as described in [8] ("ax-h" was folded into "ax") to build models for 44 phonemes. To obtain category-independent features from the cortical response, we simply set $C_1 = C$ and followed the procedure discussed in the previous sections. Phoneme samples from the TIMIT database excluding the "sa" sentences were used for the entire process of low-variance filtering, feature selection, and model training and testing. We also disregarded the within-group confusions in [8] to obtain final recognition results for 38 phonemes. For category-dependent features, 50 phonemes (excluding silence, closures, and "eng" for which there were very little samples available) were analyzed without folding, and categorized following the method discussed in Section 4.1. This resulted in the grouping shown in Table 1. The folding in [8] was applied at the final stages as part of the processing of within-group confusions in [8] to, again, obtain final recognition results for 38 phonemes.

The auditory spectrum was calculated at 10ms steps using a time constant of 8ms for the final leaky integration, consisting of 128 frequency channels ranging from 180∼7040 Hz. The cortical response was computed using 20 scale channels ranging from 0.25 cyc/oct to 4.6 cyc/oct, and 11 phase channels from $-\pi/2$ to $\pi/2$. Hence, $d = 128 \times 20 \times 11$. For the low-variance filters, e_j ranged from 957 to 2256, found by applying a fixed threshold on the gain-normalized $u_j(k)$'s. In the final dimension reduction using PCA in (15) we set $f = 12$ in all cases. Energy, delta, and acceleration coefficients were then appended to form a total length of 39. Each recognizer was modeled for isolated phoneme utterances using a five-state (three emitting) HMM with skips allowed from the second and third states to the last state in order to accommodate utterances that were only one or two frames long. While these experimental configurations were probably suboptimal, the purpose was to compare the different features rather than maximize performance.

As noticeable in the results, there is a substantial performance improvement when using the category-dependent features, especially under low SNR. The results imply that this methodology makes better use of the dimension-expanded cortical response and its noise robustness, whereas the application of a single low-variance filter incurs heavy penalties on the class discriminability as the number of phonemes increases.

Table 2: *Speech recognition rates.*

	MFCC				Cat.-Indep. Features				Cat.-Dep. Features			
	1	4	8	32	1	4	8	32	1	4	8	32
C	44.2	52.3	55.2	59.7	37.7	45.5	47.9	53.4	41.5	50.9	54.0	58.8
20	37.9	45.4	47.5	50.8	37.2	44.8	46.9	52.5	40.9	49.9	52.9	57.8
15	33.1	39.6	41.7	44.7	36.0	43.4	45.1	50.2	39.9	48.9	52.0	56.4
10	26.5	30.7	32.9	36.0	33.1	39.4	40.6	44.8	37.1	45.5	48.6	52.8
5	18.2	20.2	22.2	24.7	27.9	42.5	33.1	34.6	30.9	37.8	40.9	44.0

6. Conclusions and future work

In this study, we have introduced a method of feature selection applicable to very high-dimensional data that draws on the intuition behind Fisher LDA. By breaking the coupling of the within-class and between-class scatter into a two-step process while introducing the notion of category-dependence inspired from studies of a physiological model of the auditory system, we have developed a method that gives improved results for our cortical model. In particular, the use of category-dependent low-variance regions seems to coincide well with the noise robust properties of the auditory model, resulting in enhanced performance under noisy conditions. While we have greatly simplified the final decision making process in this study, the results motivate future study in which we can introduce the notion of hierarchical, category-dependent *decisions* in addition to the category-dependent feature selection process. Note that the method proposed here does not assume any specific mathematical models for the observations and, rather, is based more on heuristic, intuitive ideas on dimension reduction that are similar to LDA. A more rigorous, analytical framework for this feature selection method, as well as a more quantitative relation to the noise robustness of the cortical response, will also have to be developed in future study.

7. References

[1] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 5, pp. 382 – 395, Sept. 1995.

[2] W. Jeon and B.-H. Juang, "A study of auditory modeling and processing for speech signals," in *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 1, Philadelphia, PA, Mar. 2005, pp. 929–932.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley and Sons, Inc., 2001, pp. 117, 170.

[4] A. K. Jain and W. G. Waller, "On the optimal number of features in the classification of multivariate gaussian data," *Pattern Recognition*, vol. 10, pp. 365 – 374, 1978.

[5] J. H. Friedman, "On bias, variance, 0/1-loss, and the curse-of-dimensionality," *Data Mining and Knowledge Discovery*, vol. 1, pp. 55 – 77, 1997.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer-Verlag, 2001.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. Academic Press, 1990.

[8] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Mar. 1989.