Probabilistic Pursuit, Classification, and Speech

by

Upendra V. Chaudhari

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1997

Author .....
Department of Electrical Engineering and Computer Science
September 5, 1997

Certified by .....
Sanjoy K. Mitter
Professor of Electrical Engineering
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Probabilistic Pursuit, Classification, and Speech

by

Upendra V. Chaudhari

Submitted to the Department of Electrical Engineering and Computer Science
on September 5, 1997, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In this thesis, we are concerned with signal analysis toward the goals of both efficient representation and efficient classification. Thus, we consider analyses, called pursuits, that seek out the elements of a signal which characterize it best. The general structure of pursuit algorithms is extended to a probabilistic framework, thereby developing a Probabilistic Pursuit (PP) algorithm in which the search is no longer deterministic. The algorithm is based on a prior signal model in the form of a probability distribution on the search space. It is used in defining a stochastic process on that space, giving direction to the search. Accurate prior information allows us the efficiency of favoring the representation of signal over noise. In this context, we develop a Minimum-Time Decomposition principle and use it to construct an efficient, application independent classifier. Then, we describe the time-frequency analysis implications of the Probabilistic Pursuit by considering a specific search space, namely the set of Gabor functions. We apply these ideas to speech classification. Motivated by the theory of speech production, prior signal models are developed for a number of VCV utterances. These are based on novel non-stationary models for speech and use predicted formant path information. The models are used in an instantiation of the classification paradigm. Our results suggest that the new representation derived from the PP provides unique and useful information for classification.

Thesis Supervisor: Sanjoy K. Mitter
Title: Professor of Electrical Engineering

Acknowledgments

I was fortunate to have a thesis committee that was both understanding and responsive. I would like to thank my advisor, Sanjoy Mitter, for his inspired and insightful guidance. His ability to make connections among a variety of scientific disciplines in conjunction with the ability to recognize which of those are relevant contributed significantly to this thesis and my education. His generosity towards his students is something that I will remember.

I thank the other members of my committee, Ken Stevens and Alan Willsky, for taking time to read the thesis and help me better understand many of its aspects. Their thought provoking comments and suggestions were always to the point and much appreciated.

Thanks to Vivek Borkar for reading and providing helpful comments on parts of this thesis.

Thanks also to Kathleen O'Sullivan and Amy Briemer for their considerable efforts.

I would like to thank my family, Vasant, Shamala, Madhu, and my friends, for their support.

This work was supported by:

The US Army Research Office under contract DAAL03-92-G-0115.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we are concerned with signal analysis toward the goals of both efficient representation and efficient classification. For this reason, representation is not simply the computation of a fixed transformation of a given observation signal. Rather, the analysis seeks out those elements in the signal which characterize it best, a technique known as a Pursuit. Such a procedure leads to representations which are efficient in the sense that they are compact, consisting, for example, only of elements of the transformation judged to be important. However, the selection of these elements is based only on information contained in the signal being analyzed. Herein, we develop a novel probabilistic generalization of Pursuit type algorithms which we call the Probabilistic Pursuit where in addition to the information in the observation, *a priori* information is also used in the form of a probability distribution on the search space. Such a distribution constitutes for us a prior signal model.

Fundamental to the structure of Pursuits is a search algorithm and generally what is considered important is the result of the search, which is deterministic. But in our development we give meaning to the search itself by basing it on probabilistic prior signal models, in the context of which we will interpret the length of time it takes for us to search as an indication of the validity of our model. Thus we derive a Minimum-Time Decomposition Principle on which we base our novel classification scheme. The Probabilistic Pursuit representation scheme, and hence the classifier based on it, has greater efficiency because it allows us to ignore aspects of the observation consisting

of noise. These are universal results applicable to many different domains.

We give insight into how the application independent results interact with a specific application domain, in particular speech. In order to do so, a specific search space, or dictionary, must be specified. In this case, the Gabor dictionary is chosen, the use of which allows the extraction of localized time-frequency information. In particular, we show that a Probabilistic Pursuit analysis with the Gabor dictionary results in a representation containing, in the limit, more information than a Short Time Fourier analysis. Given a set of speech utterances, we show how to construct prior models for them on the Gabor dictionary via the use of novel non-stationary speech models. The arguments used to elucidate the meaning of the search are based on a consideration of the theory of speech production. The results of these application specific considerations are combined with the universal, application independent results to develop a speech classification experiment.

1.1 Background

The history of Pursuit algorithms can be traced back to Tukey, Friedman, Stuetzle, Huber, and others [13]. The basic motivation for them is that high dimensional data is very difficult to analyze. That it is difficult to visualize is obvious when one thinks about data in greater that four dimensions. But it is also true that the data is generally sparse and does not "fill up the space" as it were. Both considerations suggest that we should look for appropriate low dimensional subspaces which carry most of the information. Moreover, the representation is efficient because fewer variables are needed to describe the data.

Principal Component analysis is based on such a philosophy. One projects high dimensional data onto the space spanned by eigenvectors belonging to the biggest eigenvalues of its covariance matrix. In this case, we see that the search is characterized by looking for large eigenvalues and their associated eigenvectors. It is worth mentioning here that just because these subspaces contain most of the data, that in and of itself does not mean that these subspaces are meaningful [6]. To a large extent,

the significance of the representational elements chosen is dependent on the application. If dimension reduction in a representation is the goal, then principal component analysis is appropriate. However, if discrimination based on a representation is the goal, this type of analysis may overlook critical but insubstantial data. The connection between a given application and the search for characteristics in a signal that support it, lies in the objective function. Let us describe the general technique as follows.

1.1.1 The Pursuit Paradigm

Our definition of the paradigm is simply that given an input, we search iteratively over a dictionary of explanatory objects for the one which matches our signal the best. Since we are interested in the comparative match of the elements, the match will always be expressed as a numerical quantity with larger or smaller values, depending on the situation, indicating better matches. Then, the next time that we perform this search, our new input consists of the original input plus the previously chosen explanatory object(s). In this way we get a sequence of dictionary elements which explain the data, according to our matching criterion. We may in various contexts refer to this sequence as a decomposition as well as a construction of our input. In either context, dictionary elements are referred to as *atoms*. The decomposition's overall properties are governed by those of the dictionary, the matching criterion, as well as those of the search procedure. We will make an attempt in this thesis to indicate such connections. Already we see though that the particular explanatory objects chosen depend on the input signal and will in general be different for different inputs. This means that given two signals, the collection of elements chosen in the analysis of each separately may be different from the elements chosen in the analysis of their sum, implying non-linearity.

On the other hand, the specification of the dictionary indicates what we believe to be the important features of our signal. If we consider all of the inputs we might see and take away from that the totality of signals we could construct from our dictionary, we get what might be considered noise. Since we are specifying beforehand what might

be significant, we expect that we will require fewer parameters in our representations.

1.1.2 Previous Work

The Matching Pursuit Decomposition [18] with the Gabor dictionary is such a decomposition. The theory is developed in a Hilbert space $\mathcal{H}$. The input is assumed to be an element of $\mathcal{H}$, the dictionary is a subset of elements of $\mathcal{H}$, and the matching criterion for selecting dictionary elements is based on the inner product. When the space is L_2 and the dictionary is the Gabor dictionary, a set of modulated Gaussians, the result is an adaptive representation for signals whose time-frequency characteristics vary in time. Its adaptive nature is based on the fact that measurements of the match between the input and dictionary elements are used to select elements of the representation, and as mentioned before, this leads to non-linearity. The decomposition is based completely on the particular input signal, the dictionary, and the matching criterion. Effectively, the entire dictionary is searched and the best element is chosen at each step.

1.2 The Contributions of this Thesis

1.2.1 Probabilistic Pursuit

A careful analysis of the procedure leads us to believe that we can exploit its sequential structure. There is information in the way in which elements are picked, the nature of which can be made precise by introducing probability and prior information in the context of pursuit algorithms. A contribution of this thesis is that, rather than searching the entire dictionary, we look at elements one at a time in a principled way. We call the procedure that we define a **Probabilistic Pursuit**. We assume that we have a probability distribution defined on the dictionary, which we call a **prior signal model**, and we look at the elements in the order that they appear in an i.i.d. sequence chosen according to this distribution. When one is found that matches the signal well, it is chosen and the search is started again. Note that when a match is

found, we know in addition to the element description its location in this sequence, or its waiting time. By introducing probability into the search procedure, we are able to define a set of theoretical waiting times for the dictionary elements which will depend on the distribution that exists over these elements. These times will be used as an indication of how well the probability distribution over the dictionary models our input. In fact, a distribution on the dictionary represents our assumptions of what is signal and what is noise, and we can partition the dictionary accordingly. Then, a measure of fit for the model is that elements that are noise should have large waiting times, whereas the signal elements should have short waiting times. A related but different measure of fit deals with the distance of the chosen dictionary elements from the signal model. We will consider parametric dictionaries and this distance will be measured in the space of parameters. A more precise description requires us to be more specific about the dictionary, so we reserve the discussion until then.

Super-dictionary

Conceptually, and later practically, we deal with a set of signals $\{u_i\}$, one of which we may observe as our input, which may contain an additional random noise. For each of the u_i, we will have a prior signal model $\mathbf{P}_i$, which as mentioned before is synonymous with distribution on a dictionary, and they will be collected in $\mathcal{P} = \{\mathbf{P}_i\}$, called the *super-dictionary*. What it means to analyze a given input is that an element of $\mathcal{P}$ is chosen, and then a Probabilistic Pursuit is performed with respect to it.

Representation

In this context, we develop results showing that when we have a good prior model, the analysis should choose signal elements before it chooses noise elements as indicated by a partitioning of the dictionary. The efficiency gained here comes as a result of not matching the noise part of the observation. Also, we develop a result to show that a good prior model should in fact give small waiting times.

The following scenario illustrates the benefits of using a prior signal model when performing an analysis on an observation. We can think of a class of signals, $\{u_i\}$, each

element of which matches different parts of the dictionary. Furthermore if one took the union of all these parts, there would still be elements of the dictionary left out. That is, the complement of this union in the dictionary is non-empty. In this case, regardless of which u_i we actually encounter, in our search it would be advantageous to ignore the part of the dictionary that was left out, i.e. in the complement, because those dictionary elements would only match noise. More precisely, our observation would be one of the $\{u_i\}$ plus some noise. Though we do not know beforehand which u_i will occur, we do know that certain parts of the dictionary can only be used in representing the noise part of the observation. Thus, by concentrating probability on the union, the analysis could ignore noise.

Classification

Given an unknown input from $\{u_i\}$ we perform in parallel a separate analysis with respect to each element of $\mathcal{P}$. By comparing the results for each, we select one as explaining the input best based on waiting time and distance measures. Thus, we view classification as a pursuit over $\mathcal{P}$ where the matching criterion is based on waiting times and parameter space distances. Viewing classification as an hierarchy of finer and finer grained coverings of $\{u_i\}$, our scheme will, in general, provide information at an intermediate level of the hierarchy. The decision could then be refined, moving to a higher level, by further processing of the decomposition.

Some properties of this classification scheme are efficiency, in that it will be based on a small number of parameters, and parallelizability. The following discussion provides motivation: As mentioned before, the probability distribution on the dictionary represents our prior knowledge, or our assumption of what the best matching elements are going to be. Furthermore, each model is based on a different element of $\{u_i\}$. If the model is wrong, we will have to wait a long time to completely analyze a signal. On the other hand, if the model is good, elements will be chosen quickly. Turning this around, if indeed we have to wait a long time for a match, then this is an indication that the prior model embodies a poor assumption about the input signal characteristics. From this idea we develop a Minimum-Time Decomposition principle

by which we associate model complexity to the search time. That is, a prior model is complex if using it causes the procedure to search for a long time in order to find a matching element. On the other hand, a prior model is simple when by using it, the search procedure easily finds matching elements. Similarly, a poor model implies large parameter space distances.

Discussion

Continuing the above discussion then, one could view different distributions on the dictionary as templates. That is, when we analyze different noiseless inputs with a particularly limited dictionary, we should get different amounts of matching with the various elements. On the other hand, if we analyze different noiseless inputs with a very general dictionary, the procedure will select different sets of explanatory elements. Informally, if these sets are appreciably different, then they can serve to identify the input. Consider using each of these sets in succession as dictionaries for a set of decompositions of one of the inputs. The best matching will occur when the dictionary is the one that the input itself generated. When we introduce a random noise into the picture, this may no longer be true. Rather than considering different sets of elements, we take different probability distributions on a general dictionary, as in the Probabilistic Pursuit, and in this way we can better deal with the noise issue.

In general, the elements of $\mathcal{P}$ will not represent those of $\{u_i\}$ in an exact sense. The difference is modeling error and in essence means that the assumptions about the signal and noise parts of the dictionary are not exactly correct. But based on these assumptions the elements that we believe constitute noise will have a low probability. To ensure that we can perform an analysis even when our assumptions are wrong, we might stipulate that a fall-back scheme be used when very long waiting times are encountered. From a signal analysis point of view, we can use the procedure to obtain a more compact representation of a given signal for which we have some information. Probabilistically then, we look in the areas where we know there should be energy. In this case, we show that we are more likely to get signal atoms before noise atoms. But, since the method is probabilistic with non-zero probabilities for all of the dictionary

elements, we will eventually look in areas that are away from our model and so we will eventually pick up noise.

To summarize, the probability distribution on the dictionary is really our prior signal model and whereas before we were matching simply the input signal, we are now matching the signal plus a signal model which represents our *a priori* information. At this point we note that the previous discussion was independent of any application domain and in particular it was independent of the specific nature of the dictionary. On the other hand, it is the particular structure of the Gabor dictionary and L_2 matching criterion which will allow us to analyze speech in a meaningful way.

1.2.2 Application to Signal Analysis: Speech

Dictionary and Matching Criterion

By choosing a particular dictionary and associated matching criterion we are in effect specifying the application space. Like in the Matching Pursuit, we choose a Hilbert space, L_2 and the Gabor dictionary. In this context the expression for the inner product is shown to be identical to that for a windowed Fourier coefficient. Under these conditions, there is a result showing that the analysis, considered as a construction of the observation, will in fact converge to the observation. Thus in the frequency domain, the construction will converge to the Short Time Fourier Transform and any information obtainable from this representation, e.g. cepstra, is still present in the representation. In addition, we have time-frequency localization beyond the capabilities of the Short Time Fourier Transform, and hence more information. Cepstral parameters are used in a Hidden Markov Model framework as the components of a feature vector. With the Gabor dictionary based Probabilistic Pursuit representation, this feature vector could be extended to include the extra time-frequency localization information.

Further, we provide new analysis of the significance of the pursuit by considering our inputs to be realizations of semi-stationary oscillatory processes, the theory of which was developed by Priestley [22]. These can be thought of as being characterized

by a two dimensional function $h_t(\xi)$ which, for each time and frequency pair, specifies the relative strength of the signal. This function is called the evolutionary spectral density. The term semi-stationary indicates that the components of this density are changing slowly with time. Our analysis shows that the pursuit looks for peaks of $h_t(\xi)$.

Classification Paradigm

The ultimate goal of a speech recognition system is the mapping of a speech waveform to linguistic phonemes or words, and as such is really a subset of the speech understanding problem. It may take a rather direct route by mapping the waveform into a sequence of spectral components and then mapping that sequence into a string of words, or it may take a more sophisticated route involving the estimation of various speaker parameters from the waveform, or more commonly a collection of waveforms from the same speaker, and then using that information in the mapping from waveforms to words. Characteristic formant positions is an example of such information and can in fact be used as a form of speaker normalization.

The recognition problem is most often seen as one of classification. That is, we know beforehand the set of utterances that we may see. For each of these a model is constructed. This model information is prior information. The use of prior information itself is a characteristic of speech analysis [25] [21]. It is common for this information to be statistical, and such models must be built by training. The Hidden Markov Model is the ubiquitous paradigm for this case. However, templates are also used, and these are created based on our knowledge of speech. This may be a superficial distinction, between statistical models and templates, and really only serves to indicate how the models are created. We follow the latter method.

Non-Stationary Models for Speech

Previously, the set $\mathcal{P}$ was introduced as a super-dictionary of prior signal models in a universal setting. To translate the results into the application domain, we show how to construct the elements of $\mathcal{P}$ by using speech formant paths as approximations to $h_t(\xi)$.

Having studied the consequences of our Probabilistic Pursuit on semi-stationary oscillatory processes, we relate speech, through a development of the theory of speech production, to the defining characteristics of these processes, embodied in the evolutionary spectral densities. This is a novel contribution in that we are modeling speech as a non-stationary process rather that a set of locally stationary processes. That is, for each element of $\{u_i\}$, we essentially predict the formant paths to get say, $\{h_t^i(\xi)\}$. Normalizing each density, we get $\mathcal{P}$, a set of distributions on the Gabor dictionary. At this point, the classification method described in the application independent section can be used in a discrimination experiment. As was mentioned previously, the classification made will be at an intermediate level of an hierarchy. We could use Dynamic Programming techniques to capture, for example, the temporal correlations in the selected Gabor dictionary elements to perhaps refine the classification.

Classification Experiment

The set of utterances we choose has in it voiced stop consonants and glides. Pairwise differences in the predicted formant paths of the utterances can be great as well as small. We perform the experiment with the goal of showing that our classification procedure is efficient and can separate broad classes of data.

1.3 Thesis Outline

Chapter 1 Introduction

Chapter 2 Probabilistic Pursuit

First, background information is given describing the general ideas behind Pursuit methods. In the course of describing various existing methods, the basic Pursuit framework of interest in this thesis is developed. We then go on to develop the Probabilistic Pursuit and discuss the implications for signal representation and classification. Unique to this pursuit is the use of a prior signal model in the form of a

probability distribution on the search space. By searching in this space probabilistically, we are able to associate the length of time required in the search to a notion of model fit. An important characteristic of this discussion is that it is universal, in that it does not depend on a particular application.

Chapter 3 Probabilistic Pursuit - Application to Time Frequency Analysis

In this chapter we will instantiate the universal results presented in the first chapter. In particular, the Gabor dictionary will be studied as the search space, and the relationship to Time-Frequency analysis will be drawn out. First, deterministic signals will be looked at. But then we extend the discussion by considering non-stationary stochastic processes as well. We will consider these semi-stationary oscillatory processes to be inputs to the Pursuit and argue that the selected measurements reflect the densities of the processes. We develop a continuity result that allows the use of the application independent results in Chapter 2.

Chapter 4 Analysis of Speech

The goal of this chapter is to provide justification for building prior signal models from the resonance characteristics of speech. By looking at the theory of speech production, we argue that in many cases, the formants contain most of the energy of a signal. As they evolve in time, they trace out tracks in the time-frequency plane. A connection is then made between formants in a signal and the density of a semi-stationary process from which we argue that the formant paths give a good indication of the matching dictionary, or search space, elements.

Chapter 5 Classification Scheme and A Preliminary Experiment

The ideas from the preceding chapters are here combined into an experimental framework for classification. We describe the nature of the implementation of the Probabilistic Pursuit and the parameters extracted from it to make classification decisions based on a Minimum-Time Decomposition principle. Furthermore, the parallel nature of the method is discussed. The experiment, consisting of the identification of

stops in vowel contexts, is then described and the results are discussed. They suggest that the new representation derived from the Probabilistic Pursuit provides unique and useful information for classification.

Chapter6 Discussion, Modifications, and Extensions

After summing up, we go on to describe a number of modifications and extensions to our work. We suggest techniques to improve the accuracy of the measurements used in the selection criterion for a Probabilistic Pursuit analysis of semi-stationary oscillatory processes using the Gabor dictionary. Also, whereas the classifier we develop is efficient and parallelizable, classification accuracy could perhaps be improved by post-processing. This might include the use of Dynamic Programming to take advantage of correlations that may by present in the representation. Also, we based our models on predicted formant paths. We discuss statistical model building as a way of deriving models from training data.

Chapter 2

Probabilistic Pursuit

2.1 Pursuits - Background

Inferences drawn from observations depend largely on their salient characteristics. First, this requires us to have a definition of what a salient characteristic is and then we must be able to find these in the observed data. Thus the Pursuit is characterized by an objective function with respect to which the search is performed. Once a feature is found, the data associated with it should be removed from the observation since it may obscure the search for later features. This suggests an iterative framework.

2.1.1 Projection Pursuit

Consider the case of a high dimensional Euclidean space where the observation consists of a set of points, commonly referred to as a point cloud. In general this cloud does not fill the entire space, and is concentrated in a small number of subspaces. Projection pursuit searches over the set of projections to find these spaces. Huber [13] has given an objective function which is maximized by the eigenvector corresponding to the largest eigenvalue of the covariance matrix of the point cloud. Once found, he gives a method to find the eigenvector corresponding to the second largest eigenvalue, and so on. The result is that the data is projected onto the subspaces generated by the eigenvectors associated with the largest eigenvalues. The method is Principal

Component analysis.

2.1.2 Projection Pursuit Regression

This theory was developed by Friedman and Stuetzle [11]. The goal is to approximate a function which can vary over many variables arbitrarily by a linear combination of functions which can vary along only one direction in the space of variables. These functions are called ridge functions and are constant on hyper-planes. The idea here is that salient characteristics of a complex function can be expressed in terms of simple functions.

2.1.3 Matching Pursuit

Mallat and Zhang [18] develop a Pursuit in the context of a Hilbert space, $\mathcal{H}$. The observation, or input as we will also call it, is a function $f \in \mathcal{H}$. The objective function is derived from the inner product and the residuals are formed by removing the projections in the chosen directions. In this case, the search is restricted to a dictionary $D \subset \mathcal{H}$, which represents our generating capacity for signals. "Noise" consists of the signals we will not try to match. Before describing the procedure in more detail, we present a generalization of the framework given in [18].

A General Class of Decompositions

We work here with a function class, and impose probability when needed. So, let $\mathcal{H}$ be a Hilbert space of functions with an associated norm $\|\cdot\|$ derived from its inner product. (e.g. L_2 with $\|\cdot\|_2$) $D = \{\phi_\gamma\}_{\gamma \in \mathcal{I}}$, a complete set of elements in $\mathcal{H}$. This means that the closed linear span of the elements in D is equal to the whole space:

$$\overline{L(D)} = \mathcal{H},$$

where $L(\cdot)$ denotes the linear span of its arguments, and closure in the Hilbert space topology is indicated by the over-line. $\mathcal{I}$ is the parameter space for the dictionary D.

Furthermore, we assume a one to one correspondence between the elements of D and the elements of $\mathcal{I}$ and define the following set valued map, where for $A \subset \mathcal{I}$,

$$\phi(A) = \{\phi_\gamma : \gamma \in A\},$$

and for $A \subset D$

$$\phi^{-1}(A) = \{\gamma : \phi_\gamma \in A\}.$$

At this point let us introduce a measurable space $(\Omega, A_\mathcal{I})$, where $\Omega = \mathcal{I}$, which will be referred to in the sequel. $A_\mathcal{I}$ can be taken as the Borel sets. By decomposing $f \in \mathcal{H}$, we mean finding a set of constituent dictionary elements, not necessarily unique, that make up f as an infinite linear combination, $f = \sum_{n=1}^{\infty} c_n \phi_{\gamma_n}$, where each ϕ_{γ_n} is an element of D and each c_n is a coefficient. That we will be able to do this for every $f \in \mathcal{H}$ is a consequence of the completeness of D. A fundamental property of the decomposition is that on each iteration i, an element of D will be chosen to be the i^{th} element of this linear combination, which as $i \to \infty$, should converge in norm to the desired function. To this end, we formally define an iterative decomposition to be a sequence

$$\{(D_0, \phi_{\gamma_0}, f_0), (D_1, \phi_{\gamma_1}, f_1), (D_2, \phi_{\gamma_2}, f_2), \cdots\}$$

where we have D_i, ϕ_{γ_i}, and f_i, which we will call the i^{th} dictionary, (selected) element, and residue. Each D_i is a subset of D, the full dictionary. The function which governs each iteration is

$$\Gamma : 2^D \times \mathcal{H} \mapsto \mathcal{D} \times \mathcal{H}.$$

Given a dictionary, $\in 2^D$, and a function, $\in \mathcal{H}$, Γ determines an element of D for the approximation and a residue in $\mathcal{H}$.

$$(\phi_{\gamma_i}, f_{i+1}) = \Gamma(D_i, f_i). \tag{2.1}$$

where ϕ_{γ_i} is selected to match f_i according to the desired criterion, and

$$f_{i+1} = f_i - < f_i, \phi_{\gamma_i} > \phi_{\gamma_i}. \tag{2.2}$$

Define the projection operators $\pi^1 : D \times \mathcal{H} \mapsto \mathcal{D}$ and $\pi^2 : D \times \mathcal{H} \mapsto \mathcal{H}$ as

$$\pi^1(\phi_\gamma, f) = \phi_\gamma,$$

and

$$\pi^2(\phi_\gamma, f) = f.$$

Then equation 2.1 can be rewritten as

$$(\phi_{\gamma_i}, f_{i+1}) = (\pi^1(\Gamma(D_i, f_i)), \pi^2(\Gamma(D_i, f_i))).$$

For $i = 0$, take $f_0 = f$, the function we wish to decompose. ϕ_{γ_0} is then the first dictionary element chosen to be in the linear combination, which starts with $i = 0$. The variation of D_i with i can be arbitrary. For $m \geq 0$, the m^{th} approximation is

$$\tilde{f}_m = \sum_{i=1}^{m} < f_i, \phi_{\gamma_i} > \phi_{\gamma_i} = \sum_{i=1}^{m} \alpha_i \phi_{\gamma_i}.$$

The decomposition is successful if $\| f_i \| \to 0$ as $i \to \infty$.

Note: This formulation is a generalization of the Matching Pursuit in the sense that we allow the dictionary to vary on each iteration.

To better understand this linear combination, note that the determination of f_i requires the subtraction of the dictionary elements chosen on previous iterations. We could write our construction as

$$\tilde{f}_m = \sum_{i=1}^{m} < f_i, \Phi(f, i) > \Phi(f, i),$$

where the function $\Phi(f, i) : \mathcal{H} \times Z^+ \mapsto D$ takes a function and iteration number

as arguments and returns the dictionary element chosen on that iteration when the given function is being decomposed. In terms of the previous notation, $\Phi(f,i) = \pi^1(\Gamma(D_i, f - \alpha_1\phi_{\gamma_1} - ... - \alpha_{i-1}\phi_{\gamma_{i-1}}))$. Note that the sequence of D_is is known. This representation makes the dependence of the residues on the previous elements explicit in the function $\Phi(f,i)$, which is a non-linear function of the input f. In fact, $\Phi(f,i)$ always returns a dictionary element. Since in general it is not true that the sum of two dictionary elements is also a dictionary element, the function Φ cannot be linear.

Properties of Element Selection: In general the signal to be analyzed, f, will be an observation of some data in the presence of some noise. That is,

$$f(t) = s(t) + n(t).$$

In fact, if we assume that we are observing stochastic processes defined on a measure space $(\Omega_0, A_0, \mathbf{P}_0)$, where the sample space $\Omega_0 = \{\omega_0\}$, then we can think of the signal as

$$f(t,\cdot) = s(t,\cdot) + n(t,\cdot),$$

and the signal construction is

$$\tilde{f}_m(t,\cdot) = \sum_{i=1}^{m} < f_i(t,\cdot), \Phi(f(t,\cdot),i) > \Phi(f(t,\cdot),i).$$

Taking this view in the case when $D_i = D \ \forall\, i$, we can define for each iteration i a probability distribution on the dictionary parameter space $\mathcal{I}$, $\mathbf{P}_{induced,i}(\cdot)$. Note that because of the one to one relationship between $\mathcal{I}$ and D, it is also a probability measure on D. Let $A \subset A_{\mathcal{I}}$ be a measurable subset of the parameter space. $\mathbf{P}_{induced,i}(A)$ is the probability that on iteration i, $\phi^{-1}(\phi_{\gamma_i}) \in A$. The probabilistic nature is solely due to the process being analyzed and it is assumed that the distribution on the dictionary parameter space is derived from the distribution of the stochastic processes. Define

$$\omega_0^i(A) = \{\omega_0 \in \Omega_0 : \phi^{-1}(\pi^1(\Gamma(D, f_i(t,\omega_0)))) \in A\},$$

or more simply

$$\omega_0^i(A) \;=\; \{\omega_0 \;\in\; \Omega_0 \;:\; \phi^{-1}(\Phi(f(t,\omega_0),i)) \;\in\; A\},$$

Then provided that $\omega_0^i(A)$ is a measurable subset of Ω_0, i.e. provided that $\omega_0^i(A) \;\in\; A_0$,

$$\mathbf{P}_{induced,i}(A) \;=\; \mathbf{P}_0(\omega_0^i(A)). \tag{2.3}$$

In this context, $\mathbf{P}_{induced,i}(A)$ is the probability that $\Phi(f(t,\omega_0),i) \;\in\; A$.

Again, we point out that the probability on the dictionary defined in equation 2.3 is induced by the probabilistic nature of the function being analyzed. One must be cautious here and recognize that when a specific observation, e.g. a realization of the stochastic process, is being analyzed, there is no probability involved. This is because the function which returns dictionary elements given the residues and dictionary, Γ, is a deterministic function. Furthermore, it is important to note that the decomposition operates on the signal f, and the fact that noise is an integral part of the observation means that the selection of elements will be influenced by the noise. This is a property of the Matching Pursuit. However, in most instances we desire a decomposition of s while ignoring n. Toward this end, we develop the Probabilistic Pursuit in which the selection of dictionary elements is made probabilistic to reflect our prior assumptions of signal and noise. In essence this means that even when given a specific realization to analyze, the selection of dictionary elements, i.e. the function Γ, is probabilistic, reflecting our assumptions (see Section 2.2).

Back to Matching Pursuit

We can view the Matching Pursuit as an instance of the general paradigm, which is characterized in terms of an optimality factor $0 \;<\; \alpha \;\leq\; 1$ and choice function $C : 2^{\mathcal{I}} \;\mapsto\; \mathcal{I}$, the purpose of which is to select in a deterministic way, one element from the set of dictionary elements satisfying an optimality criterion for each iteration. If $A \;\subset\; \mathcal{I}$, then

$$C(A) \;=\; \gamma, \quad \text{where } \gamma \;\in\; A.$$

Which particular γ is chosen depends on the specifics of the choice function. Let us use the same notation for the choice function operating on subsets of D. Thus, if $A \subset D$, then

$$C(A) = \phi_\gamma, \quad \text{where } \phi_\gamma \in A.$$

Here $D_i \equiv D$, so the dictionary for each iteration is the same. Then more formally, we have a signal $f = f_0 \in \mathcal{H}$. On every iteration i, we choose an element $\phi_{\gamma_i} \in D$ such that

$$| < f_i, \phi_{\gamma_i} > | \geq \alpha \sup_{\gamma \in \mathcal{I}} | < f_i, \phi_\gamma > |,$$

where $0 < \alpha \leq 1$ is the optimality factor. Actually, there may be many elements that satisfy the optimality criterion, so the set of these is defined,

$$D_{\alpha,i} = \{\phi_\gamma : \phi_\gamma \in D \text{ and } | < f_i, \phi_\gamma > | \geq \alpha \sup_{\gamma \in \mathcal{I}} | < f_i, \phi_\gamma > |\}, \tag{2.4}$$

Which one of them is chosen is given by the choice function C:

$$\phi_{\gamma_i} = C(D_{\alpha,i}).$$

This specifies the selection criterion in Γ (equation 2.1). Convergence, i.e. that $\| f_i \| \to 0$ as $i \to \infty$, is proven in Mallat and Zhang [18] and we summarize the concepts in the proof here.

For reference, two key Lemmas from Mallat and Zhang in the convergence proof of the Matching Pursuit are restated.

Lemma 1 *(Mallat, Zhang) Let* $h_n = < f_n, \phi_{\gamma_n} > \phi_{\gamma_n}$. *For any* $n \geq 0$ *and* $m \geq 0$

$$| < h_m, f_n > | \leq \frac{1}{\alpha} \| h_m \| \| h_n \| .$$

It is from within this Lemma that the optimality factor α enters the proof of convergence.

Lemma 2 *(Mallat, Zhang) If* $(s_n)_{n \in Z}$ *is a positive sequence such that* $\sum_{n=0}^{+\infty} s_n^2 \leq +\infty$, *then*

$$\lim_{n \to +\infty} \inf s_n \sum_{k=0}^{n} s_k = 0.$$

In the pursuit, since we are removing the contribution of the elements as we choose them, it seems reasonable that on each iteration, we are decreasing the energy in the input. This can be seen in the energy conservation statement [18]

$$\| f \|^2 = \sum_{i=0}^{m-1} | < f_i, \phi_\gamma > |^2 + \| f_m \|^2,$$

derived from

$$\| f_i \|^2 = | < f_i, \phi_\gamma > |^2 + \| f_{i+1} \|^2,$$

which is a consequence of the fact that we are removing projections from the data on each iteration. So we know that $\| f_i \|$ is monotonically decreasing with i and converges to some value. To show that it converges to zero, it is shown that $\{f_i\}$ is a Cauchy sequence and that the limit is orthogonal to the elements in the dictionary, which is complete.

From equation 2.2, it is seen that, given integers $N, M \geq 0$ with $N < M$, the residues in the decomposition obey

$$f_N = f_M + \sum_{i=N}^{M-1} < f_i, \phi_{\gamma_i} > \phi_{\gamma_i},$$

or stated in terms of h_n, defined in Lemma 1,

$$f_N = f_M + \sum_{i=N}^{M-1} h_i,$$

Thus, expansion of $\| f_N - f_M \|^2$ and subsequent use of Lemma 1 allows the derivation of

$$\| f_N - f_M \|^2 \leq \| f_N \|^2 - \| f_M \|^2 + \frac{2}{\alpha} \| h_M \| \sum_{n=N}^{M-1} \| h_n \|,$$

for non-negative integers N, M. For large enough N, M the difference between $\| f_N \|^2$ and $\| f_M \|^2$ will be negligible since $\| f_i \|$ converges. Then, by using Lemma 2 one can show that the last term in the above right hand size is also negligible for large enough N, M. The application of these ideas is not as straightforward as this, but the basic ideas are here and are used to show that $\{f_i\}$ is a Cauchy sequence. Then it is shown, by using energy conservation once again, that $\lim_{n \to +\infty} | < f_n, \phi_\gamma > | = 0$ for each γ, establishing that the limit of f_i is orthogonal to the elements of the dictionary.

We note some observations and generalizations of the result. In the statement of the problem, α is required to be positive for convergence.

Observation α need not be the same on every iteration. Consider $\{\alpha_n\}$, any positive sequence of α's, one for each iteration.

Claim 1 *Convergence is guaranteed as long as* $\inf_n \alpha_n > 0$.

Proof: One simply uses the infimum in Lemma 1.

2.2 Probabilistic Pursuit

Now we define a more general pursuit decomposition, which we call the Probabilistic Pursuit, fitting in the general paradigm, which incorporates prior knowledge in a probabilistic framework. We give meaning not only to the result of the search but to its ability to find matching elements. This ability will more precisely be characterized as the length of time it takes to complete the search. We retain some of the features of the Matching Pursuit, namely the optimality factor. The fundamental idea is that we allow the dictionaries to evolve as a Stochastic Process in each iteration. Effectively, for each i we probabilistically choose a sub-dictionary $D_i \in 2^D$ one element at a time according to a probability distribution, representing our assumptions about the signal structure, until the optimality criterion is met. It will be shown that the better our assumptions are, the faster we are able to find matching elements, and vice versa. We no longer have a choice function in the sense of the Matching Pursuit, however one is implied. In the development, the input f is deterministic and an element of

$\mathcal{H}$. We could also consider $f(t, \omega_0)$ to be a realization of a process on $(\Omega_0, A_0, \mathbf{P}_0)$. In this case, we assume that

$$\mathbf{P}_0(\{\omega_0 : f(t, \omega_0) \in \mathcal{H}\}) = 1.$$

2.2.1 Dictionary Process

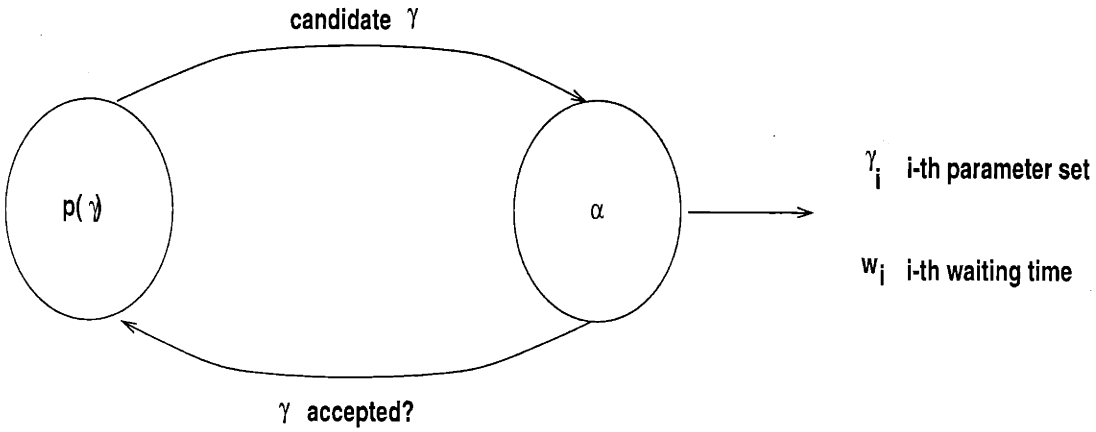

Figure 2-1: The process of choosing a dictionary element.

Formally then, we set $\Omega = \mathcal{I}$ and let $(\Omega, A_{\mathcal{I}}, \mathbf{P})$ be a probability space with $\mathbf{P}(\cdot)$ a probability measure on $\mathcal{I}$. This is the measure by which our algorithm operates. Though for the theoretical development here it is considered arbitrary, later it will in fact constitute prior knowledge. Let $p(\cdot)$ denote the density of $\mathbf{P}(\cdot)$. Recall that $\mathcal{I}$ is a space of parameters and that $\gamma \in \mathcal{I}$ is a vector of parameters. Consider a vector-valued i.i.d. discrete time stochastic process $\{\gamma_n(\cdot)\}$ defined on a measure space $(\Omega_\gamma, A_\gamma, \mathbf{P}_\gamma)$, where for fixed n, $\gamma_n(\cdot) : \Omega \mapsto \Omega$ is the random vector in $(\Omega, A, \mathbf{P})$ defined by the identity map. Since the sample space Ω is the parameter space of the dictionary, each $\gamma_n(\omega)$ is an element of $\mathcal{I} \equiv \Omega$, i.e. it is a random vector taking values in the parameter space $\mathcal{I}$ of the dictionary. For each iteration of the

decomposition, we take a realization of the process $\{\gamma_n(\cdot)\}$, and define the selected dictionary, which we call D_i, to be

$$D_i = \phi(\{\gamma_n(\omega_\gamma)\}).$$

In this case, the elements of D_i are ordered according to the time their parameters were chosen in $\{\gamma_n(\cdot)\}$. In the following, choosing an element means the same thing as choosing a parameter set. $D_{\alpha,i}$ is as in equation (2.4). Then the element chosen on a given iteration is defined as the first element in D_i that is also in $D_{\alpha,i}$. In figure 2.2.1 this is depicted as a communication between two agents. The first sends out candidate dictionary elements using the probability distribution, and the second tests whether or not the element meets the optimality requirements.

Let $\chi = \{\chi_0, ..., \chi_{m-1}\}$ be a partition of $\mathcal{I} \equiv \Omega$, and hence of the dictionary D, where each $\chi_i \in A_{\mathcal{I}}$. We define the **sub-dictionary process** associated with χ to be a discrete time i.i.d. m-ary stochastic process $\{\mathbf{d}_\chi(n,\cdot)\}$ on $(\Omega_d, A_d, \mathbf{P}_d)$ where each $\mathbf{d}_\chi(n,\cdot)$ is a random variable taking values in $\{0, ..., m-1\}$ with

$$\mathbf{P}_d(\mathbf{d}_\chi(n) = i) = \mathbf{P}(\chi_i) \quad \text{for } i \in \{0, ..., m-1\}.$$

We will use sub-dictionary processes in the following derivations.

The purpose of the following Lemma is simply to show that eventually, an element that meets the optimality criterion will occur.

Lemma 3 *$D_i \cap D_{\alpha,i}$ is non-empty w.p. 1 if $\phi^{-1}(D_{\alpha,i})$ is a measurable subset, $\phi^{-1}(D_{\alpha,i}) \in A_{\mathcal{I}}$, with $\mathbf{P}(\phi^{-1}(D_{\alpha,i})) \equiv b > 0$.*

Proof: Recall the definition of the set $D_{\alpha,i}$,

$$D_{\alpha,i} = \{\phi_\gamma : \phi_\gamma \in D \text{ and } |< f_i, \phi_\gamma >| \geq \alpha \sup_{\gamma \in \mathcal{I}} |< f_i, \phi_\gamma >|\}, \tag{2.5}$$

consisting of the set of elements satisfying the optimality criterion on iteration i. We want to show that at least one element of D_i satisfies the optimality criterion, or

in other words that there exists an n, where γ_n is the n^{th} element in the stochastic process generating D_i, such that $\phi(\gamma_n) \equiv \phi_{\gamma_n} \in D_{\alpha,i}$. We show that w.p. 1 this occurs by the following argument.

Let $\chi_1 = D_{\alpha,i}$ and $\chi_0 = D^c_{\alpha,i}$. Now $\chi = \{\chi_0, \chi_1\}$. Then the sub-dictionary process $\{\mathbf{d}_\chi(n,\cdot)\}$ is a Bernoulli sequence, modeled by $(\Omega_d, A_d, \mathbf{P}_d)$, with $\mathbf{d}_\chi(n) = 1$ corresponding to the event that $\phi_{\gamma_n} \in D_{\alpha,i}$ and $\mathbf{d}_\chi(n) = 0$ corresponding to the event that $\phi_{\gamma_n} \in D^c_{\alpha,i}$.

The sub-dictionary process is i.i.d., thus the event $\mathbf{d}_\chi(n) = 1$ is independent of $\mathbf{d}_\chi(m) = 1$ when $m \neq n$. Now $\sum_n \mathbf{P}_d(\mathbf{d}_\chi(n) = 1) = \sum_n b = \infty$. Then, the Borel - Cantelli lemma can be used to show that a 1 will occur w.p. 1 in $\{\mathbf{d}_\chi(n,\cdot)\}$, and thus the Lemma is true. $\diamond$

Lemma 4 *For each measurable subset of $\mathcal{I}$, $A \in A_\mathcal{I}$, let t_A be the waiting time for first occurrence of some element of A in the dictionary process $\{\gamma_n(\cdot)\}$. Then, t_A is well defined.*

Proof: If $\mathbf{P}(A) > 0$, then as before, define a sub-dictionary process with $\chi_1 = A$ and $\chi_0 = A^c$. Again, $\{\mathbf{d}_\chi(n,\cdot)\}$ is a Bernoulli random variable sequence with probability of success given by $b \equiv \mathbf{P}(A)$ for each element. Since a success will occur w.p. 1 in $\{\mathbf{d}_\chi(n,\cdot)\}$, the Lemma is true for this case.

If $\mathbf{P}(A) = 0$, then set $t_A = \infty$, and the Lemma is true in general. $\diamond$

We have now defined Γ, the function which explains the relationship of the decompositions elements from one iteration to the next. The probability distribution $\mathbf{P}$ will be referred to in various contexts as a **prior signal model** or **prior information**. Hence we have generalized the pursuit, which given a dictionary had worked only on correlation measurements with the observation in a deterministic fashion, to take into account prior information in a probabilistic framework. In our development, this prior distribution is independent of the actual observation signal, as well as of the iteration i. This however need not be the case, as discussed in Chapter 6. Furthermore, note the contrast with the probability given in equation 2.3, which is induced by the probabilistic nature of the observation. We compare the two in Section 2.2.2.

We develop the consequences of our formulation in two different contexts, representation and classification. If we are interested in efficient signal representation, i.e. with a small number of parameters capturing the salient characteristics, we can view the prior information as a way of finding these quickly. But, we do not want the assumptions that we have made about the structure of the input to prevent our finding important characteristics that are inconsistent with our prior knowledge. To ensure that we eventually find these, we can enforce the condition that the signal decomposition revert to a fall-back scheme when it is too difficult to find elements. On the other hand, it is the way in which the probability is distributed over D that will control, on any given input, the amount of time it takes to find matching elements. Thus, using the length of time it takes for the Probabilistic Pursuit to reduce the input signal energy, in addition to distance measures, we can construct a classification scheme.

Before continuing, we show that in our framework, the norm of the residues either goes to zero or at some iteration, the residue is inconsistent with the prior model.

Theorem 1 *Consider a Probabilistic Pursuit on $f(t,\cdot)$. With probability 1, either $\| f_i \| \rightarrow 0$ as $i \rightarrow \infty$ or for some iteration i, we have to wait infinitely long for a match.*

Proof: By assumption

$$\mathbf{P}_0(\{\omega_0 : f(t,\omega_0) \in \mathcal{H}\}) = 1.$$

To $f(t,\omega_0) \in \mathcal{H}$, we can apply the Probabilistic Pursuit. Then on every iteration, the second agent in the probabilistic search makes sure that the optimality criterion is satisfied. On a given iteration i, it is satisfied w.p. 1 on Ω_γ if $\mathbf{P}(\phi^{-1}(D_{\alpha,i})) > 0$ by Lemma 3. Previously, we saw that it was the fact that projections onto successively chosen dictionary elements were removed from the data and that the elements were chosen to satisfy the optimality criterion which guaranteed convergence. Hence the first part of the result. On the other hand, if for some i, $\mathbf{P}(\phi^{-1}(D_{\alpha,i})) = 0$, then the criterion will not be met. This means that the set of parameters corresponding

to the dictionary elements that match the residue f_i have zero measure with respect to $\mathbf{P}(\cdot)$. By Lemma 4, this means we would have to wait infinitely long for a match. $\diamond$

Let $\mathcal{P} = \{\mathbf{P}_i\}$ be the set of all the signal models, or distributions over the dictionary, that we wish to consider. Introducing some new terminology, call this set of signal models the super dictionary. We claim that the classification problem can be seen as a pursuit with respect to the super dictionary where the matching criterion is given in two parts. The first involves the waiting times for the atoms selected on each iteration whereas the second measures the distance of these from the signal model.

2.2.2 Representation

The goal is to analyze and represent the signal part of an input which we know contains signal and noise. If we denote the observation by $o(t)$, then

$$f(t) \equiv o(t) = s(t) + n(t),$$

where $s(t)$ and $n(t)$ indicate the signal and noise parts respectively. By specifying $\mathbf{P}$, we are making assumptions about $s(t)$. $\| f \|$ contains energy contributed from both the signal and noise parts. The Probabilistic Pursuit reduces $\| f_i \|$, but in a structured fashion, namely it seeks to reduce the energy due to the signal part before that of the noise part. The practical consequence of this is that here we expect that for any finite number of atoms chosen, our procedure produces more atoms that represent the signal as compared to the Matching Pursuit which does not differentiate between signal and noise processes. Put another way, to reduce the same amount of signal energy from our observation, we have do not have to carry the decomposition as far. As we will see this statement depends on the quality of the prior information we have.

We suggest with the following example to show that the framework is fairly flexible. One way to see the usefulness of this procedure then, is to consider the case where the signal part of the observation is one of a possible set or class of signals.

When the dictionary elements matching well at least one of the signals in the class are combined, it is a reasonable assumption that the result is not the entire dictionary. Assume that we know this combined set. We can think of it as an assumption about the particular signal part we see that has some error in it. Thus, perfect knowledge of the class can be seen as partial information about the signal.

To make these notions clear, we use the signal model to partition the dictionary in a meaningful way based on which we can make the arguments. Consider the following simple partition:

$$D^s \equiv \text{ Parts of } D \text{ where } p(\phi_\gamma) > \epsilon.$$

$$D^n \equiv \text{ Parts of } D \text{ where } p(\phi_\gamma) \leq \epsilon.$$

Note again that $p(\cdot)$ is used here as denoting the assumed density of $\mathbf{P}(\cdot)$ and that ϵ is a small positive number. This partition can be used to compare the probability in equation 2.3 with the prior distribution. That is, the probability of selecting a noise dictionary element can be defined in both cases. For equation 2.3, one simply integrates the probability over the noise dictionary parameters. Let $A \subset \phi^{-1}(D^n)$ such that $A \in A_{\mathcal{I}}$. Then using equation 2.3, the probability of selecting noise present in A is

$$\mathbf{P}_{induced,i}(A) = \mathbf{P}_0(\omega_0^i(A)).$$

For the prior model, recall from before that we defined $D_{\alpha,i}$, the elements that match the input signal on iteration i. Then, the critical information lies in the extent to which the elements that satisfy the optimality criterion match our assumptions of what is signal and noise. Their relative sizes give us information about the selection procedure.

The probability of choosing a signal element will be large when $D_{\alpha,i}$ overlaps significantly with D_s.

Theorem 2 *Let $\phi^{-1}(D^s)$ and $\phi^{-1}(D^n)$ be measurable, $\phi^{-1}(D^s), \phi^{-1}(D^n) \in A_{\mathcal{I}}$, and disjoint such that their union is equal to $\mathcal{I}$. Further, suppose $\phi^{-1}(D_{\alpha,i})$ is measurable, $\phi^{-1}(D_{\alpha,i}) \in A_{\mathcal{I}}$, and that $\mathbf{P}(\phi^{-1}(D^n)) = \delta$. Then for the signal model $\mathbf{P}(\cdot)$, we*

have that for each iteration i, the probability of selecting a dictionary element in D^n, $P_n(i)$, is bounded. We have,

$$P_n(i) \leq \frac{\delta}{i_n + i_s},$$

where,

$$i_s = \mathbf{P}(\phi^{-1}(D_i^s)),$$

$$i_n = \mathbf{P}(\phi^{-1}(D_i^n)),$$

$$i_o = \mathbf{P}(\phi^{-1}(D_i^o)),$$

where the sets D_i^s, D_i^n, and D_i^o are defined as,

$$D_i^s = D^s \cap D_{\alpha,i},$$

$$D_i^n = D^n \cap D_{\alpha,i},$$

$$D_i^o = D_{\alpha,i}^c.$$

Proof: The quantities $\phi^{-1}(D_i^s)$, $\phi^{-1}(D_i^n)$, and $\phi^{-1}(D_i^o)$ are well defined because of the measurability of $\phi^{-1}(D_{\alpha,i})$.

The set of waiting times $\{t_{A_i}\}_{A_i \in A_\mathcal{I}}$ may be an uncountable set, but given any finite number of subsets $\{A_i\}$, there is an ordering. These waiting times give an indication that the noise elements have a small probability of occurring. More precisely, for iteration i, the numbers i_s and i_n indicate the degrees to which $D_{\alpha,i}$ matches the signal and noise components, and as defined, they are actually the probability of occurrence for those elements in the dictionary process. The important quantity is

$$\text{Probability}(D_i^n \text{ occurs before } D_i^s) = \frac{i_n}{i_n + i_s}.$$

This is derived as follows. Consider the sub-dictionary process $\{\mathbf{d}_\chi(n,\cdot)\}$ on $(\Omega_d, A_d, \mathbf{P}_d)$ defined by $\chi = \{\chi_0, \chi_1, \chi_2\}$, where

$$\chi_0 = \phi^{-1}(D_i^s),$$

$$\chi_1 = \phi^{-1}(D_i^n), \text{ and}$$

$$\chi_2 = \phi^{-1}(D_i^o).$$

A dictionary is said to occur if an element from it occurs. That is, D_i^s occurs if $\mathbf{d}_\chi(n) = 0$, D_i^n occurs if $\mathbf{d}_\chi(n) = 1$, and D_i^o occurs if $\mathbf{d}_\chi(n) = 2$. Then,

$$\mathbf{P}_d(\mathbf{d}_\chi(n) = 0) = i_s,$$

$$\mathbf{P}_d(\mathbf{d}_\chi(n) = 1) = i_n, \text{ and}$$

$$\mathbf{P}_d(\mathbf{d}_\chi(n) = 2) = i_o.$$

Let $T_{s,1}$, $T_{n,1}$, and $T_{o,1}$ be the times for the first occurrences of the three dictionaries D_i^s, D_i^n, and D_i^o.

$$\text{Probability}(D_i^n \text{ occurs before } D_i^s) = \text{Probability}(T_{n,1} < T_{s,1})$$

$$\text{Probability}(T_{n,1} = i,\ T_{s,1} > i) = i_o^{i-1} i_n$$

$$\begin{aligned}
\text{Probability}(T_{n,1} < T_{s,1}) &= i_n \sum_{i=1}^{\infty} i_o^{i-1} && (2.6)\\
&= i_n \sum_{i=0}^{\infty} i_o^{i} && (2.7)\\
&= \frac{i_n}{1 - i_o} && (2.8)\\
&= \frac{i_n}{i_n + i_s} && (2.9)
\end{aligned}$$

By assumption, $\mathbf{P}(\phi^{-1}(D^n)) = \delta$. But,

$$\begin{aligned}
i_n &= \mathbf{P}(\phi^{-1}(D_i^n)) && (2.10)\\
&\leq \mathbf{P}(\phi^{-1}(D^n)) && (2.11)\\
&= \delta. && (2.12)
\end{aligned}$$

So,

$$\frac{i_n}{i_n + i_s} \leq \frac{\delta}{i_n + i_s}. \diamond$$

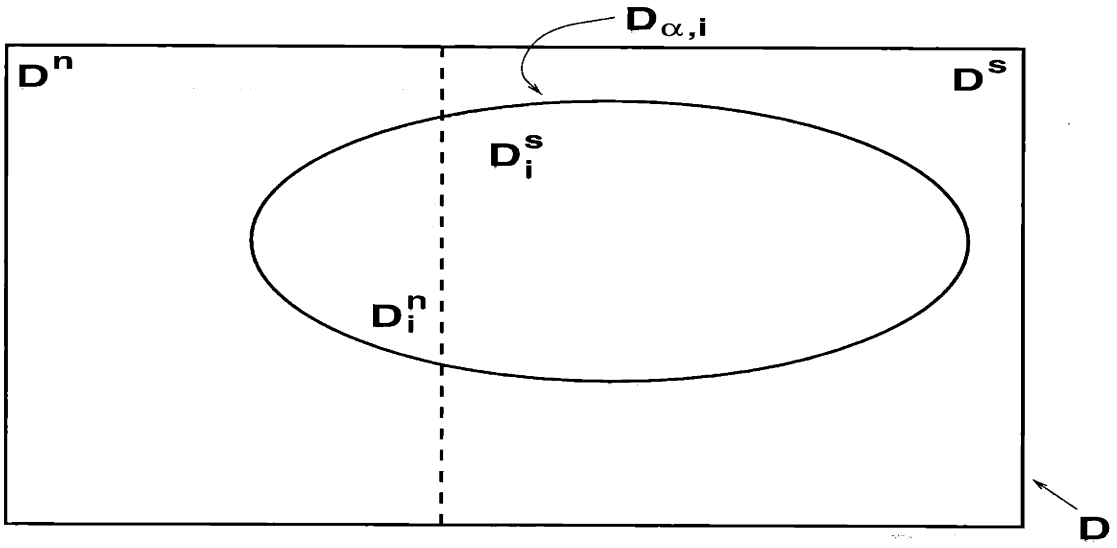

Figure 2-2: A sample partitioning of the dictionary.

These results can be made clear by considering the particular instance of the setting described above. Namely, we let D^s and D^n be defined as the sets preceding the theorem. In figure 2.2.2 these two sets are schematically represented as the two sections of the rectangle divided by the dashed line. D is the entire rectangle. The set $D_{\alpha,i}$ which represents the part of the dictionary that satisfies the optimality criterion is depicted as the oval. As seen in the theorem, the quantities of relevance are the intersections of this set with the partition of the dictionary, which are indicated by the inside portions of the oval separated by the dashed line. On any given iteration, the probability of choosing an element of the noise dictionary is given by the ratio

$$P_n(i) = \frac{i_n}{i_n + i_s},$$

the quantities being defined in the theorem. First we note that i_s will be large if

the set of elements that satisfy the optimality criterion overlap significantly with our assumptions, inherent in the prior signal model $\mathbf{P}(\cdot)$, of where the signal energy should be. This is because D^s is defined in terms of elements having a large probability of occurrence. On the other hand, even if the satisfying set overlaps significantly with D^n, i_n will be small because D^n has elements with low probability. Thus when our assumptions are correct we expect that i_s will be the dominant term in the denominator, and i_n will be small, giving a low probability of choosing a noise element.

Of course, this probability is a function of the iteration. Thinking again in terms of f being made up of contributions from the signal and noise parts, the probability of choosing noise should be small on the first iterations where there is a lot of signal energy. As this energy is reduced by the decomposition, the i_n term will become important, making it more likely to get a noise element. This supports our claim that the Probabilistic Pursuit reduces signal energy before that of noise. Then, noticing that i_s will be small when $D^s \cap D_{\alpha,i}$ is small and that $\mathbf{P}(\phi^{-1}(D^s)) = 1 - \delta$ shows that we will choose overlap or noise only if the correlation with the signal dictionary is small. Now let us rewrite $P_n(i)$ as $\mathbf{P}_i$(choosing noise element). The important thing to note here is that the probability in this case is a result of the probabilistic selection procedure operating on a deterministic function, as opposed to $\mathbf{P}_{induced,i}$(choosing noise element) which comes as a result of a deterministic procedure operating on stochastic processes.

Let the selection procedure used in deriving $\mathbf{P}_{induced,i}(\cdot)$ be that of the Matching Pursuit, i.e. one that deterministically returns a dictionary element satisfying the optimality criterion. Consider the following partition, where the signal being analyzed is $f(t,\omega_0) = s(t,\omega_0)$, i.e. a realization of $s(t,\cdot)$, defined on $(\Omega_0, A_0, \mathbf{P}_0)$, without noise (note that $\mathbf{P}_{induced,i}(\cdot)$ is a function of the iteration, but we have simplified this in the following for the sake of analysis). $\mathbf{P}_{induced,1}$ is assumed to have density $p_{induced,1}$.

$$D^s \equiv \text{ Parts of } D \text{ where } p_{induced,1}(\phi_\gamma) > \epsilon.$$

$$D^n \equiv \text{ Parts of } D \text{ where } p_{induced,1}(\phi_\gamma) \leq \epsilon.$$

But, since the observation is a realization of $f(t,\cdot) = s(t,\cdot) + n(t,\cdot)$ defined on $(\Omega_0, A_0, \mathbf{P}_0)$, the recovered dictionary elements by the Matching Pursuit will in general be different from D^s. However, with the Probabilistic Pursuit, if the prior model is given by $\mathbf{P}_{induced,1}(\cdot)$ we can correct for the presence of noise in the signal precisely because we look at dictionary elements according to $\mathbf{P}_{induced,1}(\cdot)$. That is, we note that $D_{\alpha,i}$ is dependent on the input signal f. On the first iteration, for example, $\mathbf{P}_{induced,1}(\phi^{-1}(D_{\alpha,1}))$ should be large. Here we directly see the effect that our procedure has of biasing the selection toward signal elements.

2.2.3 Classification

Hierarchical Modeling

The universe of signals is a finite set $\{u_i\}$. The observation will be one of these plus perhaps some noise. The goal of our classifier is to identify which u_i occurred. Actually the method presented below should be viewed as giving information at one level of an hierarchical classification scheme (see figure 2.2.3).

Each such level represents a covering of the space of signals with the lowest level corresponding to the coarsest description, i.e. the whole space, and the highest level to the singleton elements, or the finest covering. The nature of this covering depends on the universe of signals and on the dictionary used as well as the procedure used to classify the observation and the amount of noise it contains. Inherent in our use of a cover rather than a partition is the notion of ambiguity. At any given level of the hierarchy, one signal could be in more than one set of the covering. This implies that at that level, the given classification procedure and noise structure are such that there is not enough information extracted from an observation of the signal to make a choice between the classes, sets, of which it is a member. Furthermore, many signals can belong to the same class.

An intermediate level can be thought of as a covering of the data which is finer than the level below and coarser than the level above. At a particular level, classification means giving the set, or sets, in the cover that contains the observation. The higher up

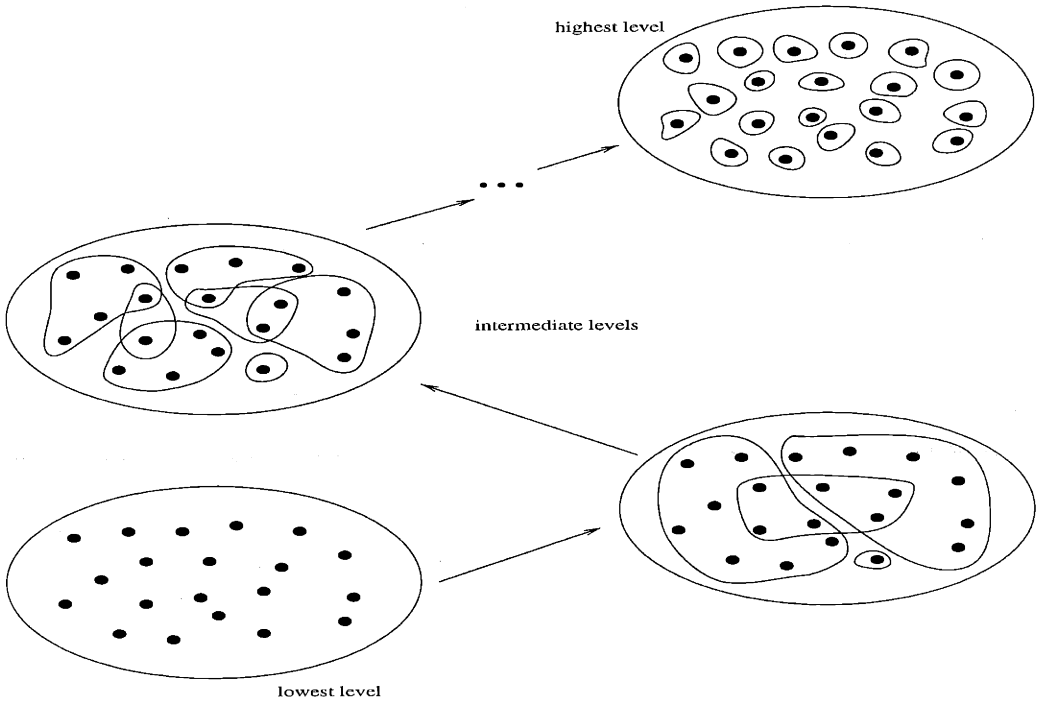

Figure 2-3: Hierarchical Classification: Each level represents a covering of the data with the lowest level giving the coarsest description and the highest level giving the finest.

in the hierarchy that we go, the more information is necessary in order to distinguish between the sets as they are more specific. In this context the base data set is the universe of signals, $\{u_i\}$ and at the lowest level, no distinction is made between the elements. Here, the different classes are identified with the elements of $\mathcal{P}$. Given an observation, the following classification scheme will identify a prior model, or set of prior models, matching it the best. That is, from the original set $\{u_i\}$, a set of subsets $\{\{u_{i_1}\}, \{u_{i_2}\}, \{u_{i_3}\}, ...\}$, are created where the indices take values in different, possibly overlapping, sets, and each, possibly empty set, corresponds to a prior signal model. To determine the structure of the covering, one would evaluate a set of test signals.

The advantage of the scheme can be seen by recognizing its parallelizability and considering the small number of parameters on which a decision is made. In fact, as will be seen in Chapter 5 Section 5.3.1 (Minimum-Time Decompositions), virtually

no post-processing of the decomposition is necessary in order to make a decision. On the other hand, in order to refine the decision and select a partition at some higher level, it is possible to take into consideration the correlations between the dictionary elements that are selected. However, to evaluate these correlations would require significant post processing, e.g. the Dynamic Programming scheme given in Section 6.2.2.

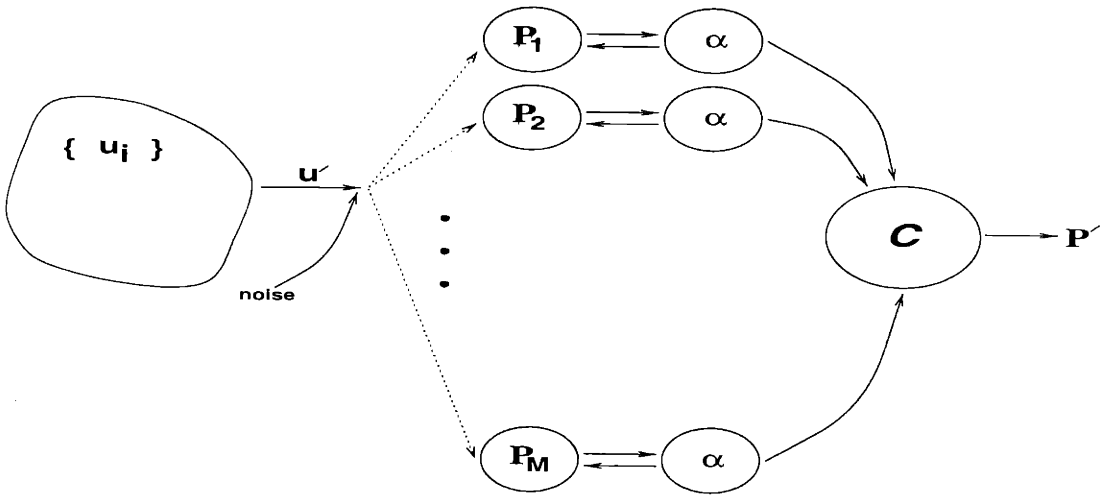

Figure 2-4: Description of the classification algorithm. The observation u' is one of the signals from the universe of possible signals. The classifier chooses a signal model $\mathbf{P}'$ based on a set of parallel decompositions.

As figure 2.2.3 shows, the classification procedure involves running a set of parallel decompositions, one for each signal model in $\mathcal{P}$, which will contain a model for each of the possible signals in $\{u_i\}$. It is the job of $\mathcal{C}$, in figure 2.2.3, to determine which model fits the observation the best. The definition of model fit is based on one of our novel contributions, namely that in the Probabilistic Pursuit we associate goodness of fit with speed of decomposition and the distance in parameter space of that decomposition from the signal model being used. The justification is given by the following.

Waiting Times

Theorem 3 *Let* $\mathbf{P}(\cdot)$ *be a signal model. Then given any input* $f \in L_2$*, the expected value of the waiting time for element selection* τ_s^i *is given by*

$$E[\tau_s^i] = \frac{1}{\mathbf{P}(\phi^{-1}(D_{\alpha,i}))}$$

for all iterations i*.*

proof: Immediate. $\diamond$

Again, we can look at figure 2.2.2. The average waiting time is controlled by the probability of occurrence of the set $D_{\alpha,i}$ in the stochastic process that gives the sub-dictionary for iteration i. If we assume for the moment that there is no noise, then in order to have low waiting times, we must give high probability to $D_{\alpha,i}$. But another way to describe this is to say that $D^s \cap D_{\alpha,i}$ must be large, or the assumptions inherent in the signal model must be consistent with what actually matches the input. On the other hand, large waiting times result when $D^n \cap D_{\alpha,i}$ is big, or when the dictionary elements that match the signal are the ones that we assume to be associated with noise. The sizes of these intersections are dependent on the iteration number i. This is important in that case where the assumptions are partially correct. Then, the appropriate way to view classification is to notice that large waiting times will be encountered before all of the energy in the input, $\| f \|^2$, is reduced.

But now assume that there is some noise. This noise may in fact have energy which overlaps with the signal part of the observation. In this case, if the correct signal model is used, the waiting times that result in the decomposition should be consistent with our earlier discussion. Another way to interpret this is to consider the case when the signal model is wrong. If there is noise energy where the signal model assumes there is signal energy, the decomposition may still have low waiting times, i.e. the waiting times in this case can indicate the wrong signal model.

Parameter Space Distances

We can define a distance measure in the parameter space of the dictionary and use this in conjunction with the waiting times to make decisions. The dictionary elements, or atoms, that the decomposition chooses are described in terms of a set of parameters, the possible values of which constitute $\mathcal{I}$. Each element of $\mathcal{P}$ is a probability measure on this same parameter space. So, we can define the parameter space distance,

$$d(\gamma, \mathbf{P}(\cdot)) \; : \; \mathcal{I} \times \mathcal{P} \; \mapsto \; R,$$

as a function which for example measures the distance from $\gamma \in \mathcal{I}$ to the nearest local maximum of the density associated with $\mathbf{P}(\cdot)$, which of course represents the assumptions of signal energy location inherent in the observation.

2.3 Finite Dimensional Signal Space

To close this chapter, we give the results in the finite dimensional case which is relevant to the experiments. We have a redundant dictionary described by $\mathcal{I}$. As in Mallat and Zhang we suppose that there exists a finite subset, indicated by $\mathcal{I}_\alpha$ such that $\sup_{\gamma \in \mathcal{I}_\alpha} | < f, \phi_\gamma > | \; \geq \; \alpha \sup_{\gamma \in \mathcal{I}} | < f, \phi_\gamma > |$ for $f \in L_2$. Since $\mathcal{I}_\alpha$ is finite the sup is really a max. Choose β such that $0 < \beta \leq 1$.

The dictionary process $\{\gamma_n(\cdot)\}$ defined on $(\Omega_\gamma, A_\gamma, \mathbf{P}_\gamma)$, resulting in D_i, is now the same as $\{\mathbf{d}_\chi(n, \cdot)\}$, a $|\mathcal{I}_\alpha|$-ary sub-dictionary process on $(\Omega_d, A_d, \mathbf{P}_d)$ where $\chi = \{\chi_i\}$ such that each χ_i is simply an element of $\mathcal{I}_\alpha$.

Analogous to equation (2.4),

$$D_{\beta,i} \; = \; \{\phi_\gamma \; : \gamma \in \mathcal{I}_\alpha \text{ and } | < f_i, \phi_\gamma > | \; \geq \; \beta \max_{\gamma \in \mathcal{I}_\alpha} | < f_i, \phi_\gamma > |\},$$

and the element chosen on a given iteration is defined as the first element in D_i that is also in $D_{\beta,i}$.

Claim 2 *If* $\mathbf{P}_d(\mathbf{d}_\chi(n) \; = \; i) \; > \; 0$ *for* $i \in \{1, ..., |\mathcal{I}_\alpha|\}$, *then the procedure above*

implies

$$\| f_i \| < \| f \| (1 - (\alpha\beta)^2\lambda^2)^{i/2}, \quad where\, 0 < \lambda = \inf_{f \in \mathcal{H}} \sup_{\gamma \in \mathcal{I}} \frac{| < f, \phi_\gamma > |}{\| f \|} \leq 1.$$

Proof: The effect of the probabilistic procedure is to modify the optimality factor from α to $\beta\alpha$. On replacing α by $\beta\alpha$ in the proof in Mallat and Zhang, the result follows. $\diamond$

So the convergence is still valid. The two theorems above will be valid as well if the dictionary is such that the sets $D_{\beta,i}$ are measurable with respect to the probability measure. Since $\mathcal{I}_\alpha$ is finite, this will be the case. This is in fact the setting which we will use for our experiments, where we will discretize and periodize the Gabor dictionary. If following [18] we included the DFT basis, it would result in a redundant, finite dictionary containing a basis. However, we choose to leave out the Fourier Basis because in the experiments to be described later, we are primarily interested in classification.

Chapter 3

Probabilistic Pursuit - Application to Time Frequency Analysis

In analyzing any real signal, we must deal with the presence of noise. The definition of noise can be simply any energy in the signal which is not contributed by the source of interest, or it can more subtly be defined as the characteristics of the source which are to be ignored in addition to extraneous noise. In light of this, the observation is most generally viewed as a *signal* part plus a *noise* part. In order to separate the two in an analysis, we must have a way to distinguish between these interesting and uninteresting parts of the signal, which for our purposes will depend on time and frequency. The interesting parts of the signal will be those that are present in a dictionary, D.

In the last chapter, we described a general class of iterative decomposition procedures for functions whose purpose was to select out the elements in the input observation which match the dictionary. A desirable property was that given any input, the procedure should be able to reconstruct it entirely, but at the same time, since elements are chosen successively, the first components chosen should be those that characterize the signal part best. A deterministic special case, the Matching Pursuit Decomposition, which possess the first of these properties was described. However, we saw that this procedure operated with no prior information with respect to noise and the selected components did not have the above ordering property. We defined

the Probabilistic Pursuit, a methodology in this broad class which operates with respect to a prior signal model in the form of a probability distribution on D. It has the property that given information about the signal part of the input, the components matching it are favored and found quickly, provided that the prior model is reasonably accurate. A bad signal model will still allow the input to be represented entirely, but the signal components will be found slowly. In practice we may have only rough information about the signal, and so the results are framed in a way that shows how performance should degrade as our information becomes weaker. It is precisely for this reason that we pursue the goal of ordered reconstruction of the signal, looking for signal elements first.

Toward the development of the application domain, which is speech classification, we present in this chapter the instantiation of the universal ideas of the last chapter in the context of time-frequency analysis. The development will be in two stages. First, we discuss analysis of deterministic signals which are elements of the Hilbert space L_2. From this we infer the kind of information that is extracted in the selection procedure. Then, to relate this specific kind of information to the signal models for the observation, we expand the analysis by considering the input to be a non-stationary stochastic process. In the next chapter, we develop the relationship of these processes to speech through the theory of speech production.

3.1 Pursuit Based Time Frequency Analysis on Deterministic Signals

3.1.1 Application Specific Definitions

Here, the developments will complement those of the last chapter since we choose a particular dictionary and associated selection criterion. By being specific about the dictionary to be used, we are making a decision about the application domain in the sense that the dictionary is meant to capture all of the variability that might be present in the input. By defining the selection criterion, we are defining a way to

quantify the amount of this variability in the input with the hope that it will be useful in finding salient characteristics. In this thesis, the application domain is speech and thus we are interested in time-frequency analysis.

The Gabor Dictionary

By performing the decomposition, we want to extract specific information from the signal. Signals such as speech are characterized by a time varying spectrum, and we might expect that we could characterize various speech sounds according to this variation. Thus, we choose as our dictionary, the Gabor dictionary which facilitates extraction of estimates of the signal energy in localized time and frequency bands. Such an analysis was proposed by Gabor [2]. We let D be the set of Gabor functions,

$$D = \{\phi_\gamma = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) e^{i\xi t}\}$$

each element of which is completely specified by an index set $\gamma = \{s, u, \xi\} = \{\text{scale, translation, frequency}\} \in R^+ \times R^2 = \mathcal{I}$. The window function $g(t)$ is a Gaussian,

$$g(t) = 2^{1/4} e^{-\pi t^2}.$$

The constant is chosen so that the L_2 norm of $g(t)$ equals 1. An important property of this dictionary is that the Gaussian is an optimal window for a particular time frequency tradeoff, because in this case the product of the variances of the window and its Fourier transform meets with equality the inequality given in the next section.

Selection Criterion

In the Hilbert space framework that we use, the matching criterion used to select dictionary elements on any given iteration is based on the magnitude of the inner product of the input $o(t) = f(t)$ and the dictionary elements,

$$< o(t), \phi_\gamma(t) > = \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} o(t) g(\frac{t-u}{s}) e^{-i\xi t} dt.$$

This is a windowed, or Short Time Fourier analysis. We elucidate the properties of a selection procedure based on this matching criterion in the following discussion.

Time Frequency Analysis

In this thesis we will be dealing almost exclusively with functions in L_2, and for these the Fourier transform, which is also in L_2, is defined as

$$F(\xi) = \int_{-\infty}^{+\infty} f(t)e^{-i\xi t}dt, \quad f \in L_2.$$

The Hilbert space structure of L_2 will be important and we denote the inner product on this space as

$$< f, g > = \int_{-\infty}^{+\infty} f(t)\overline{g}(t)dt, \quad f, g \in L_2.$$

The norm is then defined as

$$\| f \| = \sqrt{< f, f >}, \quad f \in L_2.$$

Time-Frequency analysis [17] [7] is subject to an uncertainty constraint that precludes arbitrary knowledge of the time varying energy distribution of any process. Given a function $w(t) \in L_2$ and its Fourier transform $W(\xi)$, consider the functions $|w(t)|^2$ and $\frac{1}{2\pi}|W(\xi)|^2$. Let the second moment about the mean, the variance, be defined as

$$V(f(t)) = \int_{-\infty}^{+\infty} (t - M(f(\cdot)))^2 f(t)dt,$$

where

$$M(f(t)) = \int_{-\infty}^{+\infty} tf(t)dt.$$

Then if $\| w \| = 1$, $V(|w(t)|^2)V(\frac{1}{2\pi}|W(\xi)|^2) \geq \frac{1}{4}$ [17].

To understand the implications for signal analysis we must consider the problem of extracting local information from a signal $o(t)$, which may be defined for all time $t \in R$. If we choose a time t_0 and ask about the energy in the signal at a specific frequency ξ_0, we cannot get the answer from $O(\xi)$ because it is a measure of the

frequency content for all time. So we may try to analyze $o(t)$ in a neighborhood of t_0. That is, we limit the time extent of the signal by multiplying it with a window function $w(t)$ with mean t_0. But this is the same as convolving $O(\xi)$ and $W(\xi)$. The more that we try to localize around t_0, which corresponds to making the variance of $|w(t)|^2$ smaller and smaller, the bigger the variance of $\frac{1}{2\pi}|W(\xi)|^2$ must become. The net effect is to increasingly blur $O(\xi)$. Likewise, sharpening the frequency analysis would correspond to increasing the variance of $|w(t)|^2$, which indicates a loss of time localization.

So if we have a set of times and frequencies, or more generally the entire time-frequency plane, can we assign numbers to the points (t, ξ) that uniquely describe $o(t)$ and moreover what do these numbers tell about the local properties of the signal? This question does not have a unique answer. In essence, there is a multiplicity of answers, each of which corresponds to a set of time-frequency resolution tradeoffs that one is allowed to make. It is the specific task which often motivates the tradeoffs to be made. A constant, even window yields a Short Time Fourier Transform. An analysis based on the time dilations of certain functions yields a Wavelet Transform.

Looking at it a little more carefully, we see that to each point (t, ξ) in the time frequency plane, we associate a window function whose purpose is to capture local information about the signal energy near (t, ξ). Our choice of window parameters will indicate a neighborhood around (t, ξ) and in some sense we can think of this window function as covering that neighborhood. Informally, if we choose points $\{(t_i, \xi_i)\}$ and their associated windows so that we cover the entire time frequency plane, we know that we can assign numbers to each point in $\{(t_i, \xi_i)\}$ such that we may reconstruct our observation from this data. There are many possible coverings. For a more precise discussion of the preceding, we introduce the concept of frames.

Define $\{\phi_j\}$, $j \in J$ where J is a possibly infinite index set, to be a sequence in a Hilbert space $\mathcal{H}$. Following [17], consider the signal transformation defined by the operator $T: \mathcal{H} \mapsto l_2(J)$, where for $f \in \mathcal{H}$,

$$Tf(j) = < f, \phi_j > \quad \forall\, j \in J.$$

The adjoint of T is T^*, defined by

$$T^*c(n) \;=\; \sum_{j \in J} c(j)\phi_j, \;\; \forall\, c(n) \;\in\; l_2(Z).$$

Then the pseudo inverse is

$$\hat{T}^{-1}c(n) \;=\; (T^*T)^{-1} \sum_{j \in J} c(j)\phi_j$$

$\{\phi_j\}$ is a frame for $\mathcal{H}$ if there exist $A, B \;>\; 0$ (a *tight* frame when $A = B$) such that for all $f \;\in\; \mathcal{H}$,

$$A \parallel f \parallel \leq \sum_{j \in J} | < f, \phi_j > |^2 \;\leq\; B \parallel f \parallel .$$

Then, f is reconstructed as $\hat{T}^{-1}Tf \;=\; \sum_{j \in J} < f, \phi_j > \hat{\phi}_j$, where $\hat{\phi}_j \;=\; (T^*T)^{-1}\phi_j$. The set $\{\hat{\phi}_j\}$, with $j \;\in\; J$ is called the dual frame. The key point is that a signal transformation can be viewed of as a map from $\mathcal{H}$ into $l_2(J)$ and the pseudo inverse of this map, which is well defined because of the frame condition, tells us how to reconstruct any function in $\mathcal{H}$ from its image in $l_2(J)$.

When we multiply our observation by a window and then take its Fourier transform, the equation that we get is

$$c \;=\; \int_{-\infty}^{+\infty} o(t)w(t)e^{-i\xi t}dt.$$

However, if we let $\phi \;=\; w(t)e^{i\xi t}$ and make sure that w(t) is square integrable, then this is really the same as $c \;=\; < o(t), \phi(t) >$. Accepting the fact that we cannot measure an instantaneous frequency at an instant in time, we talk about localized measurements in both time and frequency. But now we have a way of showing that these measurements are good enough for unique representations. Instead of computing numbers for each point in the time-frequency plane, we can choose a countable subset, $\{(t_i, \xi_i)\}$. In fact, the Short Time Fourier Transform (STFT) and the Discrete Wavelet Transform (DWT) are simply the result of choosing a two dimensional grid and assigning windows to each point in two different ways. Both of

these are regular in the sense that the grids are defined by two parameters $a, b > 0$. For the STFT,

$$\phi_{n,m}(t) = g(t - na)e^{imbt}, \quad m, n \in Z, \tag{3.1}$$

whereas for the DWT,

$$\phi_{n,m}(t) = \frac{1}{\sqrt{b^m}} g(\frac{t - nab^m}{b^m}), \quad m, n \in Z. \tag{3.2}$$

The corresponding grids are $\{(na, mb)\}_{n,m \in Z}$ and $\{(nab^m, b^m)\}_{n,m \in Z}$. The first component is easily seen to be a discretization of the time axis. Only in the STFT case is the second component a discretization of the frequency axis. Nevertheless, the second component for the DWT case is inversely related to the frequency variable and we can still interpret the grid as a sampling of a time-1/frequency plane. There are theorems which give conditions on $g(t)$, a, and b such that the resulting ϕ_i constitute a frame of L_2.

The STFT and DWT techniques imply that the same window functions can be used to represent any function in L_2. But a frame is, in general, different than a basis. It can be quite redundant in fact, depending on the bounds A and B. So a natural question is whether or not we need to compute the inner products with all of the frame elements and if not, how do we choose the elements? Here, we no longer deal with a pseudo inverse, but rather we generate a sequence of approximations in the space spanned by the $\{\phi_i\}$ which converges to a given $f \in L_2$. The Matching Pursuit [18] generates these approximations iteratively, where at each stage a ϕ_i is chosen based on maximizing its L_2 correlation (inner product magnitude) with the signal. This transformation of the signal is non-linear in the sense that expansion functions are not chosen before hand and will in general be different for different signals. Since we can view a frame as a dictionary in L_2, the Matching Pursuit, using a dictionary derived from the STFT or Wavelet grid (see equations 3.1 and 3.2), is a way to choose one at a time, points in a two dimensional plane, with an associated window, tailored to the signal being decomposed. One is choosing the elements of the dictionary to match the signal being analyzed. The Probabilistic Pursuit uses this

information in the form of a matching criterion in addition to *a priori* information represented by the prior signal model or probability distribution in the search.

Now, we discuss the advantages of an adaptive time-frequency representation over those derived from a grid, like the STFT (Short Time Fourier Transform) and DWT (Discrete Wavelet Transform). In an adaptive representation, the location and size of the window function in time, and hence frequency, is determined over the course of the decomposition, unlike the case of the STFT or the DWT. In the STFT, the window size remains constant, but the location of the analysis is determined by a grid in the time-frequency plane. For the DWT, the time location and size of the wavelet is given by a grid on a time-scale plane and in this analysis the frequency analyzed is related to the scale with larger scales implying analyses of lower frequencies. But this notion is made clearer if we consider the following wavelets [23]:

$$\psi_{m,n}(t) \;=\; 2^{m/2} e^{2\pi i(2^m t - n)}, \;\; 0 \;\leq\; t \;\leq\; \frac{1}{2^m}.$$

Here the scale parameter directly gives the frequency that is analyzed and the duration of the window, and thus the two are not independent.

The quality, or resolution, of the analysis in time and frequency is determined by the window size. A large window in time gives good frequency resolution but poor time resolution, a small window the reverse. We saw that it is not possible to have arbitrary resolution in both domains at once as a consequence of the uncertainty principle. With this in mind, we note that each STFT measurement is characterized by an identical resolution in time as well as in frequency, and thus these quantities are not a function of the grid location of the analysis. On the other hand, the time and frequency resolution in the DWT analysis using $\psi_{m,n}(t)$ is a function of the grid location of the measurement. In this case, a large positive m means a small duration and large frequency. A large negative m means the opposite. Thus from our discussion showing that large time windows give good frequency resolution we can make the statement that this Wavelet analysis will have good frequency resolution at low frequencies and vice versa. However, we must also make the statement that this

analysis will have poor time resolution at low frequencies and vice versa.

Thus there is a tradeoff between time and frequency resolution. In both the STFT and the DWT above, these tradeoffs are made before any analysis occurs and is completely independent on the signal to be analyzed. The various Pursuits discussed in this thesis are fundamentally different in that the tradeoffs are made on-line as the decomposition is evolving. In the Matching Pursuit, the selection criterion decides which window and modulating function to use based only on correlations with the signal. In the Probabilistic Pursuit, the selection criterion uses this information in conjunction with *a priori* information in the form of a prior signal model. Thus these representations are better in that the tradeoffs are made according to an optimality criterion. In this thesis, maximal signal energy was used as the basis for the selections. The effect is to produce a compact representation containing most of the signal energy, whereas with the STFT and DWT above, the signal energy is spread out over all of the elements and a further search would be required to determine significant components.

3.2 Pursuit Based Time Frequency Analysis on Stochastic Processes

3.2.1 Instantiating the Universal Results

The purpose of the previous discussion was to relate the selection procedure in the Probabilistic Pursuit to the general field of time-frequency analysis. Having made this connection, we would now like to use the results of the last chapter in the time-frequency context. More specifically, in the previous chapter we generalized the dictionary selection procedure to include *a priori* probabilistic information. Here, we generalize the discussion in a different direction by considering the prior signal model to be a model for a stochastic process which generates the input. Having done this, we will develop probability distributions on the Gabor dictionary, and then later discuss ways to partition the dictionary.

3.2.2 Stochastic Model for Input

In order to gain some insight we consider the class of semi-stationary oscillatory processes [22], appropriate in modeling sources whose energy distribution over frequency is changing slowly with time. We then approach the signal analysis problem as the analysis of a realization of such a stochastic process. Having defined this framework, we can see explicitly what our methods are measuring.

Let $X(t) = X(t, \omega_0)$, $\omega_0 \in \Omega_0$, $t \in T$, be a non-stationary stochastic process that is 0-mean and has finite variance. Also, we will assume that $X(t, \cdot) \in L^2$ w.p. 1.

Oscillatory Processes

The heart of what is important for our analysis is a two dimensional function, called the evolutionary spectral density, which for each time and each frequency specifies the relative strength of the process. In order to show how this function is related to a non-stationary stochastic process, we must look at the definition of an oscillatory process.

Let $\mathcal{F} = \{A_t(\xi)e^{i\xi t}\}$ be a set of functions indexed by ξ, such that for each ξ, $A_t(\xi)e^{i\xi t}$ is a function of time t. Define the spectrum of $A_t(\xi)$ with respect to t to be $dK_\xi(\omega)$ so that

$$A_t(\xi) = \int_{-\infty}^{+\infty} e^{it\omega} dK_\xi(\omega).$$

If for each ξ, $|dK_\xi(\omega)|$ has an absolute maximum at the origin, $\omega = 0$, then $\mathcal{F}$ is called a family of **oscillatory functions**. Each $A_t(\xi)e^{i\xi t}$ has a Fourier transform centered around ξ.

An **oscillatory process** $X(t, \cdot) \equiv X(t)$ (Priestley) is defined as one for which there exists a family of oscillatory functions $\mathcal{F}$ and a measure $\mu(\xi)$ on the real line such that the covariance kernel of the process can be written as

$$E[\overline{X}(t_1)X(t_2)] = \int \overline{A_{t_2}}(\xi)A_{t_1}(\xi)e^{i\xi(t_2-t_1)}d\mu(\xi). \tag{3.3}$$

Then, the process can be represented by

$$X(t) = \int A_t(\xi) e^{i\xi t} dZ(\xi) \tag{3.4}$$

where $d\mu(\xi) = E[|dZ(\xi)|^2]$ and $dZ(\xi)$ has orthogonal increments. Thus an oscillatory process has a time varying spectral representation.

For any given oscillatory process $X(t)$, the family $\mathcal{F}$ may not be unique and in general there will be a collection $\mathcal{C}_X = \{\mathcal{F}_1, \mathcal{F}_2, ...\}$ of families in terms of which $X(t)$ has a representation of the form given above.

Evolutionary Spectral Density

The evolutionary power spectrum is defined at time t as $dH_t(\xi) = |A_t(\omega)|^2 d\mu(\xi)$. This is a two-dimensional function which characterizes the time varying energy distribution over frequency. In this thesis, $\mu(\xi)$ is assumed to be absolutely continuous w.r.t. Lebesgue measure. Thus we can write $dH_t(\xi) = |A_t(\omega)|^2 d\mu(\xi) = h_t(\xi) d\mu_L(\xi)$, where $h_t(\xi)$ is a density on R^2, the evolutionary spectral density, and $\mu_L(\xi)$ is the one-dimensional Lebesgue measure ($\mu_L(\cdot)$ in $d\mu_L(\cdot)$ is replaced with the appropriate integration variable when needed). We assume therefore, that $H_t(\xi)$ is a measurable function on R^2 which has a density with respect to $\mu_L \times \mu_L$. Further, it is assumed that

$$\int \int h_t(\xi) d\xi dt < \infty.$$

Since $\int h_t(\xi) d\xi$ gives the power in the process at time t, the above integral is the total energy over all time, and we require it to be finite. The equations for discrete time oscillatory processes are essentially the same except that the frequency integrals go from $-\pi$ to $+\pi$.

As previously mentioned, there may be a number of families, $\mathcal{C}_X = \{\mathcal{F}_1, \mathcal{F}_2, ...\}$, in terms of which the process can be defined. Since the evolutionary spectral density $h_t(\xi)$ is defined by the family, each $\mathcal{F}_i$ will define a separate density $h_t^{\mathcal{F}_i}(\xi)$. When we restrict our attention to the class of semi-stationary oscillatory processes, defined below, we will see that this non-uniqueness is not significant from the point of view

of signal analysis.

Semi-Stationary Oscillatory Processes

These processes have a characteristic width W_X associated with them which indicates the length of time over which the process can be viewed as stationary.

$$W_X \;=\; sup_{\mathcal{F}\,\epsilon\,\mathcal{C}} \left\{ \left[sup_{\xi} \{ \int_{-\infty}^{+\infty} |\omega| |dK_{\xi}^{\mathcal{F}}(\omega)| \} \right]^{-1} \right\},$$

where $K_{\xi}^{\mathcal{F}}(\omega)$ is the spectrum of $A_t(\xi)$, and $\{A_t(\xi)\} \;=\; \mathcal{F}$.

Generally, there will be a subset of $\mathcal{C}_X$ which contains the families with the largest characteristic width. Then let us simply define $\mathcal{C}^X$ to be this subset. Though it is true that an evolutionary spectral density can be defined for each element $\mathcal{C}^X$, the important point is that measurement of the process characteristics will not reveal the differences between them. In the discussion of time-frequency analysis, we noted that the operations of localizing in time and frequency blurred the characteristics of the signal. The following discussion will show that when we analyze semi-stationary oscillatory processes using the Gabor dictionary and associated matching criterion, there is an analogous blurring and error introduced into the measurement. Thus, though from a representational point of view, the covariance of the process may have a number of representations, each associated with a different evolutionary spectral density, from an analysis point of view, the localized measurements obtained will not allow us to distinguish between them.

The measurements we make do not depend significantly on our assumptions regarding the family. In the discussion following equation 3.11 we will see that regardless of which family in $\mathcal{C}^X$ is used to represent the process, the expected measurements do not differ significantly.

3.2.3 Selection Criterion on Semi-Stationary Input

Our method of analyzing the realizations of a process is to compute their inner products with elements in our dictionary. Let us now analyze the implications. We will

use the Short Time Fourier Transform as our inner product model, because as seen previously, when we use a Gaussian function as our window, this implies that we are decomposing over a Gabor dictionary.

Let $O(t)$ be a semi-stationary oscillatory process with family $\mathcal{C}^O = \{\mathcal{F}_1, \mathcal{F}_2, ...\}$ and the associated set of evolutionary spectral densities $\{h_t^{\mathcal{F}_1}(\xi), h_t^{\mathcal{F}_2}(\xi), ...\}$.

Continuous Time Short Time Fourier Transform

Let $o(t)$ be the observed signal, a realization of $O(t)$. The STFT is a two dimensional function of time and frequency,

$$S(u,\xi) = \int_{-\infty}^{+\infty} o(t)w(t-u)e^{-i\xi t}dt.$$

But if we let $w(t)$ be a normalized Gaussian, then also,

$$\begin{aligned} S(u,\xi) &= < o(t), \phi_\gamma(t) > & (3.5)\\ &= \frac{1}{\sqrt{s}}\int_{-\infty}^{+\infty} o(t)g(\frac{t-u}{s})e^{-i\xi t}dt. & (3.6)\\ & & (3.7) \end{aligned}$$

Windowed Analysis of Oscillatory Processes

When we substitute $O(t)$ for $o(t)$ in the above, we get

$$\begin{aligned} < O(t), \phi_\gamma(t) > &= \frac{1}{\sqrt{s}}\int_{-\infty}^{+\infty} O(t)g(\frac{u-t}{s})e^{-i\xi t}dt \quad \text{by symmetry} & (3.8)\\ &= \frac{1}{\sqrt{s}}\int_{-\infty}^{+\infty} O(u-t)g(\frac{t}{s})e^{-i\xi(u-t)}dt. & (3.9)\\ & & (3.10) \end{aligned}$$

The last expression is a convolution which is a function of γ, and the expression on the right hand side shows where the three parameters enter. The shift by u indicates the location in time around which the analysis will occur. The scale s determines the size of the window, and the frequency ξ locates the frequency neighborhood of

interest. So in particular, the window function is centered at the origin. Now,

$$\int_{-\infty}^{+\infty} |t||g(\frac{t}{s})|dt \;=\; \frac{s^2}{\pi}.$$

Let $G(\xi)$ be the Fourier transform of $g(\frac{t}{s})$, so that

$$G(\xi) \;=\; sg(s\xi).$$

Then, from Priestley [22] we have the following, for $\mathcal{F}_i \in \mathcal{C}^O$.

$$E[\,|<O(t),\phi_\gamma(t)>|^2\,] \;=\; \int_{-\infty}^{+\infty} |G(w)|^2 dH_u^{\mathcal{F}_i}(w+\xi) \;+\; \mathbf{O}(\frac{s^2}{\pi W_O}), \qquad (3.11)$$

or,

$$E[\,|<O(t),\phi_\gamma(t)>|^2\,] \;=\; \int_{-\infty}^{+\infty} |G(w)|^2 h_u^{\mathcal{F}_i}(w+\xi)dw \;+\; \mathbf{O}(\frac{s^2}{\pi W_O}),$$

where $\mathbf{O}(x)$ goes to 0 with x.

We note now that $\int_{-\infty}^{+\infty} |G(w)|^2 h_u(w+\xi)dw$ is a smoothed estimate of $h_t(\xi)$. This is because $|G(w)|^2$ is centered at $w \;=\; 0$ with a variance proportional to $\frac{1}{s^2}$. The bigger s is the better the estimate of the evolutionary spectral density at frequency ξ. However, this means that the window function is increasing in length because its variance is proportional to s^2.

To expand on the consequences of a non-unique representation, we consider the above expression for all members of $\mathcal{C}^O$. The two quantities $E[\,|<O(t),\phi_\gamma(t)>|^2\,]$ and $\mathbf{O}(\frac{s^2}{\pi W_O})$ are independent of which particular family in $\mathcal{C}^O$ is used. Furthermore $E[\,|<O(t),\phi_\gamma(t)>|^2\,]$ is a fixed value. Thus the difference between $\int_{-\infty}^{+\infty} |G(w)|^2 h_u^{\mathcal{F}_i}(w+\xi)dw$ and $\int_{-\infty}^{+\infty} |G(w)|^2 h_u^{\mathcal{F}_j}(w+\xi)dw$ is only $\mathbf{O}(\frac{s^2}{\pi W_O})$. Therefore, it is of no practical consequence which family is used in the representation.

Moreover, the important point is that $|<o(t),\phi_\gamma(t)>|^2$ is an approximately unbiased estimator of the smoothed evolutionary spectral density up to an error term that is $\mathbf{O}(\frac{s^2}{\pi W_O})$. Thus we see that our selection criterion searches on average for peaks

in the evolutionary spectral density. We analyze the consequences of this result in the following section.

3.3 Probabilistic Pursuit and Non-Stationary Signal Analysis

We will assume that the observed signal is $f(t) \equiv f(t,\omega_0) = s(t,\omega_0) + n(t,\omega_0)$, all of the processes being defined on a common probability space $(\Omega_0, A_0, \mathbf{P}_0)$ with ω_0 a sample point. $s(t)$ and $n(t)$ are oscillatory, so we also have $\mathcal{F}_s = \{A_t^s(\xi)e^{i\xi t}\}$ and $\mathcal{F}_n = \{A_t^n(\xi)e^{i\xi t}\}$. Though we deal practically with deterministic signals, they are assumed to be realizations of a well defined process. Corresponding to $\mathcal{F}_s$ and $\mathcal{F}_n$, we have the evolutionary spectral densities $h_t^s(\xi)$ and $h_t^n(\xi)$. We choose an interval, say $[0,T]$, on which to decompose the signal.

3.3.1 Defining a Probability Measure: Prior Signal Model

Prior knowledge in our framework is embodied in the probability distribution on the dictionary that we use in the decomposition. In the discussion above, we identified the signals that might be observed with evolutionary spectral densities of the stochastic processes that produce them and went on to show that the Probabilistic Pursuit, in its search for salient characteristics, looks for the peaks of these functions. In the following discussion, we derive the distribution on the dictionary based on this density, which will be used to partition our dictionary D, on $[0,T]$, into D^s, D^n, and D^o, the signal, noise, and other dictionaries. These result by observing first that $dH_t^s(\xi) = h_t^s(\xi)d\mu_L(\xi)$ on $[0,T]$.

Let us make a few more definitions.

$$Z_{h^s} \equiv \int\int h_t^s(\xi)d\xi dt < \infty,$$

and

$$Z_{h^n} \equiv \int\int h_t^n(\xi)d\xi dt < \infty.$$

Using these normalization constants, define the densities

$$p_s(t,\xi) = \frac{1}{Z_{h^s}} h_t^s(\xi),$$

and

$$p_n(t,\xi) = \frac{1}{Z_{h^n}} h_t^n(\xi),$$

from which we defines the measures

$$\mathbf{P}_s(A) = \int\int I_A p_s(t,\xi)d\xi dt,$$

and

$$\mathbf{P}_s(A) = \int\int I_A p_s(t,\xi)d\xi dt,$$

where I_A is the indicator function of $A \subset \mathcal{B}^2$, the Borel sets in R^2. Define a probability measure on $\mathcal{I} = \mathcal{R}^+ \times R^2$ by taking the density of a distribution on R^+, for example the exponential distribution on R^+, and combining it separately with p_s and p_n so that the marginal distributions on R^2 are given by $\mathbf{P}_s(\cdot)$ and $\mathbf{P}_n(\cdot)$. The resulting measures are prior signal models for $s(t)$ and $n(t)$.

3.3.2 Partitioning the Dictionary

Rather than using the probability distribution, we first partition the dictionary using the evolutionary spectral densities to show more concretely the contents of the noise and signal parts.

$$O_s = \{(t,\xi) : h_t^s(\xi) > 0) \text{ and } t \in [0,T]\}$$

which is the support on the time-frequency plane, restricted to $[0,T]$, for the evolutionary spectral density $h_t^s(\xi)$ and

$$O_n = \{(t,\xi) : h_t^n(\xi) > 0) \text{ and } t \in [0,T]\},$$

the support on the time-frequency plane, restricted to $[0,T]$, for the evolutionary spectral density $h_t^n(\xi)$. Let

$$O = O_s \cap O_n,$$

which is the set of time-frequencies that the signal and noise have in common. Choose $\epsilon > 0$. Let O_ϵ be the subset of O such that $h_t^n(\xi) < \epsilon$ for $(t,\xi) \in O_\epsilon$. On O_ϵ, we will approximate the signal as being noise free. This is justified as follows.

Fix t_0. Let $\hat{h}(\xi)$ be a bandpass filter with frequency support given by the set of frequencies in $\{\xi : (t_0,\xi) \in O_\epsilon\} = O_{\epsilon,t_0} \cap O_{band}$, where $\mu_L(O_{band}) = C_{band}$ is small but otherwise O_{band} is an arbitrary measurable subset of R. Consider a stationary noise with spectral density $h_{t_0}^n(\xi)$. If this noise is input to the filter, the variance (energy) of the output would be

$$\begin{aligned} \int_{O_{\epsilon,t_0} \cap O_{band}} h_{t_0}^w(\xi) d\xi &< \int_{O_{\epsilon,t_0} \cap O_{band}} \epsilon d\xi & (3.12) \\ &= \epsilon \mu_L(O_{\epsilon,t_0} \cap O_{band}) & (3.13) \\ &\leq \epsilon C_{band}. & (3.14) \end{aligned}$$

Let

$$D^s = \phi(\{\gamma = (s,u,\xi) \in \mathcal{I} : (u,\xi) \in (O_s - O_n) \cup (O_\epsilon)\}),$$

$$D^n = \phi(\{\gamma = (s,u,\xi) \in \mathcal{I} : (u,\xi) \in O - O_\epsilon\}), \text{ and}$$

$$D^o = \phi(\{\gamma = (s,u,\xi) \in \mathcal{I} : (u,\xi) \in O_n - O_s\}).$$

Note that this is simply one way of partitioning the dictionary, and we could have done it differently if we wished. The connection to evolutionary spectral densities is

important when we consider the relationship to speech, where we will show that these densities actually describe physically meaningful phenomena.

We see that the two sets D^s and D^n are in fact partitioned according to the probability densities by noting that these quantities differ from the evolutionary spectral densities, used in defining D^s and D^n only by a constant factor. Thus, we can apply Theorem 2.

Measurability of Generated Sets

The conditions of the Theorem 2 require that $\phi^{-1}(D^s)$ and $\phi^{-1}(D^n)$ along with $\phi^{-1}(D_{\alpha,i})$ be measurable. $\phi^{-1}(D^s)$ and $\phi^{-1}(D^n)$ are measurable sets because of the assumptions on $h_t(\xi)$. The following theorem guarantees that for the Gabor dictionary and L_2 based matching criterion, $\phi^{-1}(D_{\alpha,i})$ is measurable.

Theorem 4 $|< f_i, \phi_\gamma >|$ *is continuous and measurable on* $\mathcal{I} = \mathcal{R}^+ \times R^2$.

Proof: First, we show that $|< f, \phi_\gamma >|$ is continuous in $\gamma \in \mathcal{I}$. Each γ in $\mathcal{I}$ is of the form $\{s, u, \xi\}$ where $s \in R^+$ and $u, \xi \in R$. Consider a sequence $\{\gamma_n\} = \{s_n, u_n, \xi_n\}$ with $|s - s_n| < \delta_{s,n}$, $|u - u_n| < \delta_{u,n}$, and $|\xi - \xi_n| < \delta_{\xi,n}$ such that $\delta_{s,n}$, $\delta_{u,n}$, $\delta_{\xi,n} \to 0$ as $n \to \infty$. For each n,

$$< f, \phi_{\gamma_n} > = \int f g(\frac{t - u_n}{s_n}) e^{-i2\pi\xi_n t} dt \tag{3.15}$$

Since $f \in L_2$, we can multiply it by another function in L_2, and the resulting product will be in L_1. We will use this fact to construct a dominating function as follows. First, $|g(\frac{t-u_n}{s_n})e^{-i2\pi\xi_n t}|$ is $|g(\frac{t-u_n}{s_n})|$, which is bounded, i.e. $|g(\frac{t-u_n}{s_n})| < B_{g,n}$, where $B_{g,n}$ is a constant depending on n. Choose an N. Let $B_g = \sup_{n>N} B_{g,n}$ and $s' = \sup_{n>N} s_n$. Since $\{s_n, u_n, \xi_n\} \to \{s, u, \xi\}$, $\exists M > N$ such that for $n > M$ $|g(\frac{t-u_n}{s_n})| \leq |Z_g B_g g(\frac{t-u}{s'})|$, where Z_g is a normalization constant that forces $g(\frac{t-u}{s'})$ to equal 1 at $t = u$. Then $|f| Z_g B_g g(\frac{t-u}{s'})$ is an integrable function bounding the magnitude of the integrand in (3.15) for all $n > M$. This follows because f, and $Z_g B_g g(\frac{t-u}{s'})$ are both in L_2, which implies that their product is in L_1. This in

turn implies that the proposed dominating function is integrable. By the Dominated Convergence theorem,

$$lim_{n\to\infty} \int fg(\frac{t-u_n}{s_n})e^{-i2\pi\xi_n t}dt \;=\; \int fg(\frac{t-u}{s})e^{-i2\pi\xi t}dt$$

Therefore $< f, \phi_{\gamma_n} >$ is continuous in γ. Since $| < f, \phi_{\gamma_n} > |$ is a continuous function of $< f, \phi_{\gamma_n} >$, $| < f, \phi_{\gamma_n} > |$ is continuous in γ. Hence the function is also measurable. $\diamond$.

With this result, we conclude that the universal properties developed in the previous chapter are applicable here as well. In this chapter, we have made a connection between the Probabilistic Pursuit and any application domain that may involve time-frequency analysis. First, the time-frequency consequences of pursuit analysis on deterministic signals was given. Then we generalized by considering the input to be a non-stationary process. It is this generalization that allows us in the next chapter to develop a specific time-frequency application based on the theory of speech production.

Chapter 4

Analysis of Speech

We desire that the theoretical results we have obtained fit naturally into an experimental framework, and furthermore, that the corresponding results give us more insight into our theoretical claims. In this chapter, we explain how speech is a natural choice for our experimental domain. Briefly summarizing the main points of the theoretical model, its knowledge base is a set of probability distributions $\mathcal{P}$ which we call the super dictionary. Each member of $\mathcal{P}$ is considered to be derived from a non-stationary stochastic process whose time frequency characteristics are different. The observation is a realization of one of these processes plus perhaps some noise, and by performing a time-frequency analysis, the Probabilistic Pursuit, of this observation, the goal is to efficiently identify the element in $\mathcal{P}$ which explains the observation best. In fact the observation need not be a realization of one of the processes described in $\mathcal{P}$, and instead could be arbitrary. The notion of finding the best element to explain the observation still remains valid. It was hypothesized that we could use waiting times and the distance measure to optimize over the elements. In this chapter, the construction of $\mathcal{P}$ is justified by considering the physical speech process for a number of Vowel-Consonant-Vowel (VCV) sequences. Then, given utterances of speech which are instances of the VCV sequences, we try to find the best matching element in our super dictionary, which must be constructed, by combining our waiting time and parameter space distance data.

4.1 Time-Frequency Information in Speech

In the preceding chapters we developed a signal decomposition which requires a signal model in the form of a time-frequency energy distribution. It should be reiterated that $\mathcal{P}$ should have elements that are significantly different for the method to be effective. This means that the portions of the search space that one model gives high probability to must not overlap completely with the same such portions defined by another model. Here we describe in some detail, the nature of speech as it relates to the construction of our signal models, the elements of $\mathcal{P}$. In doing so we hope to provide justification for their use as well. We begin by noting that the physical speech process is the result of airflow through the vocal cavities, which are most often modeled as tubes with time varying cross sectional areas. The resonances of the vocal tract, and their variation in time, describe the time frequency energy distribution in speech.

4.1.1 Modes of Vocal Tract Models

Uniform Tube

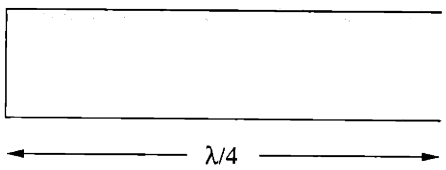

Figure 4-1: Uniform tube: closed at one end, open at the other.

Length $= l$, Cross-sectional area $=$ constant, closed at one end and open at the other.

$$F_n = \frac{c}{4l}(2n-1), \quad n = 1, 2, 3 \ldots,$$

where $c = 35,400 \ \frac{\text{cm}}{\text{s}}$. The approximate length from the glottis to the lips is $l = 17.7$ cm. This gives

$$F_n = 500(2n-1).$$

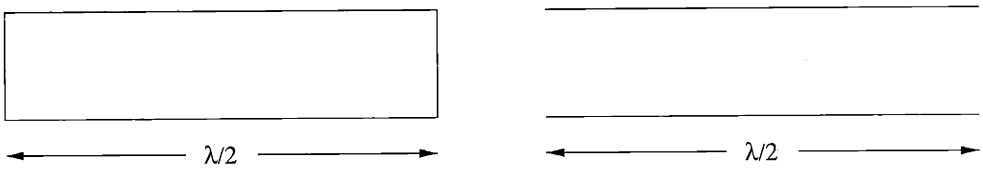

Figure 4-2: Uniform tube: closed or open at both ends.

If the tube is open at both ends,

$$F_n = \frac{c}{2l}(n), \quad n = 0, 1, 2 \ldots,$$

or

$$F_n = 1000(n), \quad n = 0, 1, 2 \ldots.$$

Helmholtz resonator

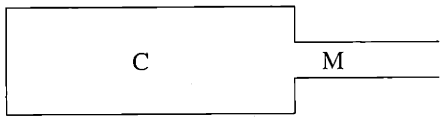

Figure 4-3: Helmholtz resonator

The natural frequency is

$$f_1 = \frac{1}{2\pi\sqrt{MC}},$$

where M is the acoustic mass and C the acoustic compliance of the resonator.

These can be considered as building blocks from which more complex vocal tract shapes are constructed by interconnection. However, in that case we must deal with the effects of coupling the tubes together which are often seen as slight modifications to the uncoupled resonances. To form a more complete model of the vocal tract, effects of excitation sources must be added.

4.2 The VCV Environment, where C is a Stop Consonant

This environment [28] is a sequence of articulatory movements, depicted in Figure 4-4, that take the vocal tract from a vowel like open position through a closure to a fully closed position, from which a release is initiated to bring the vocal tract back to the vowel configuration. This is manifest in our coupled tube model of the vocal tract as changes in lengths and cross sectional areas of the various sections and results in the movement of the natural frequencies of the entire model.

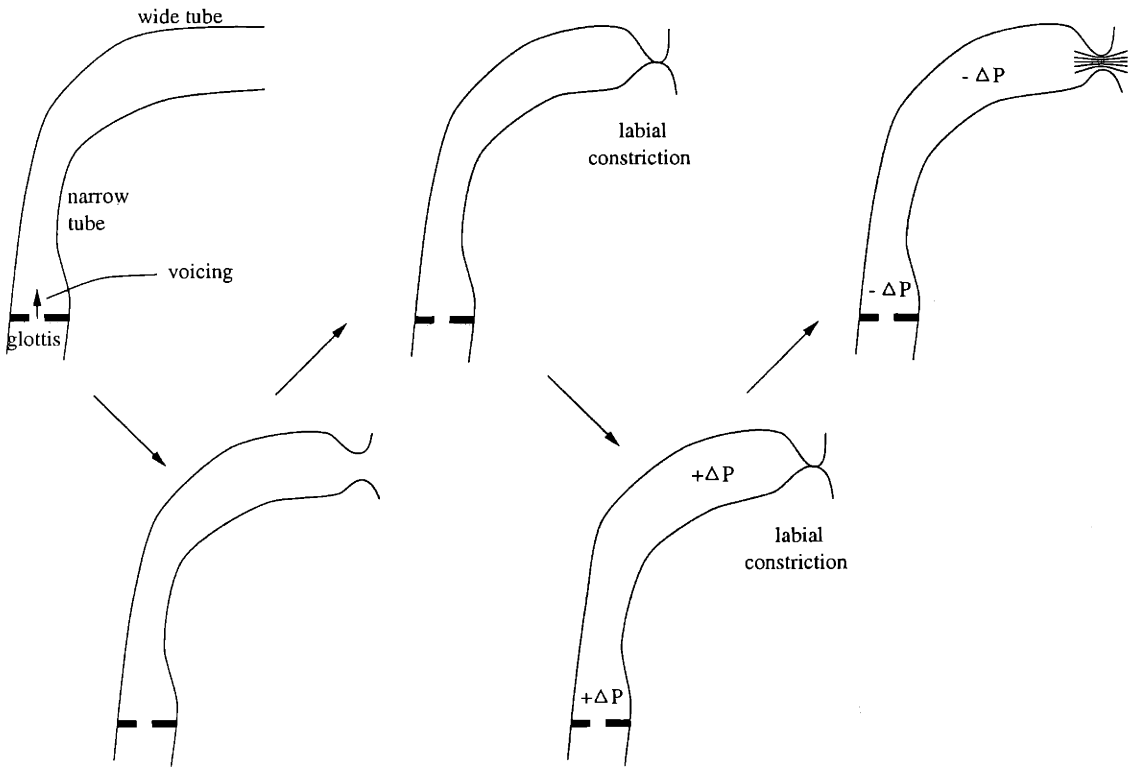

Figure 4-4: Sequence of Vocal Tract configurations for Vowel-Labial Stop Consonant-Vowel

During this sequence of articulatory movements, there is a corresponding sequence of excitation sources which appear in varying degrees and contribute to displaying or hiding these natural frequencies in the output. We can observe the movement of the natural frequencies to the extent that the sources excite them and in accordance with our ability to resolve fast spectral changes.

In order for a particular source to be active, certain conditions must be met by the vocal tract configuration. Often times, most of these sources do not exhibit an on/off behavior but rather they are active to various degrees. Voicing, turbulence noise, and transients are the major categories. The voicing and turbulence categories are further subdivided depending on differences in the vocal tract configurations.

Resonances of the vocal tract are best observed in the vowel environment where there is little mobility of the natural frequencies and where the excitation is modal voicing. This type of source, if modification by the radiation is included, has a $\frac{1}{f}$ behavior in magnitude, so we use preemphasis to correct the spectrum. The source is situated at the glottis and thus excites all of the natural frequencies of the vocal tract. For a voiced consonant, this excitation is strong for most of the utterance. In the case of a voiceless consonant however, as the vocal tract moves into and out of the closed position, this mode of excitation is weak.

On the other hand, at the release of the stop, the turbulence excitation is strong. This type of excitation is formed when there is a constriction in the vocal tract through which air is forced at a high speed. The constriction separates the vocal tract into back and front cavities. In the case of a glottal constriction, the back cavity is called the sub-glottal cavity. The net effect of this separation is an introduction of zeros into the vocal tract transfer function near poles of the back cavity transfer function. Thus at the output, we see mainly the front cavity resonances. But when the noise source is at the glottis, a case called aspiration, the front cavity resonances are those of the vocal tract proper, and the zeros that are introduced correspond to the poles of the sub-glottal cavity, which also appear.

Under the conditions that the supra-glottal constriction has both a small length and a rapid rate of increase in its cross sectional area, there is an initial transient in volume velocity in the first millisecond following the release. This source may contribute to the low frequency spectrum of the radiated sound.

4.2.1 Some General Comments

During periodic voicing, the fundamental frequency F_0 is determined by the vocal folds. For males, $F_0 = 125$ Hz whereas for females, $F_0 = 230$ Hz, approximately, on the average. There are changes spectrum amplitude which are roughly correlated with the movements of the formants, especially $F1$. Also, it is important to keep in mind that the lower formant frequencies usually increase faster than the higher ones at the release of a consonant.

4.2.2 VCV:Low Back Vowel - Labial Stop Consonant - Low Back Vowel

The sequence of vocal tract states as well as the approximate formant positions for a typical VCV utterance with a low back vowel and a labial stop consonant are given in Fig. 4-5. As can be seen, the vocal tract state for the vowel is modeled by a narrow tube near the glottis connected to a wider tube ending at the lips. This configuration is that of two resonators which are coupled together. The lowest natural frequency of the configuration, in this case $F1$, will be associated with the longer of the two sections. For the low back vowels, this is mostly the back tube, or cavity. As such, the state of the glottis will affect the bandwidth of $F1$. Typical values of $F1$ are in the range of 750-800 Hz.

When the tongue is in the back position, the value of $F2$ is close to the value of $F1$, and $F3$ and $F4$ are also close together. This can be seen in the figure, in the sections where voicing is active.

The transition to a labial stop consonant is achieved by closing the lips, and is indicated by the transition to the middle configuration in Fig. 4-5.

In the closure interval, $F1 \approx 200$ Hz. From the rate of change of the constriction at the lips (≈ 100 cm^2/s for the labial), it is estimated that most of the $F1$ movement, as well as that of the higher formants, is completed in about 10 ms. Here we have assumed that $F1$ is a front cavity resonance, and consequently is affected a great deal by the lip movement. On the other hand, this implies that $F2$ is a back cavity

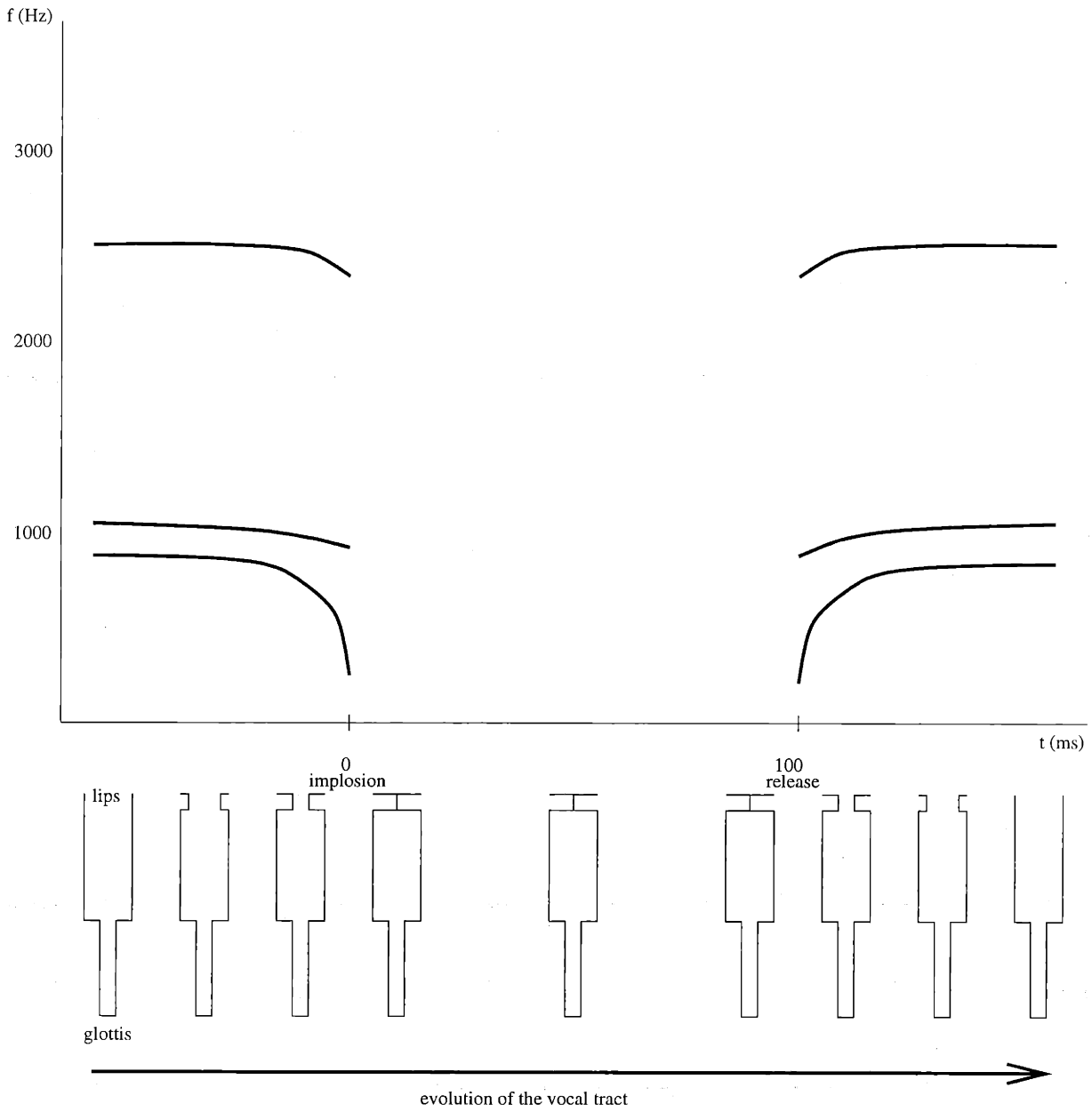

Figure 4-5: Vocal Tract Deformations in a VCV where V is a low back vowel and C is /p/.

resonance which is relatively unaffected by the constriction at the lips. A typical range of $F2$ would be 1100-1200 Hz. $F3$ is then a front cavity resonance, and moves as well, though less than $F1$. It may go from 2500Hz down to about 2200 Hz.

4.2.3 VCV:High Front Vowel - Labial Stop Consonant - High Front Vowel

Fig. 4-6 shows the vocal tract configuration sequence and the associated formant values. In this case $F1$ is a Helmholtz resonance, and the frequency in the closure should be $\approx$ 200 Hz. $F2$ is a front cavity resonance, (due to narrow lip opening) so it is affected by the lip movement. It may start at 2100 Hz, and transition to about 1400 Hz. Again much of the transition occurs within are approximately 10 ms, for the 100 cm^2/s rate. $F3$ goes from approximately 2250 Hz to about 2000 Hz. The first Back cavity resonance is about 2000 Hz. It starts as $F2$ and as the lips close, it becomes $F3$.

The two cases presented describe the articulatory effects during the rendering of labial stop consonants. Also, the figures serve to show the formant positions for the low back "aa" and high front "iy" vowels.

Effect of Place of Articulation

In addition to the labial, the data set used in the experiments contains alveolar and velar stop consonants. One essential difference lies in the rate that the articulators move, but another is the direction of movement and the starting points of the F2 and F3 transitions. The tongue moves slower for the alveolars than for the labials, and slower still for the velar. Consequently, the formants move more slowly in these cases.

4.3 Transition to Models

In this section we argue that in the context of a Probabilistic Pursuit analysis based on the Gabor dictionary, signal models for speech utterance observations can be

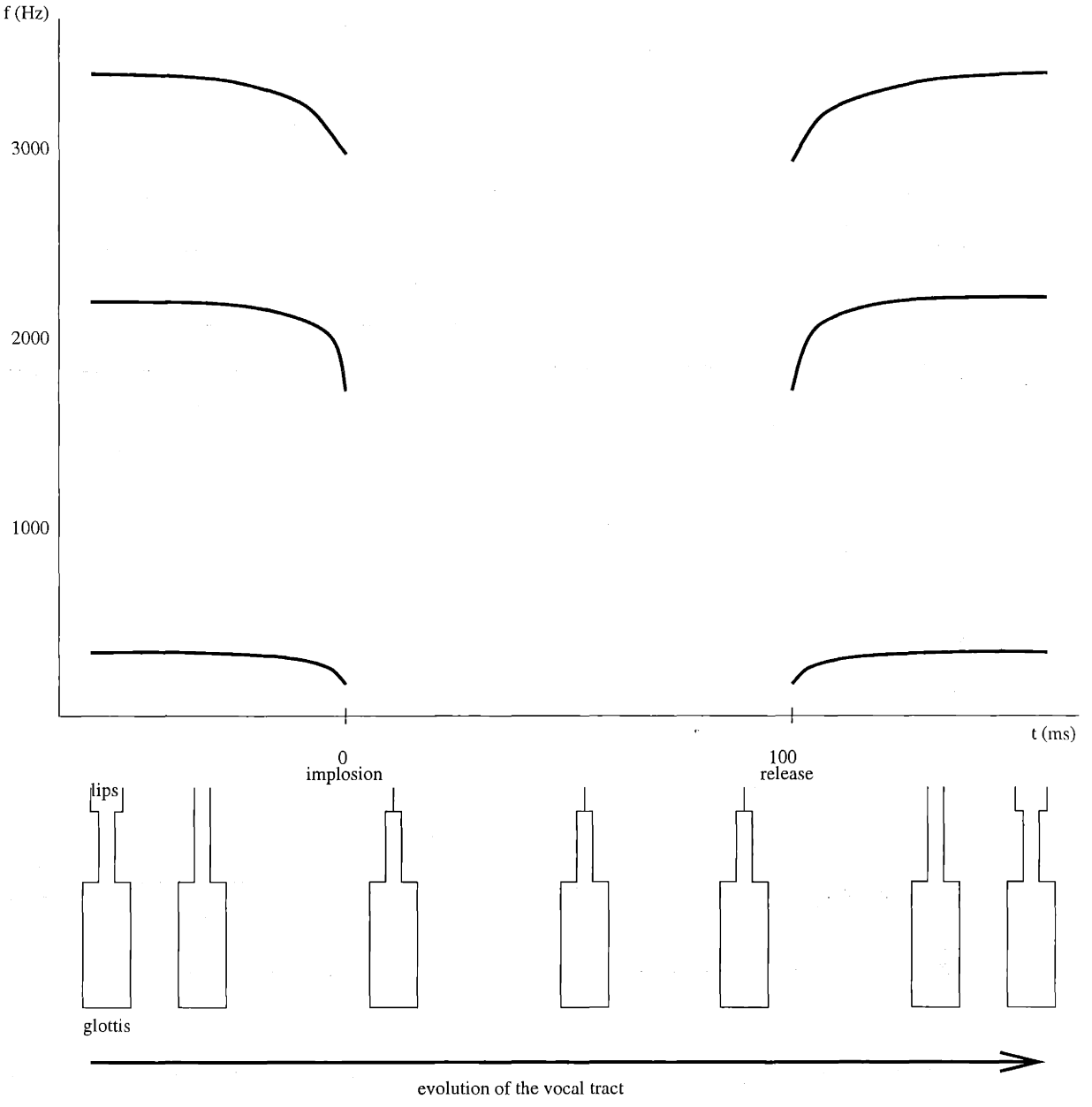

Figure 4-6: Vocal Tract Deformations in a VCV where V is a high front vowel and C is /p/.

derived from the paths of the formant frequencies as they evolve over time. As discussed in Section 3.2.3, when this type of analysis is performed on a realization of a semi-stationary oscillatory process $O(t)$ with evolutionary spectral density $h_t^{\mathcal{F}}(\xi)$, the criterion for selecting dictionary elements looks for those with large values of the correlation with the observation, whose expected value is (see equation 3.11)

$$E[\ |< O(t), \phi_\gamma(t) >|^2\ ] \ = \ \int_{-\infty}^{+\infty} |G(w)|^2 h_u^{\mathcal{F}}(w+\xi)dw \ + \ \mathbf{O}(\frac{s^2}{\pi W_O}).$$

Thus, the dictionary element selection procedure in effect searches for peaks in the smoothed evolutionary spectral density. Practically, we work with realizations and so let us consider the inner product of a dictionary element and the observation $o(t)$,

$$< o(t), \phi_\gamma(t) > = \ \frac{1}{\sqrt{s}} \int_{-\infty}^{+\infty} o(t) g(\frac{t-u}{s}) e^{-i\xi t} dt,$$

which is actually an estimate of the smoothed $h_t^{\mathcal{F}}(\xi)$ in the neighborhood of the time and frequency parameters of $\phi_\gamma(t)$ (see equation 3.11). The time and frequency parameters that result in large correlations correspond to the locations of the resonances, or formants, in the signal.

The idea is that if speech observations are modeled as realizations of some semi-stationary oscillatory process, the peaks of the evolutionary spectral density of that process can be estimated from the formant paths, which contain most of the energy in the signal. Thus the locations of the formants give us a good idea of which Gabor dictionary elements, in terms of their time and frequency parameters, will match the observed signal.

4.3.1 Prior Signal Model Construction

The models themselves are probability distributions on the Gabor dictionary. Here we specify how we construct such distributions based on knowledge of where the formant paths should be. Consider again the universe of signals $\{u_i\}$ and the corresponding super-dictionary $\mathcal{P} \ = \ \{\mathbf{P}_i\}$. Now each u_i is a speech utterance for which we know

the formant trajectories in the time-frequency plane. More precisely, we have for each u_i, a two dimensional function $\hat{h}_t^{u_i}(\xi) : R^2 \mapsto R$, whose peaks occur at t, ξ corresponding to the predicted formants paths. As the notation suggests, $\hat{h}_t^{u_i}(\xi)$ could be viewed as an estimate of the evolutionary spectral density of a process, a realization of which is u_i. Normalize $\hat{h}_t^{u_i}(\xi)$ so that $p_i(t,\xi) = Z_i \hat{h}_t^{u_i}(\xi)$, is such that

$$\int \int p_i(t,\xi) dt d\xi = 1,$$

Z_i being a constant. Then we follow the procedure in the previous chapter to form $\mathbf{P}_i$.

In the cases that we have described, the movements of the resonances are the result of a physical process that is not discontinuous. This process can produce rapid movements in the formants, but they are continuous. We claim that this continuity justifies our modeling of these utterances as semi-stationary oscillatory processes, which as we have noted before have the property that their energy distribution over frequency is changing slowly over time. The more rapidly the formants move, the less this is true however. But, the point is that the locations of the formants, wherever they may be, give a good indication as to which dictionary elements will match the observation and therefore should serve in creating a good signal model.

Chapter 5

Classification Scheme and A Preliminary Experiment

In this chapter, the application independent classification methodology is instantiated in the form of a speech classification experiment using the results on time-frequency analysis. First, we discuss the main ingredient of the analysis, which is $\mathcal{P}$. It is derived from the predicted formant paths of the set of utterances $\{u_i\}$ that constitute the data set. Then we describe in greater detail the specific waiting time and distance measures that we will use in making the decision, a process most easily viewed as a race in terms of the set of discrimination measures which we define in this chapter. Then we present the results, where our goal is to gain an understanding of the significance of the decision parameters.

5.1 Experiment

The purpose of the experiment is twofold. First, from the results, we want to interpret the significance of each of the components of the discriminant function, to be described. Then, we will provide a proof of concept of the classification paradigm. The set of data we use is based on these considerations and consists of utterances of voiced stop consonants in the context of a vowel.

5.1.1 Construction of $\mathcal{P}$

Though the theoretical development in the previous chapter gives us an idea of what should occur, we must ultimately look to the actual data to determine appropriate rates of movement for the formants and the models based on them. Denote the set of signals that we may observe by $\{u_j\}$. For each of these, we construct a model of the time-frequency behavior, based on our knowledge of speech and an analysis of the actual waveform. We then compare the models produced by our estimates with the formant tracks provided by the Entropics LPC tracker. By this we mean that the LPC tracks are used only to verify that the models are reasonable. This is because we want the models to be rough in order to account for the fact that in practice we will not have perfectly accurate models.

LPC Analysis

Let $s(n)$ be the discrete time speech signal. The assumption made is that the value of the speech signal at time n can be predicted as a linear combination of the past N speech samples,

$$s(n) \;=\; a_1 s(n-1) \;+\; a_2 s(n-2) \;+\; \cdots \;+\; a_N s(n-N).$$

In the Z-transform domain, this homogeneous difference equation gives the characteristic equation

$$C(z) \;=\; 1 - \sum_{j=1}^{N} a_j z^j,$$

from whose roots we can estimate the formant frequencies. The analysis consists of estimating the coefficients a_j based on differences between the real and predicted signals. The Entropics LPC tracker uses these roots in determining the formant paths through a signal. The problems with this method of obtaining formant estimates are well known and are a result of the fact that an all pole transfer function as above is not always a good model for the vocal tract. We use the tracks only to test the validity of the predicted models.

5.1.2 Hypothesized Models

We assume that for each utterance, we know the formant frequencies for the vowel environment, the start and finish locations for the stop transitions, and parameters, described below, for the formant paths during the stop transitions. Given this information, we model the formant paths as piecewise linear [29]. Thus the additional parameters that we need are the slopes and ranges of the line segments used. For the transition into the closure, the starting point is given by knowledge of the formant frequency and start location. Then, in succession, we linearly move this frequency through the given range using the given slope. The ending point for this segment is used as the starting point for the next. A similar technique is used for the transition from the closure except that now our starting point is the known ending formant frequency and finish time. In this case we build the model using the same procedure as for the transition into the closure, but going backwards in time. We note that the lengths of the transitions for different elements of $\{u_j\}$ are different. Thus, after the models are constructed with the above procedure, they are gated to fit within the transition. Recall that we assume the location and length of the transitions are known. For the experiment, each u_i is a speech waveform, and the known quantities were determined by looking at this waveform. See Appendix A for the parameter values for the speakers that we consider: cb, a female and ks, a male. The lengths of the models actually used were different from the ones derived from the parameters because they were gated. We present in figure 5-1 of one of the models, in the low back vowel case. The sampling rate for the signal is 16129 Hz, and the frequency axis is in Hz. In the figure, the thin lines are the LPC tracks and the thick superimposed lines constitute the model. In the transition, the model is at 0 Hz. However, the middle section of the utterance is not considered in the experiment.

As can be seen in the figure, we do not have perfect matches. One reason for this is that we force all formants to start and end at the same times, whereas the LPC tracks do show some variation. But, it seems more reasonable to assume a starting and ending time for the closure as a whole. These values are obtained by looking at

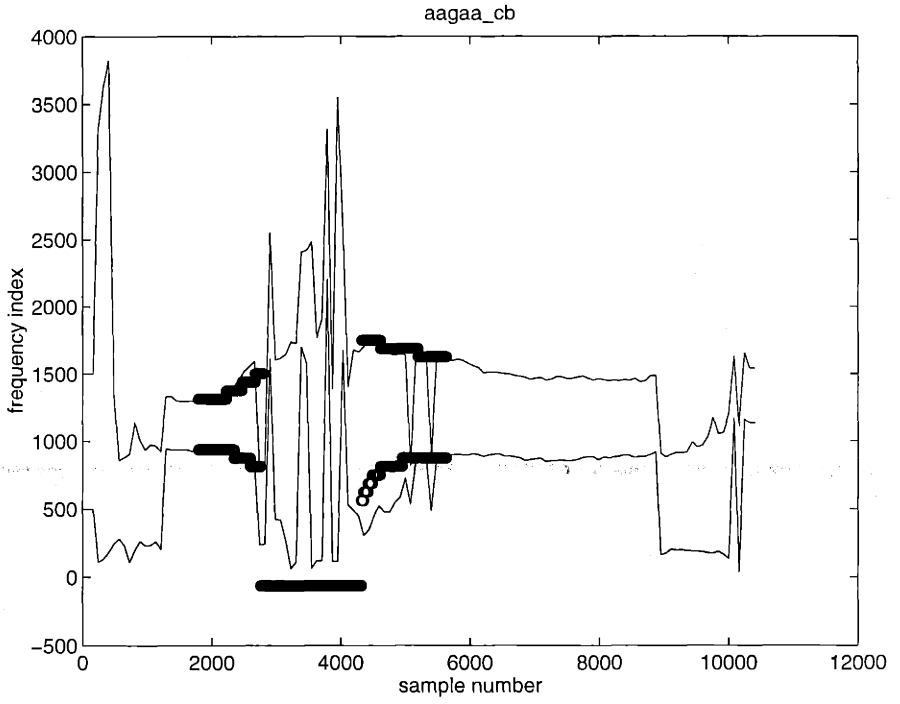

Figure 5-1: The prior signal model for the utterance aagaa overlapping the LPC derived tracks.

the waveform. Another reason is that the LPC tracks are sometimes erroneous. The models are thus a combination of what we expect to see, in light of the discussion in the previous chapter, and what we do see, from the tracks.

Probability Distribution

We assume that for each time and each formant, its likely location is distributed as a Gaussian in frequency with mean equal to our model value, and a small variance to account for errors. Then, we normalize the entire set of values to get a probability distribution which is one of the models in $\mathcal{P}$.

5.2 Frame Based Analysis

For reasons of efficiency, we choose to perform a frame based analysis, though the theory does not require it. Let us briefly explore the relationship between a windowed

Probabilistic Pursuit analysis and the Short Time Fourier Transform (STFT). In the case of the STFT, we can view the window position parameter as specifying a frame of data because the time frequency plane is sampled at uniform intervals in time with the same window size. But when we adaptively select the window size and position, as in the Probabilistic Pursuit, a frame based analysis must be defined more carefully. To do this, we note the following duality property: While in the time domain, the Probabilistic Pursuit can be viewed as constructing a function of time, in the frequency domain, this procedure builds the Fourier transform of this function.

$$o(t) = \sum_{n=0}^{m-1} a_n \phi_{\gamma_n}(t) + r^m o(t). \tag{5.1}$$

As a consequence of the linearity of the Fourier transform, we have

$$O(\xi) = \sum_{n=0}^{m-1} a_n \Phi_{\gamma_n}(\xi) + R^m o(\xi), \tag{5.2}$$

where $\Phi_{\gamma_n}(\xi)$ is the Fourier transform of $\phi_{\gamma_n}(t)$ and $R^m o(\xi)$ is the Fourier transform of $r^m o(t)$.

Both sums above converge to their respective left hand sides. We know that

$$\| r^m o_t \| \to 0.$$

But

$$\| R^m o(\xi) \| = K \| r^m o(t) \|, \ \forall m.$$

Therefore,

$$\| R^m o(\xi) \| \to 0.$$

So now rather than letting the window length determine the size of the frame, we choose this beforehand and treat each frame as a finite time signal to be analyzed by the Probabilistic Pursuit. Then as the decomposition proceeds in each frame the

STFT is being built. In discrete time,

$$s(m,n) = \sum_l a_{m,l}\phi_{m,l}(n) \; + \; r_l(n), \tag{5.3}$$

where the right hand side is an MPD. Transforming both sides, we obtain

$$S(m,k) = \sum_l a_{m,l}\Phi_{m,l}(k) \; + \; R_l(k). \tag{5.4}$$

So, based on the above duality, in the limit as l increases we have not lost anything, as compared to the STFT, by doing this. One might argue that when l ranges over a finite set, the components that are left out are in fact noise. We can obtain temporal information: derivatives of spectral components, cepstra, etc., as before, but in addition, we have gained information about the time-localization of the various frequency components in an analysis frame. This information is potentially useful in front-end processing.

5.2.1 Windowing

The choosing of frames for the Probabilistic Pursuit is a windowing operation in itself. The construction of these frames is not necessary to the signal analysis, but does increase the efficiency of the procedure.

Let us assume that we have an utterance of length M samples. One iteration of the MPD takes on the order of $M \log M$ operations. For the sake of argument let us arbitrarily fix the number of atoms we wish to compute, N. Computation of the total representation will require on the order of $NM \log M$ operations. On the other hand, say we wish to window our signal so that it has length $W \; << \; M$. Then, the procedure will take on the order of $NW \log W$ operations. This is a considerable saving when $W \; << \; M$. Further, as M increases, so does $NM \log M$, which is not the case for $NW \log W$.

Another reason for windowing is that it gives us the ability to use a different dictionary at each frame, allowing us to provide information on the location of significant

components in the signal. The representation can thus be adaptively controlled.

However, there may be adverse effects as well. Assume we have a signal $o(t)$, and that we split it into beginning, $b(t)$, and ending, $e(t)$, parts. We can compute decompositions for $b(t)$ and $e(t)$ independently, which when combined appropriately will converge to $o(t)$. Separately, we could compute a decomposition for $o(t)$, which also converges. The two separate procedures would most likely yield non-identical sets of atoms. Most might be the same, but, for example, those atoms in the last decomposition with scales greater than the length of $b(t)$ and $e(t)$ would disappear and appear as combinations of smaller atoms in the first two decompositions, as would those on the border of $b(t)$ and $e(t)$.

5.2.2 Window Size

The window size chosen was 256 points, or approximately 16 ms. The reason for this was to obtain a reasonable amount of frequency resolution. This does not have a serious consequence for our analysis since we are optimizing over window length. However, the scale was limited to be at most 128, which is approximately the duration of the periodic source waveform during voicing for males. The rationale is that we are interested only in components that are within a pitch period.

Having made the connection to frame based analysis, we mention briefly that the output of the Probabilistic Pursuit can be viewed in such a way that it gives complementary information. Namely, we could continue with the frame based approach and calculate a (variable length) vector of features for each frame. That is, we augment the cepstra and their derivatives with time localization information and bandwidth information now available. One could argue that the added information could improve performance.

5.2.3 Implementing the Probabilistic Pursuit

In the experiments we deal with discrete time signals of a certain length N, and consequently we use a sampled and periodized version of the Gabor dictionary. Recall

from before that the continuous time dictionary is defined as

$$D = \{\phi_\gamma = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) e^{i\xi t}\}$$

each element of which is completely specified by an index set $\gamma = \{s, u, \xi\} = \{\text{scale, translation, frequency}\} \in R^+ \times R^2 = \mathcal{I}$ where the window function $g(t)$ is a Gaussian,

$$g(t) = 2^{1/4} e^{-\pi t^2}$$

and the constant is chosen so that the L_2 norm of $g(t)$ equals 1. A sampled and periodized version of the window function is given as

$$g_{sp}(n) = \frac{K_{sp}}{\sqrt{s}} \sum_{j=-\infty}^{+\infty} g(\frac{n - jN}{s}),$$

K_{sp} being a normalization constant. The dictionary element is then defined as

$$\phi_\gamma(n) = g_{sp}(n-p) e^{i(2\pi k n)/N}.$$

The new index set is $\gamma = \{s, p, k\}$, where $s \in (1, N)$ and $p, k \in [0, 1, 2, ..., N-1]$. This is the dictionary we use. However, the signal we decompose is real, and thus in our selection process, we maximize the norm of the inner product of the signal, f_i, with the real atoms

$$\phi_{\gamma,\varphi}(n) = K_{\gamma,\varphi} g_{sp}(n-p) \cos(\frac{2\pi k}{N} n + \varphi),$$

where φ is the phase of $< \phi_\gamma(n), f_i >$ and $K_{\gamma,\varphi}$ is a normalization constant.

Consider a subset of the index set given by $(2^j, p2^{j-1}, k\pi 2^{-j})$, where $0 < j < \log_2 N$, $0 \leq p < N2^{-j+1}$, and $0 \leq k < 2^{j+1}$. We first maximize over this set and then find a local maximum in the complete dictionary in a neighborhood of the resulting element.

This defines the best matching element with respect to which the Probabilistic

Pursuit must operate. The prior signal models are thus distributions on this discretized and periodized Gabor dictionary. In our experiments we chose to optimize over the scale parameter, and thus the distribution is on the time-frequency plane. We simply choose the best scale, from those possible, for each time frequency point that occurs in the dictionary process. Then in implementing the Probabilistic Pursuit, the two step selection procedure is carried out after having chosen a value β, $0 < \beta \leq 1$, used in the criterion, i.e. the candidate is chosen if its *match value* is greater than $\beta \times$ *best match value.* Using the same utterance and signal model in figure 5-1, we show the effect of using a prior distribution. First in figure 5-2 the results of a decomposition when we do not use any prior information is given. Then in figure 5-3 we see the effect of using a prior signal model. In these figures, the pluses denote the locations of the selected dictionary elements.

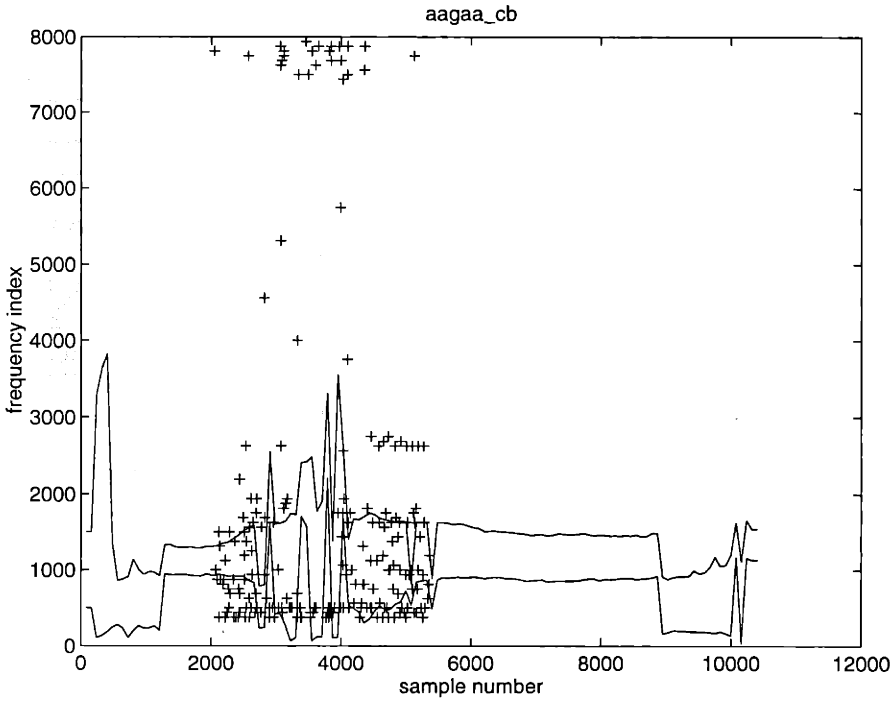

Figure 5-2: Frame based decomposition of the utterance aagaa where no prior knowledge is used.

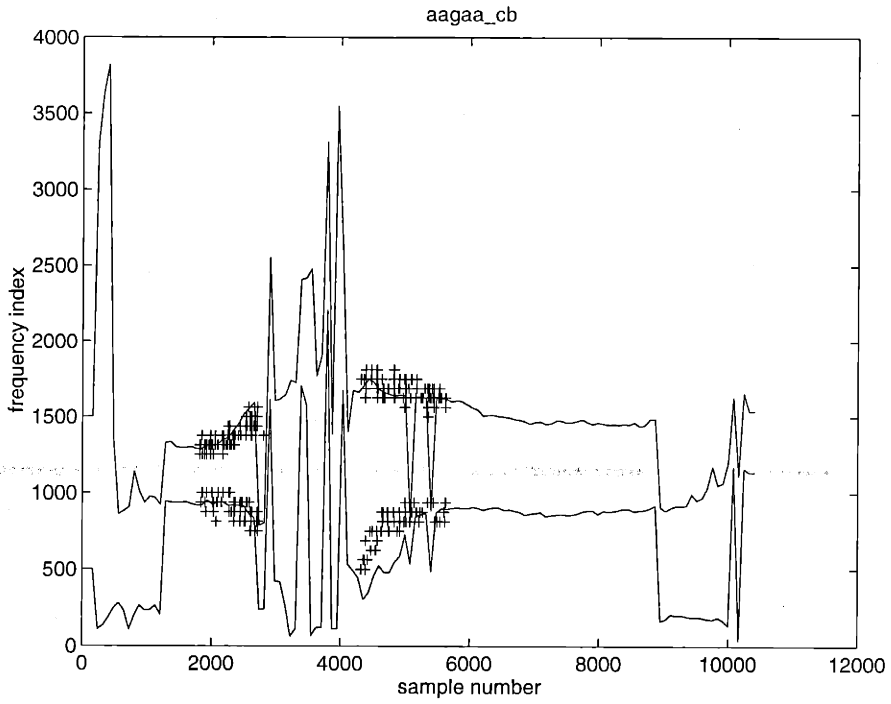

Figure 5-3: Frame based decomposition of the utterance aagaa where a prior signal model is used.

5.3 Frame based Discriminant Function

Consider that there is a set of utterances that we might observe which make up the set $\{u_j\}_{j=1,\ldots,M}$, and to each of them we have associated a prior signal model included in $\mathcal{P}$. The observation in discrete or continuous time, call it f now, is a realization of one of the utterances, say u_{j_0}. The signal restricted to the k^{th} frame is $f^k = f_0^k$, where f_i^k is the i^{th} residue.

Assume that for frame k, there are N_k atoms, a number which can be specified beforehand or determined over the course of the decomposition by stopping if the waiting time gets too big. The signal energy consumed by these atoms is

$$E_k = \sum_{i=1}^{N_k} | < f_i^k, \phi_{\gamma_i^k}(t) > |^2.$$

The total waiting time for the frame is taken to be the sum of the waiting times,

$$W_k = \sum_{i=1}^{N_k} \tau_i^k.$$

The total distance is

$$d_k = \sum_{i=1}^{N_k} d(\gamma_i^k),$$

and in the experiment we use the L_1 distance of the selected elements to the nearest model peak at the same time index. We define a frame discriminant function, $\varrho(\cdot)$, to be a function of all the

$$\overline{\rho}_k = \begin{bmatrix} \frac{W_k}{E_k} \\ \frac{W_k}{N_k} \\ \frac{d_k}{N_k} \\ N_k \end{bmatrix},$$

for each k. The method that is used in the later experiments stops the decomposition in a frame if the waiting time for that element is too large. As a result, the number of atoms in each frame is not known beforehand. For this reason, the first three components of ρ_k are given as per atom quantities. The smaller each component of ρ_k, the better the match for the first three components. But we use the last component, the number of atoms in the frame, to indicate the amount of the signal decomposed and thus a larger number here is better. For these reasons, we choose

$$\varrho = \frac{\sum_k w_1 \overline{\rho}_k(1) + w_2 \overline{\rho}_k(2) + w_3 \overline{\rho}_k(3)}{1 + \sum_k \overline{\rho}_k(4)}.$$

5.3.1 Minimum-Time Decomposition

The decision procedure based on the function ϱ defined above is closely related to the Minimum Description Length (MDL) principle developed by Rissanen [26]. The principle is based on joint consideration of model and data complexity. More specifically, given a parametric model class $\{M_j\}$ defined over a variable length set of parameters θ^j taken from $\{\theta_i,\ i = 1, 2, 3, ...\}$ and a sequence of observations $\{x_1, ..., x_n\} \equiv x_1^n$, let $\hat{\theta}^j$ be the Maximum Likelihood estimates of the parameters given the data x_1^n for

model M_j. Then, the principle would select a model from the class by minimizing the quantity

$$L(x_1^n|M_j) \;=\; \text{code-length}(x_1^n|M_j(\hat{\theta^j})) \;+\; \text{code-length}(\hat{\theta^j}).$$

Essentially,

$$L(x_1^n|M_j) \;=\; \text{complexity}(x_1^n|M_j(\hat{\theta^j})) \;+\; \text{complexity}(\hat{\theta^j}),$$

where complexity in this case is associated with the length of the codewords necessary to transmit the components. Our formulation however requires a somewhat different notion of model complexity. The models we use below, in $\mathcal{P}$, are still parametric. However each has the same finite parameter set, consisting of the values of the probability mass function on the time frequency plane. While the notion of code-length still has relevance in evaluating the data complexity, the first term on the right hand side above, the notion has little meaning for our model complexity, given that each is simply a finite, same-sized set of numbers. The notion of model complexity that we develop uses the time required to analyze our observation given a model. Thus our operating principle is *Minimum-Time Decomposition.* The numerator of ϱ is $w_1 \sum_k \overline{\rho}_k(1) + w_2 \sum_k \overline{\rho}_k(2) + w_3 \sum_k \overline{\rho}_k(3)$. Since $\sum_k \overline{\rho}_k(3)$ is the sum of distances $d(\gamma_i^k)$ between the prior model and the decomposition of the observation data, we can associate it, minus a normalization constant, with $-log(p(\text{observation}|\text{model}))$, and thus the code-length interpretation is valid. $\sum_k \overline{\rho}_k(1)$ and $\sum_k \overline{\rho}_k(2)$ are a sums of per/energy and per/atom waiting times, respectively, and these together constitute our model complexity. That is, a model is complex if using it requires long searches through the dictionary in order to find matching components.

5.3.2 The Parallel Nature of the Decision Procedure

Since each frame can be analyzed independently, and moreover, since no post-processing is necessary in the calculation of $\varrho(\{\rho_k\})$, a decision can be made at any time simply

by pooling the current results in each frame. In this case, the decision is based on a set of vectors,

$$\begin{bmatrix} \frac{W_1}{E_1} \\ \frac{W_1}{N_1} \\ \frac{d_1}{N_1} \\ N_1 \end{bmatrix}, \begin{bmatrix} \frac{W_2}{E_2} \\ \frac{W_2}{N_2} \\ \frac{d_2}{N_2} \\ N_2 \end{bmatrix}, \begin{bmatrix} \frac{W_3}{E_3} \\ \frac{W_3}{N_3} \\ \frac{d_3}{N_3} \\ N_3 \end{bmatrix}, \cdots, \begin{bmatrix} \frac{W_k}{E_k} \\ \frac{W_k}{N_k} \\ \frac{d_k}{N_k} \\ N_k \end{bmatrix}, \cdots,$$

which, in principle, are all simultaneously being constructed for each model. In fact, we need not use all of the frames in making a decision. That is, an incomplete decomposition of the observation will not necessarily hurt the classification. We might in fact choose to alternate the decomposition iterations with the decision making so as to be able stop when a clear choice becomes apparent.

5.4 Decomposition Results

The experiment is as follows. An utterance is seen and it is to be decided which of /b/, /d/, or /g/ occurred. The decisions will be made based on the frame based discriminant function described in the last section and so for each utterance frame and each signal model, a decomposition is computed. Speaker cb is a female.

For speaker cb, the following models were created:

- /b/, /d/, or /g/ in the context of /aa/
- /b/, /d/, or /g/ in the context of /ih/

For speaker cb, the following utterences are tested:

- /b/, /d/, or /g/ in the context of /aa/
- /b/, /d/, or /g/ in the context of /ih/

For comparison we also include some results for a male speaker, ks.

For speaker ks, the following models were created:

- /b/, /d/, or /g/ in the context of /aa/

For speaker ks, the following utterences are tested:

- /b/, /d/, or /g/ in the context of /aa/

In analyzing the results of the experiments, we want to find out the significance of the parameters used in the frame based discriminant function as well as verifying the usefulness of the paradigm. The experiment was performed so as to be able to analyze separately the transitions into and out of the closures for each formant.

5.5 Classification Results

The experiments we performed can be thought of as being 36 cases of a three way decision problem. For each frame of data that is analyzed, three basic values are computed:

$$E_k = \sum_{i=1}^{N_k} | < f_i^k, \phi_{\gamma_i^k}(t) > |^2,$$

$$W_k = \sum_{i=1}^{N_k} \tau_i^k, \text{ and}$$

$$d_k = \sum_{i=1}^{N_k} d(\gamma_i^k).$$

Recall from before that these are used in making up three of the four components of the vector $\overline{\rho}_k$, the fourth being the number of atoms in the frame N_k, and the actual decision is made based on

$$\varrho = \frac{\sum_k w_1 \overline{\rho}_k(1) + w_2 \overline{\rho}_k(2) + w_3 \overline{\rho}_k(3)}{1 + \sum_k \overline{\rho}_k(4)},$$

where $w_1 = 1000$, $w_2 = 100$, and $w_3 = 10$ were chosen so as to bring all of the values to the same order of magnitude. Let us write the above as

$$\varrho = w_1 \frac{\sum_k \overline{\rho}_k(1)}{1 + \sum_k \overline{\rho}_k(4)} + w_2 \frac{\sum_k \overline{\rho}_k(2)}{1 + \sum_k \overline{\rho}_k(4)} + w_3 \frac{\sum_k \overline{\rho}_k(3)}{1 + \sum_k \overline{\rho}_k(4)} \tag{5.5}$$

$$= w_1 \frac{\varrho_1}{1 + \varrho_N} + w_2 \frac{\varrho_2}{1 + \varrho_N} + w_3 \frac{\varrho_3}{1 + \varrho_N}. \tag{5.6}$$

It might be the case that we do not get any atoms for the frame, that is $N_k = 0$, and for this k, $\overline{\rho}_k(1)$, $\overline{\rho}_k(2)$, $\overline{\rho}_k(3)$, and $\overline{\rho}_k(4)$ are all zero. However, in this case we want to add a baseline value to the each of the three ϱ quantities in order to penalize the fact that no atoms were found in a reasonable time. We chose to set $\overline{\rho}_k(i) = \frac{10}{w_i}$ for $i = 1, 2, 3$ when $N_k = 0$. The results from the decomposition can be found in Appendix B. There, for speaker cb, the results for the /aa/ context are given, followed by the /ih/ context, after which come the /aa/ context results for speaker ks. The tables are organized with the various parameters described above across the top and the different models along the side. Thus, each table represents the results of a parallel decomposition using the three models /b/, /d/, and /g/. As we are interested in low waiting times and small distances, our decisions consist of choosing the model that gives the lowest values for the parameters. In the appendix, each utterance has two tables associated with it. All of the parameters in the headings are described above. The second table, i.e. the normalized parameters, are what we base our decisions on. The normalization accounts for the differences in the total number of atoms in each of the decompositions. The data is presented so that we can see which decision would be made if each parameter was used independently, as well the combined score ϱ. The tables in the appendix are further analyzed below.

Table 5.1: Speaker: cb, Classification results in the context of /aa/. A 1 in the matrix indicates correct classification, a 0, incorrect classification, and an i indicates an indeterminate case. In the model column, a 1 corresponds to the first formant in the transition to the closure, a 2 to the first formant in the transition from the closure, a 3 to the second formant in the transition to the closure, and a 4 to the second formant in the transition from the closure.

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b1	i	i	i	i
b2	1	1	1	1
b3	1	1	i	1
b4	1	1	0	1
d1	1	1	1	1
d2	1	1	1	1
d3	1	i	0	1
d4	1	1	i	1
g1	1	1	1	1
g2	1	1	0	1
g3	1	1	0	1
g4	0	0	0	0

Table 5.2: Speaker: cb, Classification results in the context of /ih/. A 1 in the matrix indicates correct classification, a 0, incorrect classification, and an i indicates an indeterminate case.

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b1	1	1	1	1
b2	0	0	1	0
b3	1	0	i	1
b4	0	0	0	0
d1	1	1	0	1
d2	1	1	0	1
d3	1	1	1	1
d4	1	1	i	1
g1	1	1	0	1
g2	0	0	0	0
g3	1	0	0	0
g4	1	1	0	1

Table 5.3: Speaker: ks, Classification results in the context of /aa/. A 1 in the matrix indicates correct classification, a 0, incorrect classification, and an i indicates an indeterminate case.

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b1	0	0	1	0
b2	0	0	1	0
b3	1	1	1	1
b4	1	1	0	1
d1	1	1	1	1
d2	1	1	1	1
d3	1	1	1	1
d4	1	1	0	1
g1	1	1	0	1
g2	1	0	0	1
g3	1	1	0	1
g4	1	0	0	1

Table 5.4: Speaker:cb, Correct classification percentages for the /aa/ environment.

	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
% correct	83	75	33	83

Table 5.5: Speaker:cb, Correct classification percentages for the /ih/ environment.

	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
% correct	75	58	25	67

5.5.1 Analysis

Let us consider each parameter making up ϱ independently, and then ϱ itself. The three tables 5.1, 5.2 and 5.3 give separately the classification results based on each parameter as well as those based on their linear combination. In each table, a 1 indicates correct classification, a 0 indicates incorrect classification, and an i indicates that the case was indeterminate. In the first column, the number after the stop indicates the particular transition and formant looked at in making the decision. A 1 corresponds to the first formant in the transition to the closure, a 2 to the first formant in the transition from the closure, a 3 to the second formant in the transition to the closure, and a 4 to the second formant in the transition from the closure. In tables 5.4, 5.5, the rounded percentages of correct classifications are given for the two contexts for speaker cb. And in table 5.6, percentages for the /aa/ context for speaker ks are given.

These results are intuitively satisfying and reinforce the observations we have made. Namely, the waiting time parameters give very good results whereas the distance parameter seems to have a lot of variance. Rather than training, we constructed

Table 5.6: Speaker:ks, Correct classification percentages for the /aa/ environment.

	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
% correct	83	67	50	83

models. But again we reiterate that the models were made purposefully rough for the sake of experiment. We made 36 classifications, taking into account the independence of the decisions for each transition and formant. Though the absolute percentages are not very high (They could have been improved by constructing more specific models.), here we directly see the advantage that our waiting time approach gives us. We know from the theory that spectral estimators have a large variance. To reduce the effect of this variance, we used parameter space distances. However, they are still based on the exact location of the spectral peaks encountered and thus still are affected by the variance. We point out that since the atoms are selected probabilistically, there is an additional variance in the distance. The consequences of these effects are evident in our results. On the other hand, the waiting time parameters, which are based on a *distribution* of where the peaks might occur, work very well.

Chapter 6

Discussion, Modifications, and Extensions

We have developed a Probabilistic version of a pursuit type algorithm with the goals of both efficient representation and efficient classification in mind. The basic notion of the pursuit, that the way to analyze a signal is to search for its most salient characteristics, is generalized to include a requirement that the matching in the search be done with respect to a probabilistic prior signal model. The results of the decomposition in such a scheme indicate whether or not the prior signal model is a good one. Namely, we developed the idea of a Minimum-Time Decomposition. Moreover, when the prior model is good, the representation favors signal over noise. When there are a number of models, we can select one as explaining our signal best if it results in the best match. In our case, this was indicated by small decomposition (waiting) times and parameter space distances for the selections. Also, we noted a potential gain in the information obtained when this decomposition is used in the front-end processing stage of an analysis.

A point of note is that our decomposition procedure needs only the distribution on the dictionary, yet in the thesis we considered partitions of the dictionary. These partitions were conceptual entities which we used to argue that correct signal models yield low waiting times. That is, a distribution is an assumption of what in the dictionary constitutes signal and noise. If our input has energy on the components

that we consider to be signal, the waiting times in the decomposition will be small.

There is a generalization here which takes us one level above that which is normally dealt with. Namely, the distributions that we call templates constitute in and of themselves a dictionary: The dictionary of distributions on the primary dictionary D, called $\mathcal{P} = \{\mathbf{P}_i\}$. The matching criterion is based on small waiting times and parameter space distances. Classification is choosing the distribution on the dictionary that explains the input best, i.e. a Pursuit with respect to $\mathcal{P}$.

The above are properties of the pursuit which are essentially independent of the application domain. Once we choose a dictionary though, we are explicitly stating the nature of the inputs that we will consider. In this thesis, we chose to deal with the Gabor dictionary and developed some results that show the consequences for time-frequency signal analysis. We then generalized to the case when the signals analyzed were realizations of semi-stationary oscillatory processes. Then we argued that speech is a good candidate for our application domain. Towards this end, we discussed the nature of speech formants and used them to derive probability distributions on the Gabor dictionary. This was based on the claim that the resonances contain most of the energy. Although, they do not capture all of the energy, since for example, there is noise. By noting that the formants move smoothly, we argued that it is reasonable to model speech utterances as semi-stationary processes. Thus, the distributions that we obtain from the formant paths are, to a constant factor, approximations to the evolutionary spectral density. Thus, based on our results, we concluded that they give a good indication of the dictionary elements that will match the signal. These were the distributions that made up $\mathcal{P}$.

We then experimented with classifying utterances based on such models and our probabilistic generalization of the structure of Pursuits. Our results, based on 36 classifications, indicate that the procedure is efficient and useful. In particular, we saw that based on a relatively small number of parameters, and using a technique, based on a Minimum-Time Decomposition principle, that required no post-processing, good classification results were possible. Further, the Minimum-Time Decomposition based classification was much better than that based on the parameter space distance,

suggesting that it is a worthwhile measure providing unique information.

6.1 Modifications

6.1.1 Bigger Frame Size

Clearly, we get better results if we use a bigger frame size. This is quite unlike the Short Time Fourier Transform where there are serious consequences such as poor time localization. In our case, we are optimizing over the window length. Because in our experimental implementation this optimization is time consuming, we chose a relatively small frame. An obvious improvement is to use a more efficient implementation and use bigger frames. Based on the results in [18], we can safely claim that such an implementation does indeed exist. Bigger frames would mean better frequency resolution and would allow for more accurate prior models.

6.1.2 Time Averaging

In chapter 3 we developed an analysis of the decomposition selection criterion in the context of Time-Frequency analysis. Not only did we consider deterministic inputs, but we extended the analysis to non-stationary oscillatory processes. The framework was based on the inner product and to restate a result from Priestley,

$$E[\ |< o(t), \phi_\gamma(t) >|^2\ ] \ = \ \int_{-\infty}^{+\infty} |G(w)|^2 h_u(w+\xi)dw \ + \ e(\frac{s^2}{\pi W_X}). \qquad (6.1)$$

It was noted that the quantity has a very large variance. It is possible to reduce the variance by time averaging. That is, given a point in the parameter space, γ, instead of looking at the quantity $|< o(t), \phi_\gamma(t) >|$, which we recall is a function of the set of parameters, and thus u, as explained in chapter 3, we look at

$$\int |< o(t), \phi_\gamma(t) >|^2 v(s-u)du.$$

Here $v(\cdot)$ is another (weighting) window function with the property that $\int_{-\infty}^{+\infty} v(u)du = 1$. This is useful when the prior signal models that we have are good estimates of the evolutionary spectral densities of the processes we are trying to identify. Recall that the selection procedure looked for elements in a set $D_{\alpha,i}$, but that the contents depended on $| < f_i, \phi_\gamma > |$. To use the time averaging, a new set must be defined.

$$D_{\alpha,i,tav} = \{ \phi_\gamma : \phi_\gamma \in D \text{ and } \int | < f_i, \phi_\gamma > |^2 v(s-u)du \geq \alpha \sup_{\gamma \in \mathcal{I}} \int | < f_i, \phi_\gamma > |^2 v(s-u)du \ \}. \quad (6.2)$$

Then, the selection procedure must wait for an element, or parameter set, that is in $D_{\alpha,i,tav}$. However, this does not necessarily guarantee that the construction of the signal converges. If on the other hand, the agent also checked that the parameter set was in $D_{\epsilon,i}$, for an appropriate small positive ϵ, convergence would be guaranteed.

6.2 Extensions

6.2.1 Dictionary Evolution

There are some natural extensions to this probabilistic paradigm. Since the analysis is frame based, and it is likely that the spectral components in adjacent frames are correlated, a scheme that adapts the distribution for the next frame given the elements chosen in the current frame is reasonable. In this case, templates and classification loose their significance. Instead we infer that our decomposition will be very fast and would adapt the distribution to the signal structure.

6.2.2 Dynamic Programming - Extracting Components

In this thesis we have discussed representation and classification. However, we can also talk about extracting components from the results of a Pursuit. In extracting components, the goal is to group together a subset of the atoms in the decomposition to form descriptors of the input signal. Furthermore, this grouping need not be static.

One could decompose the signal into N_1 atoms, perform the grouping, and then do it again with $N_2 > N_1$ atoms, etc. The information we gain from grouping at one iteration can be used in the next, an so on. This approach is outlined below.

We can restate the above in the framework of dynamic programming. We assume we have a signal to be analyzed and that a (Probabilistic) Pursuit for it has been computed. We decide a priori the number of descriptors we will look for. Let us consider the case of a one descriptor. After N iterations we get N elements of the dictionary, each element of which is described by a set of parameters. Consider the set of points in the parameter space which correspond to the selected atoms. The state space for the dynamic programming algorithm will be built upon this set. We take the product space of this with a set of states which describe a local measure of fit, which for example can be the derivatives of the parameters. At any given stage, the control space of the system consists of the possible assignments of an atom to the descriptor. This is a deterministic problem.

Assume we are in state $\overline{x}_k$, which is a vector containing the atom and local fit parameters. The result of control u_k will be an assignment of one atom to the descriptor. This procedure directly gives the new atom parameters $k+1$. The local fit measures for stage $k+1$ are determined as a function of the new atom parameters and the old parameters in $\overline{x}_k$. This will be denoted as $\overline{x}_{k+1} = f_k(\overline{x}_k, u_k)$.

The cost at stage k, $g(\overline{x}_k, u_k)$, is determined as a function of the differences between the parameters in stage $k+1$ and stage k. The cost can, in general, be more complicated. For example, the statistical correlations we may discover could be employed.

Then the Dynamic Programming equations are:

$$J_N(\overline{x}_N) = 0 \quad \forall\, \overline{x}_N$$

for any k,

$$J_k(\overline{x}_k) = \min_{u_k}(g(\overline{x}_k, u_k) + J_{k+1}(f_k(\overline{x}_k, u_k)))$$

In terms of the hierarchical description of classification, the structure of these

components could be viewed as extra information allowing a refinement of the classification decision.

6.2.3 Statistically Derived Prior Models

In our experiments, we constructed prior signal models based on predicted formant paths and the idea was that we were trying to model the densities of non-stationary processes. In future work, we could however take a different point of view. Given a database that contains many repetitions of the signals to be distinguished, we could perform a deterministic decomposition of all of its elements. Then by combining all of the elements that are instances of a particular signal, we could derive the induced distribution, $\mathbf{P}_{induced,i}$, on the dictionary elements, and use it as a prior signal model.

Appendix A

Tables of Parameter Values

Table A.1: Speaker cb: parameter values for the voiced stop consonants in the context of /aa/.

aabaa_cb F1				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	10	-1	10	-0.5
2	70	-4	200	-10
3	50	-4	100	-15
4	100	-4	150	-15

aabaa_cb F2				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	60	6	50	2
2	60	6	50	1.75
3	100	7	50	1.5
4	50	3	50	1.5

Table A.2: Speaker cb: parameter values for the voiced stop consonants in the context of /aa/.

aadaa_cb F1				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	10	-0.5	10	-0.4
2	70	-3.5	100	-3.5
3	50	-4	100	-15
4	50	-4	200	-15

aadaa_cb F2				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	70	4.5	50	2
2	70	4.5	50	2
3	70	4.5	50	2
4	50	4.5	100	2

Table A.3: Speaker cb: parameter values for the voiced stop consonants in the context of /aa/.

aagaa_cb F1				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	10	-1	10	-0.4
2	100	-4	100	-3
3	10	-5	100	-15
4	20	-5	200	-15

aagaa_cb F2				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	40	4	50	1.75
2	40	4	50	1.75
3	100	5	50	1.75
4	50	5	50	1.75

Table A.4: Speaker cb: parameter values for the voiced stop consonants in the context of /ih/.

ihbih_cb F1				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	10	-0.4	10	-0.4
2	100	-2	100	-3
3	10	-0.5	60	-9
4	150	-15	100	-9

ihbih_cb F2				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	10	-2	10	-0.5
2	80	-7	0	-0.5
3	100	-7	15	-0.5
4	200	-7	50	-4

Table A.5: Speaker cb: parameter values for the voiced stop consonants in the context of /ih/.

ihdih_cb F1				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	1	-0.5	10	-0.4
2	70	-2	100	-3
3	100	-20	60	-10
4	200	-20	100	-10

ihdih_cb F2				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	1	0.1	50	1.25
2	1	0.1	50	1.25
3	1	0.1	50	1.25
4	1	0.1	50	1.25

Table A.6: Speaker cb: parameter values for the voiced stop consonants in the context of /ih/.

ihgih_cb F1				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	50	-3	10	-0.4
2	50	-3	100	-3
3	100	-3	60	-12
4	50	-3	100	-12

ihgih_cb F2				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	70	2.5	50	1.75
2	70	2.5	50	5
3	100	3	50	5
4	100	3	150	5

Table A.7: Speaker ks: parameter values for the voiced stop consonants in the context of /aa/.

aabaa_ks F1				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	10	-1	10	-0.7
2	50	-3	50	-5
3	50	-4	50	-5
4	100	-6	50	-5

aabaa_ks F2				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	50	-3	50	-3
2	50	-3	50	-3
3	20	-3	20	-3
4	10	-3	10	-3

Table A.8: Speaker ks: parameter values for the voiced stop consonants in the context of /aa/.

aadaa_ks F1				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	10	0.5	10	-0.4
2	70	-2	100	-3
3	100	-5	100	-3
4	200	-5	100	-3

aadaa_ks F2				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	50	2.5	50	1.5
2	100	2.5	50	2
3	100	4	50	2
4	150	6	100	2

Table A.9: Speaker ks: parameter values for the voiced stop consonants in the context of /aa/.

aagaa_ks F1				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	10	-0.5	10	-0.5
2	50	-3	100	-3.5
3	100	-4	10	-3.5
4	200	-5	250	-3.5

aagaa_ks F2				
	to		from	
segment	range (Hz)	slope (Hz/ms)	range (Hz)	slope (Hz/ms)
1	10	2	50	2.5
2	100	2	50	2.5
3	100	6	50	2.5
4	150	7	100	2.5

Appendix B

Tables of Decomposition Results

Table B.1: Summed results for F1 going into the closure in aabaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	626.432560	57.125000	3.083333	15
d	1364.077490	113.325000	1.050000	20
g	626.432560	57.125000	3.083333	15

Table B.2: Normalized results for F1 going into the closure in aabaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	39.152035	3.570313	0.192708	0.941260
d	64.956071	5.396429	0.050000	1.239204
g	39.152035	3.570313	0.192708	0.941260

Table B.3: Summed results for F1 coming out of the closure in aabaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	1524.323389	74.855556	1.952778	37
d	2998.701729	155.985714	1.814286	31
g	2580.077303	152.838636	1.722727	29

Table B.4: Normalized results for F1 coming out of the closure in aabaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	40.113773	1.969883	0.051389	0.649515
d	93.709429	4.874554	0.056696	1.481246
g	86.002577	5.094621	0.057424	1.426912

Table B.5: Summed results for F2 going into the closure in aabaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	782.803703	17.700000	1.400000	60
d	821.316346	19.800000	1.400000	60
g	807.418776	19.800000	1.350000	60

Table B.6: Normalized results for F2 going into the closure in aabaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	12.832848	0.290164	0.022951	0.180296
d	13.464202	0.324590	0.022951	0.190052
g	13.236373	0.324590	0.022131	0.186954

Table B.7: Summed results for F2 coming out of the closure in aabaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	542.828757	12.050000	1.850000	60
d	633.791788	14.050000	1.700000	60
g	646.671789	14.450000	1.700000	60

Table B.8: Normalized results for F2 coming out of the closure in aabaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	8.898832	0.197541	0.030328	0.139070
d	10.390029	0.230328	0.027869	0.154802
g	10.601177	0.236885	0.027869	0.157569

Table B.9: Summed results for F1 going into the closure in aadaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	3351.791488	154.000000	3.100000	42
d	1959.254256	105.826389	2.333333	37
g	4285.877607	247.000000	12.100000	40

Table B.10: Normalized results for F1 going into the closure in aadaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	77.948639	3.581395	0.072093	1.209719
d	51.559323	2.784905	0.061403	0.855487
g	104.533600	6.024390	0.295122	1.942897

Table B.11: Summed results for F1 coming out of the closure in aadaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	4147.031042	355.825000	31.600000	38
d	4570.130950	284.565934	14.173077	49
g	4554.080318	322.888889	13.065874	44

Table B.12: Normalized results for F1 coming out of the closure in aadaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	106.334129	9.123718	0.810256	2.785969
d	91.402619	5.691319	0.283462	1.766620
g	101.201785	7.175309	0.290353	2.019901

Table B.13: Summed results for F2 going into the closure in aadaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	1186.311596	23.250000	1.850000	80
d	1119.359480	23.250000	1.950000	80
g	1147.398409	23.250000	1.900000	80

Table B.14: Normalized results for F2 going into the closure in aadaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	14.645822	0.287037	0.022840	0.198001
d	13.819253	0.287037	0.024074	0.190970
g	14.165412	0.287037	0.023457	0.193815

Table B.15: Summed results for F2 coming out of the closure in aadaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	2457.819271	52.000000	3.450000	120
d	2424.431824	50.750000	3.350000	120
g	2453.611724	51.050000	3.350000	120

Table B.16: Normalized results for F2 coming out of the closure in aadaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	20.312556	0.429752	0.028512	0.274613
d	20.036627	0.419421	0.027686	0.269994
g	20.277783	0.421901	0.027686	0.272654

Table B.17: Summed results for F1 going into the closure in aagaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	2674.778263	100.600000	2.027778	33
d	1035.475867	48.250000	2.027778	15
g	1527.675650	42.600000	2.361111	47

Table B.18: Normalized results for F1 going into the closure in aagaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	78.669949	2.958824	0.059641	1.142222
d	64.717242	3.015625	0.126736	1.075471
g	31.826576	0.887500	0.049190	0.456206

Table B.19: Summed results for F1 coming out of the closure in aagaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	5058.965155	283.995192	13.323077	43
d	4306.600793	262.025000	3.366666	44
g	4311.310589	189.153097	3.661839	47

Table B.20: Normalized results for F1 coming out of the closure in aagaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	114.976481	6.454436	0.302797	2.098006
d	95.702240	5.822778	0.074815	1.614115
g	89.818971	3.940690	0.076288	1.368547

Table B.21: Summed results for F2 going into the closure in aagaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	785.676726	12.900000	1.700000	60
d	773.371474	12.950000	1.450000	60
g	719.186180	12.150000	1.650000	60

Table B.22: Normalized results for F2 going into the closure in aagaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	12.879946	0.211475	0.027869	0.177816
d	12.678221	0.212295	0.023770	0.171782
g	11.789937	0.199180	0.027049	0.164867

Table B.23: Summed results for F2 coming out of the closure in aagaa_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	2068.107637	46.750000	2.650000	81
d	2008.339282	44.300000	2.500000	81
g	2485.344293	60.300000	2.600000	81

Table B.24: Normalized results for F2 coming out of the closure in aagaa_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	25.220825	0.570122	0.032317	0.341538
d	24.491942	0.540244	0.030488	0.329432
g	30.309077	0.735366	0.031707	0.408335

Table B.25: Summed results for F1 going into the closure in ihbih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	250.572697	11.104545	1.363636	51
d	581.155847	29.750000	1.766667	52
g	370.906028	17.066667	1.766667	46

Table B.26: Normalized results for F1 going into the closure in ihbih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	4.818706	0.213549	0.026224	0.095766
d	10.965205	0.561321	0.033333	0.199117
g	7.891618	0.363121	0.037589	0.152817

Table B.27: Summed results for F1 coming out of the closure in ihbih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	988.637496	85.166667	3.716667	86
d	884.716451	45.550000	4.100000	82
g	1644.031205	190.200000	4.400000	82

Table B.28: Normalized results for F1 coming out of the closure in ihbih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	11.363649	0.978927	0.042720	0.254250
d	10.659234	0.548795	0.049398	0.210869
g	19.807605	2.291566	0.053012	0.480245

Table B.29: Summed results for F2 going into the closure in ihbih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	7994.739667	287.600000	11.400000	21
d	8034.568591	288.000000	11.400000	21
g	8015.127682	287.450000	11.500000	21

Table B.30: Normalized results for F2 going into the closure in ihbih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	363.397258	13.072727	0.518182	5.459427
d	365.207663	13.090909	0.518182	5.479349
g	364.323986	13.065909	0.522727	5.472558

Table B.31: Summed results for F2 coming out of the closure in ihbih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	12582.899710	440.250000	21.050000	41
d	12554.647392	437.550000	20.850000	41
g	12972.106252	448.550000	20.850000	41

Table B.32: Normalized results for F2 coming out of the closure in ihbih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	299.592850	10.482143	0.501190	4.545333
d	298.920176	10.417857	0.496429	4.527416
g	308.859673	10.679762	0.496429	4.653001

Table B.33: Summed results for F1 going into the closure in ihdih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	1911.765505	135.550000	2.550000	64
d	1215.997150	74.383333	3.200000	66
g	1212.377731	87.150000	3.450000	65

Table B.34: Normalized results for F1 going into the closure in ihdih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	29.411777	2.085385	0.039231	0.541887
d	18.149211	1.110199	0.047761	0.340273
g	18.369360	1.320455	0.052273	0.368012

Table B.35: Summed results for F1 coming out of the closure in ihdih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	3639.235690	325.937500	23.462500	80
d	2384.904482	221.475000	23.075000	78
g	3516.687905	267.400000	22.450000	81

Table B.36: Normalized results for F1 coming out of the closure in ihdih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	44.928836	4.023920	0.289660	1.141341
d	30.188664	2.803481	0.292089	0.874323
g	42.886438	3.260976	0.273780	1.028742

Table B.37: Summed results for F2 going into the closure in ihdih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	5412.562093	210.742857	12.114286	13
d	5420.985407	209.476190	11.880953	14
g	6093.632107	303.142857	21.714286	8

Table B.38: Normalized results for F2 going into the closure in ihdih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	386.611578	15.053061	0.865306	6.236728
d	361.399027	13.965079	0.792064	5.802562
g	677.070234	33.682540	2.412698	12.551655

Table B.39: Summed results for F2 coming out of the closure in ihdih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	14205.243874	652.700000	41.950000	35
d	12503.188151	477.416667	22.700000	51
g	12593.517657	479.316667	22.700000	51

Table B.40: Normalized results for F2 coming out of the closure in ihdih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	394.590108	18.130556	1.165278	6.924234
d	240.445926	9.181090	0.436538	3.759107
g	242.183032	9.217628	0.436538	3.780132

Table B.41: Summed results for F1 going into the closure in ihgih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	2607.669858	127.100000	4.250000	105
d	3394.617280	309.973077	23.315385	55
g	1311.366342	63.190757	4.050420	93

Table B.42: Normalized results for F1 going into the closure in ihgih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	24.600659	1.199057	0.040094	0.406007
d	60.618166	5.535234	0.416346	1.576051
g	13.950706	0.672242	0.043090	0.249821

Table B.43: Summed results for F1 coming out of the closure in ihgih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	3814.371194	340.700000	33.200000	85
d	4031.780036	369.866667	33.433333	86
g	3700.596843	329.450000	32.650000	81

Table B.44: Normalized results for F1 coming out of the closure in ihgih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	44.353153	3.961628	0.386047	1.225741
d	46.342299	4.251341	0.384291	1.272848
g	45.129230	4.017683	0.398171	1.251231

Table B.45: Summed results for F2 going into the closure in ihgih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	8160.938661	473.800000	32.000000	44
d	10959.593591	435.681818	22.063636	56
g	5536.859648	423.100000	41.100000	40

Table B.46: Normalized results for F2 going into the closure in ihgih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	181.354192	10.528889	0.711111	3.577542
d	192.273572	7.643541	0.387081	3.074171
g	135.045357	10.319512	1.002439	3.384844

Table B.47: Summed results for F2 coming out of the closure in ihgih_cb

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	10097.423211	548.050000	41.600000	62
d	9904.635802	415.766667	22.533333	85
g	8881.003473	374.166667	23.300000	85

Table B.48: Normalized results for F2 coming out of the closure in ihgih_cb

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	160.276559	8.699206	0.660317	3.133004
d	115.170184	4.834496	0.262016	1.897167
g	103.267482	4.350775	0.270930	1.738683

Table B.49: Summed results for F1 going into the closure in aabaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	890.417720	34.250000	1.450000	40
d	579.090718	22.450000	1.550000	40
g	579.090718	22.450000	1.550000	40

Table B.50: Normalized results for F1 going into the closure in aabaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	21.717505	0.835366	0.035366	0.336077
d	14.124164	0.547561	0.037805	0.233803
g	14.124164	0.547561	0.037805	0.233803

Table B.51: Summed results for F1 coming out of the closure in aabaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	1433.245265	76.941667	1.519444	45
d	953.249085	48.123684	2.378070	45
g	1144.564061	67.372549	2.078431	35

Table B.52: Normalized results for F1 coming out of the closure in aabaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	31.157506	1.672645	0.033031	0.511871
d	20.722806	1.046167	0.051697	0.363542
g	31.793446	1.871460	0.057734	0.562815

Table B.53: Summed results for F2 going into the closure in aabaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	389.804044	10.100000	1.100000	40
d	632.517969	15.300000	1.300000	40
g	632.517969	15.300000	1.300000	40

Table B.54: Normalized results for F2 going into the closure in aabaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	9.507416	0.246341	0.026829	0.146538
d	15.427268	0.373171	0.031707	0.223297
g	15.427268	0.373171	0.031707	0.223297

Table B.55: Summed results for F2 coming out of the closure in aabaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	305.182253	9.850000	1.700000	60
d	421.428993	13.400000	1.900000	60
g	438.400660	13.650000	1.600000	60

Table B.56: Normalized results for F2 coming out of the closure in aabaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	5.002988	0.161475	0.027869	0.094046
d	6.908672	0.219672	0.031148	0.122201
g	7.186896	0.223770	0.026230	0.120476

Table B.57: Summed results for F1 going into the closure in aadaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	2441.432986	178.739286	12.946104	61
d	2076.325968	93.525000	2.583333	79
g	3652.266738	146.275000	3.350000	96

Table B.58: Normalized results for F1 going into the closure in aadaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	39.377951	2.882892	0.208808	0.890877
d	25.954075	1.169062	0.032292	0.408739
g	37.652234	1.507990	0.034536	0.561857

Table B.59: Summed results for F1 coming out of the closure in aadaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	4994.853257	387.550000	31.872222	58
d	1846.762668	80.761765	3.902941	101
g	2685.967642	112.629412	4.576471	117

Table B.60: Normalized results for F1 coming out of the closure in aadaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	84.658530	6.568644	0.540207	2.043657
d	18.105516	0.791782	0.038264	0.298497
g	22.762438	0.954487	0.038784	0.361857

Table B.61: Summed results for F2 going into the closure in aadaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	2312.038684	208.950000	21.350000	60
d	1538.803709	36.250000	2.616667	95
g	1599.213837	40.600000	2.971429	94

Table B.62: Normalized results for F2 going into the closure in aadaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	37.902274	3.425410	0.350000	1.071564
d	16.029205	0.377604	0.027257	0.225309
g	16.833830	0.427368	0.031278	0.242353

Table B.63: Summed results for F2 coming out of the closure in aadaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	4499.095767	350.916667	31.650000	52
d	4716.812818	157.550000	3.736364	92
g	6059.474492	200.800000	3.150000	86

Table B.64: Normalized results for F2 coming out of the closure in aadaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	84.888599	6.621069	0.597170	2.108163
d	50.718417	1.694086	0.040176	0.716769
g	69.649132	2.308046	0.036207	0.963503

Table B.65: Summed results for F1 going into the closure in aagaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	4219.032709	323.411538	24.325962	74
d	2437.737936	109.546079	4.487255	124
g	1966.952847	89.817647	4.772374	124

Table B.66: Normalized results for F1 going into the closure in aagaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	56.253769	4.312154	0.324346	1.318099
d	19.501903	0.876369	0.035898	0.318554
g	15.735623	0.718541	0.038179	0.267389

Table B.67: Summed results for F1 coming out of the closure in aagaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	4536.239403	366.166667	31.721568	40
d	3749.955866	178.721710	4.058991	96
g	2287.964334	144.784066	3.499634	67

Table B.68: Normalized results for F1 coming out of the closure in aagaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	110.639985	8.930894	0.773697	2.773186
d	38.659339	1.842492	0.041845	0.612688
g	33.646534	2.129177	0.051465	0.600848

Table B.69: Summed results for F2 going into the closure in aagaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	3440.629810	313.600000	32.100000	72
d	1798.749936	39.400000	4.000000	140
g	1463.869800	34.675000	4.000000	136

Table B.70: Normalized results for F2 going into the closure in aagaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	47.131915	4.295890	0.439726	1.340634
d	12.757092	0.279433	0.028369	0.183883
g	10.685181	0.253102	0.029197	0.161359

Table B.71: Summed results for F2 coming out of the closure in aagaa_ks

model	ϱ_1	ϱ_2	ϱ_3	ϱ_N
b	2532.168974	216.500000	149.700000	47
d	4540.803860	97.850000	3.250000	120
g	3904.388192	98.913636	3.718182	111

Table B.72: Normalized results for F2 coming out of the closure in aagaa_ks

model	$\varrho_1/1+\varrho_N$	$\varrho_2/1+\varrho_N$	$\varrho_3/1+\varrho_N$	ϱ
b	52.753520	4.510417	3.118750	4.097327
d	37.527305	0.808678	0.026860	0.483000
g	34.860609	0.883157	0.033198	0.470120

Bibliography

[1] Peter F. Assmann. The role of formant transitions in the perception of concurrent vowels. *J. Acoust. Soc. Am.*, 97(1), January 1995.

[2] Martin J. Bastiaans. Gabor's expansion of a signal into gaussian elementary signals. *Proceedings of the IEEE*, 68(4), April 1980.

[3] John J. Benedetto and Michael W. Frazier, editors. *Wavelets: Mathematics and Applications.* Studies in Advanced Mathematics. CRC Press, 1994.

[4] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control.* Athena Scientific, 1995.

[5] Leo Breiman. *Probability.* Classics in Applied Mathematics. SIAM, 1992.

[6] Herman Chernoff. Notes. August 1977.

[7] Ingrid Daubechies. *Ten Lectures on Wavelets.* CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, 1992.

[8] John R. Deller, Jr., John G. Proakis, and John H. L. Hansen. *Discrete-Time Processing of Speech Signals.* Macmillan Publishing Company, 1993.

[9] J.L. Doob. *Measure Theory.* Graduate Texts in Mathematics. Springer-Verlag, 1994.

[10] Richard O. Duda and Peter E. Hart. *Measure Theory.* John Wiley & Sons, 1973.

[11] Jerome H. Friedman and Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376), December 1981.

[12] Mark Allan Hasegawa-Johnson. *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification.* PhD thesis, Massachusetts Institute of Technology, 1996.

[13] Peter J. Huber. Projection pursuit. *The Annals of Statistics*, 13(2), 1985.

[14] Lee K. Jones. On a conjecture of huber concerning the convergence of projection pursuit regression. *The Annals of Statistics*, 15(2), 1987.

[15] Gary E. Kopec. Formant tracking using hidden markov models and vector quantization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4), August 1986.

[16] Shan Lu and Peter C. Doerschuk. Nonlinear modeling and processing of speech based on sums of am-fm formant models. Technical report, School of Electrical Engineering, Purdue University, 1995.

[17] Stéphane Mallat. *Wavelet Signal Processing.* Academic Press, 1996.

[18] Stéphane G. Mallat and Zhang Zhifeng. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12), December 1993.

[19] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing.* Signal Processing Series. Prentice-Hall, 1989.

[20] Emanuel Parzen. Statistical inference on time series by hilbert space methods, i. Technical report, Applied Mathematics and Statistics Laboratory, Stanford University, January 1959. Technical Report No. 23.

[21] Joseph W. Picone. Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9), September 1993.

[22] M. B. Priestley. *Spectral Analysis and Time Series.* Probability and Mathematical Statistics. Academic Press Inc., 1981.

[23] M. B. Priestley. Wavelets and time-dependent spectral analysis. Technical report, Department of Statistics, Stanford University, April 1995. Technical Report No. 311.

[24] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals.* Signal Processing Series. Prentice-Hall, 1978.

[25] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Signal Processing Series. Prentice-Hall, 1993.

[26] J. Rissanen. Shannon-wiener information and stochastic complexity. Technical report, IBM Almaden Research Center. From a talk.

[27] Stephanie Seneff. An auditory-based speech recognition strategy: Application to speaker independent vowel recognition. In *From the Proceedings of the Speech Recognition Workshop*, 1986.

[28] Kenneth Stevens. 6.541 course notes. To be published.

[29] Kenneth Stevens. Personal communication.

[30] Bart M. ter Haar Romeny, editor. *Geometry-Driven Diffusion in Computer Vision.* Computational Imaging and Vision. Kluwer Academic Publishers, 1994. David Mumford, pp. 135-146.