A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features

Li Deng [a)b)] *and Don X. Sun* [c)]

Spoken Language Systems Group, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 [a)]

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1 [b)]

Department of Applied Mathematics and Statistics State University of New York at Stony Brook, NY 11794-3600 [c)]

Abstract

We have in the recent years been involved in the development of a feature-based general statistical framework for automatic speech recognition via novel designs of minimal or atomic units of speech, aiming at a parsimonious scheme to share the inter-word and inter-phone speech data and at a unified way to account for the context-dependent behaviors in speech. Our basic design philosophy has been motivated by the theory of distinctive features (Chomsky and Halle, 1968; Stevens, 1986) and by a new form of phonology which argues for use of multi-dimensional articulatory structures (Browman and Goldstein, 1992). In this paper, we present the feature-based recognizer developed most recently, which is capable of operating on all classes of English sounds. We provide detailed descriptions of the design considerations for the recognizer and of key aspects of the design process. This process, which we call lexicon "compilation", consists of three elements: 1) establishing a feature-specification system; 2) constructing a probabilistic and fractional temporal overlapping pattern across the features; and 3) mapping from the feature-overlap pattern to a state-transition graph. A standard phonetic classification task from the TIMIT database is used as a test-bed to evaluate the performance of the recognizer. The experimental results provide preliminary evidence for the effectiveness of our feature-based approach to speech recognition.

Introduction

For the past several years, an approach to automatic speech recognition has been pursued in which consistent and reliable speech knowledge at a relatively high level is utilized to explicitly determine the general properties and the global structure of the recognizers, while at a more detailed level where speech knowledge tends to be less reliable, statistical methods and automatic learning procedures are called for (e.g., Deng and Erler, 1992; Deng, et al., 1992; Deng, Lennig, and Mermelstein, 1989; Phillips, Glass, and Zue, 1992; Zue et al., 1990) The statistical component of the recognizers is intended to accommodate the "residual" speech variability which is difficult to be accounted for by the general properties. This strategy is robust in performance — free from catastrophic types of errors due to inadequacy in speech knowledge — compared with the purely knowledge-based approaches on the one hand, and it imposes only a minimal demand on the training data compared with the purely statistical approaches on the other hand. As more and more reliable speech knowledge is accumulated and incorporated, the statistical component of the recognizers is expected to account for a diminishing amount of variability.

Advances in speech science over the past four decades have significantly enhanced our knowledge of the speech production mechanisms, and of the phonological aspects of features and articulatory gestures that serve a basis for describing the underlying regularities of speech. We believe that the knowledge acquired from these advances is sufficiently reliable and can hence be utilized for the design of automatic speech recognizers. This forms the principal motivation for the study to be described in this paper.

Utilization of speech knowledge advanced from speech science has been argued to be beneficial for automatic speech recognition (Zue, 1985); this should especially be the case when dealing with large vocabularies where highly confusible word pairs abound and acoustic variability due to phonetic contexts is particularly conspicuous. In earlier works on the development of a large-vocabulary word recognition system, such acoustic variability was found to be one major obstacle to the achievement of high word recognition accuracy (Deng et al., 1991; Deng et al., 1990). We believe that solutions to this contextual variation problem should lie in devising ways in which acoustic data can be intelligently shared across a large number of lexical items at a microsegmental level. That is, the atomic or the minimally sharable microsegmental (sub-phonemic) speech units are to be selected in such a way that the contextual information is automatically assimilated into the chosen units and their acoustic manifestations should show a maximal degree of consistency; also, the total number of such units must be sufficiently small so as not to impose too stringent a demand on the training data. In this study, we present and evaluate one general framework aiming at such a goal. The development of this framework has been guided by knowledge of acoustic phonetics (e.g. Zue, 1991), and inspired by the advantages of using phonological features for lexical access (e.g. Stevens, 1986) and by a new form of phonology which advocates the use of multi-dimensional articulatory gestures to unify phonetic and phonological descriptions of speech (e.g. Browman and Goldstein, 1986; Browman and Goldstein, 1992). In particular, the framework is built on the basis of using constellations of overlapping articulatory features as the atomic speech unit and using statistical methods to provide mapping from these symbolic speech primitives to the acoustic observations of speech.

One first attempt to model microsegments of speech for the purpose of large vocabulary word recognition was made in the study of (Deng, Lennig, and Mermelstein, 1990), where each word-initial and word-final stop consonant was decomposed into several, possibly optional, microsegments. Although many acoustically similar events associated with the production of stop consonants are shared across varying phonetic contexts, the phoneme-to-microsegment decomposition process lacks generalizability for wider classes of consonant sounds. Further, for the transitional microsegments from consonant to vowel (CV) and from vowel to consonant (VC), the intended data-sharing goal cannot be easily achieved. In a further study, improvement of the data-sharing mechanism for the transitional microsegments was made where the parameters in the statistical model associated with the acoustic transitions in each CV and VC were constrained to be the interpolated values between the "target" parameters associated with the consonant and those with the vowel (Deng et al., 1992). In both of the studies, moderate improvement in recognition accuracy for the vocabulary of the size of 86,000 words was obtained.

The above microsegment modeling approaches were recently extended and recast into a different scheme where the microsegments for stop CV syllables were defined automatically based on a small set of empirically chosen, multi-valued phonetic and articulatory features (Deng and Erler, 1992). Although the scope of that study was necessarily limited to only the stop CV syllables, due mainly to the rather rigid way in which the syllables are represented in terms of the feature compositions, the accuracy for stop discrimination obtained by using such feature-based speech units was shown to be consistently higher than using several conventional segmental speech units. The same superiority in the discrimination accuracy has been more recently extended to a somewhat wider set of utterances — stop-vowel-stop (CVC) words (Erler and Deng, 1993).

To be able to create a general framework in which the idea of using feature-based speech units to design the recognizer's structure can be applied to unlimited-vocabulary speech recognition, we need to solve three grave problems. First, a full set of features need to be defined which, together with the hypothesis of asynchronous feature overlap, are capable of describing a considerably richer set of co-articulatory phenomena than those encountered in our previous CV and CVC (C = stop) cases. The challenge is to choose the features and to assign them to the phonemic units in such a way that they are appropriate to distinguish different articulatory structures but are sufficiently orthogonal to minimize the representational redundancy. Second, extension of the vocabulary requires that more intelligent and efficient strategies than that described by Deng and Erler (1992) be devised for lexical representation, i.e., representing each lexical item in terms of a distinct collection of the atomic speech units each being a constellation of the overlapping articulatory features. An issue closely related to lexical representation is how to determine the extent to which the various articulatory features are coordinated and are allowed to spread and to overlap. Third, it remains essential to design evaluation tests to determine whether the feature-based speech recognition strategy is practically advantageous over the conventional acoustic modeling strategies for general, large vocabulary speech recognition applications.

The principal aim of this study is to address all these three issues to the extent that a general feature-based computational framework can be established for implementing practical speech recognition systems. The organization of this paper is as follow. In Section I, we provide an overview of the general structure of the feature-based speech recognizer experimented in this study, with emphasis on the similarities and differences between the recognizers utilizing the standard segmental speech units and the feature-based units. ¿From Sections II to IV, we give detailed treatment of the above three issues sequentially — feature specification, lexical representation, and experimental evaluation. Finally, in Section V, we summarize the salient properties of our feature-based recognizer design and point out future research directions.

I. Overview of the Feature-Based Speech Recognizer Design Considerations

As for the design of any speech recognizer, two major issues are of concern in our current design of the feature-based speech recognizer. First, we are concerned with lexical representation; that is, we desire to choose a small — compared with the size of the vocabulary — set of speech units such that in terms of these symbolic representations each lexical item can be specified and be made distinguished from each other. The second issue is the mapping from the abstract speech units, which are of symbolic nature, to the corresponding physically measurable, continuous acoustic stream in the speech signal. Given the decisions made to address these two issues, the standard Bayesian strategy (which we take) can be used to implement "optimal" lexical access in a computationally efficient manner.

Desirable choices for lexical representation and for acoustic mapping should satisfy the two requirements: (1) Each speech unit is mapped into a (relationally) invariant, "consistent" acoustic region, and consequently distinct speech units are mapped into maximally separate acoustic regions; (2) The inventory of the speech units is sufficiently small such that different lexical items are allowed to maximally (and beneficially) share the available acoustic data and hence the demand for training data can be kept minimal. Speech units based on entire words satisfy the first requirement well but satisfy the second requirement poorly . The opposite is true for the context-independent phonemic speech units. The popular triphone-like speech units, or context-dependent allophonic units (Deng et al., 1990; Lee and Hon, 1990; Schwartz et al., 1984), appear to satisfy both of the requirements well, but only for small to medium-sized vocabularies (up to 1000 words or so). The decision-tree method appears to be another type of compromise but fulfillment of the first requirement above is questionable. A recent proposal of using "senones" (Hwang, Huang, and Alleva, 1993) as the atomic speech unit is moving towards the direction of providing maximally sharable acoustic information among varying lexical items, a goal in common with our feature-based speech recognition framework. The senone approach achieves such a goal via a data-driven semi-automatic method. In contrast, our feature-based approach is built on an elaborate articulatory timing and feature spreading model constructed in accordance with the theories of speech production and of phonology.

Although lexical representation and acoustic mapping should be viewed as integrated components in the speech recognizer design, the system developed during the span of the current study has its main novelty mainly in the lexical-representation component. The recognizer utilizes the mathematical structure known as the hidden Markov model (HMM) (Baum, 1972; Rabiner, 1989) to link the lexical-representation component and the acoustic-mapping component. In contrast to the conventional use of the HMMs having a fixed left-to-right topology, the underlying Markov chain of our feature-based HMM serves the critical role of tracking the temporal evolution of the articulatory features — a heavily quantized, simplistic picture of the phonologically meaningful articulators' motions within a specialized "coordinate" system (see Section II). This is accomplished by using constellations (bundles) of overlapping articulatory features as the atomic speech unit represented by the states in the HMM. In our current implementation, we use five multi-valued features, Lips, Tongue Blade (TB), Tongue Dorsum (TD), Velum, and Larynx, which are assigned uniquely to each quasi-phonemic unit with intended "minimal" redundancy and "maximal" separability. Then the major contextual variations in speech are modeled as a natural result of overlap or asynchronous spread of the "intrinsic" values of one or more of these features across adjacent phonetic units. Given a fixed feature-overlap pattern, a one-to-one mapping can be made from such a pattern to a state-transition graph which forms the topology of the underlying Markov chain of the HMM. The graph is constructed in such a way that each node in the graph (or the state in the HMM) represents a unique composition of the five features (see Section III for detail). Each individual lexical item is represented by a distinct state-transition graph, which is formed by concatenating a sequence of sub-graphs associated with the phone sequence in the phonetic transcription according to the feature-overlap patterns constructed specifically from these phones-in-context. Since each node in this global graph contains a distinct composition of the features, we can also view the representation of a lexical item described above as an organized set of feature-bundle collection.

For the acoustic-mapping component, we adopt the conventional statistical methods: for the frame-based recognition scheme which we will describe, the acoustic mapping is accomplished by sequentially sampling from mixture Gaussian distribution(s) whose parameters are dependent on the underlying feature composition. (Similar mechanisms can be applied to the segment-based recognition schemes where the distribution is made associated with a single segmental speech unit and the sampling rate is fixed at one per segment.) This is implemented in the HMM framework by associating each HMM state with a mixture Gaussian output distribution. In principle, the acoustic mapping component of the recognizer can be made deterministic provided that the acoustic correlates of the feature constellations as the speech unit are fully understood and can be extracted with no error. This, however, is obviously not the case and hence the statistical mapping as an ignorant but practical implementational tool is called for. The statistical distribution, as conditioned on each HMM state (for a fixed feature composition), is used to accommodate the residual speech variability that cannot be accounted for by the symbolic feature-overlapping construction in the recognizer (which is intended to account for contextual variability in speech). At least two clear sources of such residual variability can be identified: (1) the variability due to speaker variation and (2) the variability due to use of simplistic and uniform acoustic measurements (see Section IV for detail) independent of the speech class hypothesized during the recognition process. Being easily extractable, these measurements nevertheless lack clear perceptual correlates. Both of these factors, in additional to possibly many others, necessarily introduce variability in the acoustic parameter values. Thus, until these various sources of the residual variability are adequately accounted for in the future research, flexible ways of modeling the statistical variation in the state-dependent acoustic parameters are necessary.

Compared with the conventional ways of modeling co-articulation using the allophonic speech units, our feature-based strategy is more parsimonious in that the co-articulatory or context dependent phenomena can be accounted for in a more natural manner and thus fewer model parameters are required. The conventional speech recognition methods assume that the acoustic variability in speech is expressed in terms of non-overlapping segments and that direct correspondence is made between the segment and its acoustic manifestation. Our feature-based framework removes the above assumptions and treats the speech as a result of continuously overlapping sequences of articulatory commands which are represented by the temporal evolution of features. In this way, the acoustic variation observed in speech is accounted for by the unified overlapping mechanism operative on the various articulatory features, rather than relying on enumeration of the explosively large number of contextual factors based solely on acoustic descriptions.

II. A System for Specifying Articulatory Features

As mentioned in Section I, the design of the lexical-representation component of the feature-based recognizer begins with specification of a full set of articulatory features which are to be assigned to each phoneme-like unit. Our choice of the feature set is motivated by the work of Browman and Goldstein in their development of articulatory phonology (Browman and Goldstein, 1986; 1992), enhanced by practical considerations for the recognizer implementation. The five articulatory features we choose to use are Lips, Tongue blade (TB), Tongue Dorsum (TD), Velum, and Larynx, similar to the five tract variables (Lip protrusion/aperture, Tongue Tip constrict location/degree, Tongue Body constrict location/degree, Velic aperture, and Glottal aperture) described in Browman and Goldstein (1990) for a gestural framework in speech synthesis. (For convenience, we shall call the above first three features as the principal-articulator features.) Like speech synthesis, we believe that these five articulatory features are also adequate for the purpose of speech recognition, as they are capable of describing major co-articulatory phenomena once appropriate assumptions are made about asynchronous overlap of these features over temporal and spatial dimensions (see examples in Section III.D).

Unlike models for speech synthesis, however, where the articulators' positions are usually specified in terms of geometric coordinates (e.g. Mermelstein, 1973; Saltzman, 1986), in speech recognition applications, the detailed values that the articulatory features are allowed to take must be connected, either directly or implicitly, to the phonological labels for the utterances. Although it may be desirable to accomplish specification of the features with binary values enhanced by some kind of hierarchical (tiered) structures such as those proposed in (Clements, 1985; Halle and Stevens, 1991; McCarthy, 1988), in this study we decide to adopt a more straightforward but highly economical feature specification system: multi-valued features with only a single-layered structure.

Such a feature specification system is illustrated in Table 1 for the consonantal segments and in Table 2 for the non-consonantal segments in English [1]. In the following, we provide six notes on the conventions used and of key properties of the feature specification system summarized in Tables 1 and 2.

First, the special feature value denoted by "0" for the phonetic or quasi-phonemic segment in Tables 1 and 2 indicates that in attempting to produce the sound this particular articulatory dimension is unmarked (or irrelevant). For instance, Lips feature is unmarked for all vowel segments except for those with lip rounding (feature value "3" for /uw/ and "4" for /ao/), and labial consonants are all unmarked in both TB and TD features. Unmarked features are most susceptible to feature spreading (assimilation) from the adjacent phonetic segments; that is, segments having unmarked feature(s) tend to inherit these features from the adjacent segments.

Second, for each of the five articulatory features, we provide a range of feature values (i.e. multiple quantization levels) — five for Lips, seven for TB, 20 for TD, two (binary) for Velum, and three for Larynx. Within the scope of the present study, the feature assignments should be interpreted as of a symbolic nature only without ordering and relational information (see Section V for discussions on a more general interpretation for future research).

Third, the specified features for each quasi-phonemic segment is context-independent, which are related to the intended target articulatory gestures, or the abstract "hidden" messages to be conveyed through the speech production process. For example, the intended gestures for phoneme /t/ is to move the tongue blade in contact with alveolar ridge anteriorly (value "1" in *TB* feature in Table 1), to close the velopharyngeal port (value "1" in *Velum* feature), and to abduct the vocal cords (value "2" in *Larynx* feature), all independent of the contexts [2]. Despite allophonic variations, the above intended gestures for /t/ tend to be invariant [3].

Fourth, in the current feature specification system, the systematic nonstationary segments (or complex segments) are decomposed into relatively stationary sub-segments. Specifically, affricates are decomposed into their associated stop and fricative sub-segments (last two rows in Table 1), and diphthong specification is accomplished by sequential connection of the two target vowels (see double labels in either Lips or TD feature for diphthongs in Table 2). Such specifications for the complex segments imply that their feature spread is extended over only the left or right, rather than both, adjacent segments.

Fifth, all vowels are marked in the TD feature as the tongue dorsum is always active in producing these sounds; for lip-rounding vowels, the Lips feature is marked as well. The TB feature is unmarked for vowels except for /el/ and /er/. Semi-vowels, on the other hand, are marked by two or all three principal articulator features.

Finally, as TIMIT database will be used for performance evaluation of the recognizer (Section IV), we follow the TIMIT convention and adopt the TIMIT labels to denote all the phonetic segments in Tables 1 and 2. The TIMIT labels have a small degree of deviation from the standard phonemic labels; for example, the closure interval of a stop is made distinguished from the stop release, and some rather systematic allophones — flaps /dx/ and /nx/, glottal stop /q/, voiced-h /hv/, fronted-u /ux/, and devoiced-schwa /ax-h/ — are added as separate labels. We call this expanded set of phonetic segments as quasi-phonemic ones. Since the additional allophonic segments are of a small number and they also tend to possess rather consistent acoustic manifestations, it is felt that there is no need, at least in the current system to be evaluated on the TIMIT task, to convert them back to the purely phonemic segments for the purpose of constructing a high-performance speech recognizer.

III. Lexical Representation and Context-Dependence Modeling

A. Fractional overlap of articulatory features

Once the feature-specification system is established, we have accomplished an "abstract" lexical representation, which is in the same form as the feature-matrix representation described in Chomsky and Halle (1968) and in Stevens (1986). The next step is to "compile" such an abstract representation, based on the concept of probabilistic and fractional overlap of features or articulatory gestures, so as to associate the lexical representation with the acoustic attributes and to allow the process of "learning from data" to take place in constructing an automatic speech recognizer. Allowing the overlap to be fractional and providing a probabilistic structure to the fractional overlap in our current feature-overlap model generalize the notion of feature spreading discussed in Stevens (1986) and that of gestural overlap discussed in Browman and Goldstein (1992).

Figure 1 is a schematic showing results of the feature-overlapping process for an example phrase /t eh n k ae t s/ (*ten cats*). This process starts with a feature-matrix representation of the utterance according to the feature specification system described in Section II; that is, for each quasi-phonemic segment (adjacent ones are segregated by the hypothetical vertical dashed lines for illustration purposes) and for each of the five features, the related entry ("symbolic" feature value) in Tables 1 or 2 is entered into the matrix as shown in the figure. We then allow these feature values to spread from a "home" segment into the adjacent segments, with a varying degree of spread depth (hence fractional overlap), resulting in asynchronous alignment or temporal overlap among different features. The shaded horizontal boxes schematize the pattern of the feature overlap, which, in general, can be in both left and right directions. Left spreading of features gives rise to anticipatory co-articulation while right spreading gives rise to carry-over co-articulation. For instance, the TD feature associated with /k/ (value "1" as in Figure 1) is spread along both directions: the spread into the left segment /n/ causes palatalization of the alveolar nasal /n/; while the spread into the right segment /ae/ results in acoustic transition (e.g. formant transition) within the /ae/ segment. In this example, the palatalization of /n/ can be either complete or partial, depending on whether or not the TD feature "1" is spread through the entire /n/ segment (Figure 1 shows only partial or fractional palatalization, i.e. only a fraction of the /n/ segment adjacent to /k/ is palatalized into /ng/-like sound). Likewise, the acoustic transition in /ae/ depicted in Figure 1 spans about half of its entire segment. This flexibility of allowing for fractional or partial feature spread from the home segment into others appears to have included the feature-modification scheme discussed in Stevens et al. (1992) as a special case where it was assumed that feature modification, when implemented, is necessarily complete.

Note that Figure 1 is intended to illustrate only one specific realization of the feature overlap, which is for one particular token of the example utterance. In our feature-overlap model which is intended to describe the statistical behavior of a collection of all instances of an utterance, we model the degree to which features are allowed to spread into the adjacent segments (i.e. the depth of feature spread) as being probabilistic. That is, the features associated with different tokens of the same utterance are modeled to spread from the home segment into the adjacent segments with variable depths, where the exact fraction of the spread depth is governed by a probability distribution. In the current model, such a probability distribution is only implicit in that it is determined indirectly by the collection of the output distributions each associated with a distinct feature composition (i.e. a state in the HMM), and by the organization of these output distributions according to the Markov chain's topology. Computationally, the actual realization of the feature-overlap depth for any given token of an utterance can be estimated by a Viterbi-like decoding algorithm [4].

In addition to the examples of palatalization of /n/ and formant transition of /ae/ just discussed, Figure 1 also illustrates following obvious context-dependent phenomena. First, TB feature associated with the phrase initial /t/ (value "1") is spread into the /eh/ segment, resulting in formant transition in /eh/ on its left portion. A similar formant-transition pattern also appears on the right portion due to the same TB feature spreading from right. Note that /t/ and /n/ are designed to share the same TB feature value "1" (see Table 1) as their production involves the same underlying gesture for the tongue blade articulator: being in touch with alveolar ridge anteriorly. As such, acoustic transitions induced by them are modeled to be the same (except that the directions of the transitions can be constrained to be in reverse). Also, it is worth pointing out that the acoustic transitions induced by spreading the TB feature are mainly a result of the motion of the tongue dorsum after the constriction formed by tongue-blade is released. This interpretation of the TB feature spread might offer some opportunities for constraining the spread of both TB and TD features by considering their possible interactions [5]. Second, the Larynx feature associated with the /t/ (value "2" denoting devoicing according to Table 1) is spread into /eh/ segment, resulting in an aspiration microsegment in /eh/. Third, the Velum feature associated with the /n/ (value "2" denoting plus-velum according to Table 1) is spread from right into /eh/ segment, resulting in a portion of nasalized vowel. Combining the consequences of the above three separate ways of feature spreading, we can easily identify the feature compositions and their associated acoustic properties for each time slice within the /eh/ segment. As two examples, the time slice denoted by S1 contains the feature composition (0 1 9 1 2) (one for each feature organized in the same order as in Tables 1 and 2) and can be interpreted as a sample from the initial portion of /eh/ which exhibits /t/-induced transitional acoustics and is devoiced. Similarly, the time slice denoted by S2 contains the feature composition (0 1 9 2 1) and is sampled from a final portion of /eh/ which exhibits /n/-induced (same as /t/-induced) transitional acoustics and is nasalized. Fourth, the TD feature associated with the /ae/ (value "10") is spread from right into /k/ segment, resulting in a fronted /k/ burst [6]. This is illustrated as the time slice S4 with feature composition (0 0 10 1 2). Finally, time slice S3, which contains feature composition (0 0 1 2 1), corresponds to palatalized /n/ with the palatalization caused by TD feature spread from the /k/ segment. Referring to Table 1, we see that the velar nasal /ng/ is specified identically to the feature composition associated with time slice S3.

B. Constraints on feature overlaps

The scheme for feature overlaps across quasi-phonemic segments described in Section III.A necessarily generates many feature-overlapping patterns which are not permissible in reality. To avoid such over-generation, we must impose constraints on feature spreads. With regard to the nature of the constraints, they appear to be in the form of the relative timing or phasing relationship among the various features representing the actions of the corresponding articulatory gestures. For example, the study of Zue and Sia (1982), based on the accelerometer measurement data, suggests that in producing vowel-nasal-fricative (homorganic in nasal-fricative) clusters in American English, the spread in the Velum feature (action of velum open and closure) from the nasal segment to the vowel segment is closely phased with the voicing character (the Larynx feature) in the nasal segment and with the oral-closure gesture (which determines the boundary between the vowel and the nasal segments). The effect of the constraints is to weed out the feature compositions (or the states in the HMM) which reflect no physical reality of articulation in terms of the possible coordination among the various articulatory gestures. This is necessary for the design of robust recognizers, which are characterized by only the relevant and useful free parameters and by having these free parameters organized as tidily as possible. A rich source of knowledge in speech production and in speech perception can be utilized as the basis for deriving such constraints.

One critical issue to be addressed is how to devise mechanisms by which feature spreading is permitted in the model in realistic terms. There are at least three ways the constraints can be imposed. First, as a most straightforward approach, a set of rules can be provided to stipulate what kinds of feature spreading are permitted and what are not. Second, since choice of the features is motivated by articulatory gestures (Section II), parametric models which describe coordination of various articulators' motions during the speech production process can be used as a basis for providing the feature-overlap constraints [7]. Third, effective constraints can be achieved algorithmically by "soft tying" parameters associated with large sets of feature compositions according to some known quantal properties in the articulatory-to-acoustic relations of speech production (Stevens, 1989).

To restrict the scope of this study, we will present the result of the first method. A simple set of rules implemented for constraining feature overlaps and the phonetic classification results obtained using various combinations of these rules will be described in Section IV.D.

C. Construction of state-transition graph from feature-overlap pattern

Given the "abstract" lexical representation by the feature matrix form according to the feature-specification system and given the derived feature-overlap pattern described above, the final step for completing the "compiled" lexical representation for an utterance is to establish a mapping from the feature-overlap pattern to a state-transition graph. Such a graph can be used to serve as a general phonological component of a speech recognizer. For the HMM-based recognizer as has been implemented in this study, the graph has been used as the topological structure of the (hidden) Markov chain (with self-transition loops added), which represents the possible manipulation of the underlying articulatory structure responsible for producing the utterance.

We have implemented a recursive procedure to efficiently implement the mapping from a feature-overlap pattern to a state-transition graph. The implemented mapping has the properties 1) that the mapping is unique and 2) that all the temporal phasing relations represented in the feature-overlap pattern are respected among the various articulatory features. The first property of uniqueness implies that the mapping makes neither overgeneration nor undergeneration; that is, the feature composition of each state (or node) in the resulting state-transition graph can always find its presence in the feature-overlap pattern *and* conversely each feature composition which can be identified in the allowable feature-overlap pattern can also find its corresponding node in the resulting graph. The second property ensures further that each transition in the state-transition graph is lawful; that is, any pair of feature compositions (representing two distinct articulatory configurations) which must be reached from one to another through any possible intermediate feature composition(s) according to the temporal sequencing relations expressed in the feature-overlap pattern are necessarily disconnected.

The recursive procedure in an algorithmic form is outlined below in this subsection. The next subsection will show the results of this procedure for an example utterance.

Let the feature matrix for an n-phone utterance be

phone 1	phone 2	$\cdots$	phone n
f_{11}	f_{21}	$\cdots$	f_{n1}
f_{12}	f_{22}	$\cdots$	f_{n2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$
f_{1D}	f_{2D}	$\cdots$	f_{nD}

Then the following procedure produces the state-transition graph for each phone-in-context:

1. Initialization:

   (a) Initialize the state-graph to be empty and set *depth* =0.

   (b) Form an initial dummy state $S^* = (f_{11}, f_{12}, , \cdots f_{1D})$.

2. Recursion:
   take a dummy state as input;

   (a) Apply all the feature-overlap rules to the incoming dummy state S^*.

   (b) If all the rules are satisfied by the dummy state, then

      i. The dummy state turns to an active state.

      ii. The state-graph is updated by this most current active state with the current value of *depth*.

      iii. $depth \leftarrow depth + 1$.

   (c) For each feature dimension, move one step right in the feature array and form a set of new dummy states S_d^*, $d = 1, \cdots, D$.

   (d) For each dummy state S_d^*, $d = 1, \cdots, D$ with the current value of *depth*, assign $S^* = S_d^*$ and go to step 2.

3. Termination: Stop if no dummy states can be generated in step (c) of step 2.

D. State-transition graph for an example utterance and its phonetic interpretation

In this subsection, we use the example of Figure 1 to demonstrate the construction of the state-transition graph from the feature-overlap pattern. The construction is made one for each phone-in-context, which can be interpreted as the feature-defined allophone conditioned on the context of the entire utterance (in principle). For our example utterance /t eh n k ae t s/ (*ten cats*), the state-transition graphs for all the phones-in-context can be easily constructed according to the procedure described in the previous subsection; the graphs for the first phones-in-context are shown in Figures 2a and 2b. The construction is based on the feature-overlap pattern obtained by assembling all possible feature alignments (with one specific instance of the feature alignment illustrated in Figure 1), and it utilizes the constraint that feature spreading does not extend beyond the immediately adjacent phones.

Let us focus on the state-transition graph for /eh/-in-context (Figure 2b). Each node encompasses a five-tuple feature composition, which is encircled by a box. All possible phasing relations of the features shown in the first three columns of Figure 1 uniquely form seven distinct nodes and their connections as depicted in Figure 2b. For instance, node A is formed by spreading the TB feature from left and simultaneously spreading the Larynx feature from the left into the /eh/ segment. A straightforward way of spreading TB feature from /t/ into /eh/ would result in a feature composition (0 1 9 1 2) for node A; but in Figure 2b, we denote node A instead as (0 L(1) 9 1 2), where the TB feature L(1) symbolizes *Left-Influence* of value "1" and is distinguished from the *intrinsic* value "1" associated with the /t/ segment. The need for this distinction arises from the consideration that spread of the TB feature into a vowel segment necessarily induces acoustic transitions, which should be modeled using a nonstationary process in the output distribution (rf. Deng, 1992) rather than using a conventional stationary process in the HMM state. Likewise, the TB feature R(1) in node F and node G symbolizes *Right-Influence* of value "1" which is distinguished both from the *intrinsic* value "1" and from L(1). (Some obvious constraints on parameters can be applied to L(1) and R(1), however; e.g. the slopes of the transitional trends can be constrained to have opposite signs.) In this study, we adopt the assumption that when a principal-articulator feature (Lips, TB, or TD) is spread into a vowel, they will turn their intrinsic value x into L(x) or R(x) [8]. Different ways of and different depths in the spread of various features from both left and right give rise to the remaining nodes in Figure 2b.

The example of Figure 2b also illustrates the generality of the state-transition graph in representing the phasing or timing relations among the various articulatory features. As an example, let us examine the relative phasing between the TB feature and the Larynx feature (voicing characteristics in the "larynx" function). If the relative phasing is such that the spread of the Larynx feature leads the spread of the TB feature, feature composition of (0 0 9 1 2) will be implemented and node C will be visited. The acoustic manifestation of this feature composition will be the aspiration (voiceless) microsegment of the /eh/ segment without appearance of acoustic transition. Since this microsegment has to follow the diminishing effect of the motion of the principal articulator (the tongue dorsum in this case, which is initiated or induced by releasing the vocal tract constriction formed by the tongue blade), the transition from node A to node C is represented as unidirectional. Conversely, if the spread of the Larynx feature lags that of the TB feature then the implemented features will be (0 L(1) 9 1 1) as represented in node B. The associated microsegment will be the voiced portion of the acoustic transition in /eh/. Again, this has to follow the microsegment for the voiceless portion of the acoustic transition and hence node B is forbidden from connecting back to node A. It is noted that for a single token the phasing between the two features must be either leading or lagging; hence, the transitions between node B and node C are not allowed and the state transition must go through only one path. But for the state-transition graph to be general enough to represent an ensemble of tokens, both possibilities for the phasing relations are allowed and hence both nodes B and C are present (in parallel) in the same state-transition graph.

The way in which the state-transition graph is constructed as illustrated in the above example ensures that all the nodes within each graph for each particular phone-in-context are distinct. However, for the same phone across a large number of different contexts, a large degree of sharing of acoustic data can be accomplished by identifying the same feature compositions across the state-transition graphs representing these phones-in-context [9]. To demonstrate this, we choose an example allophone /eh/ in context /p eh n/ (*pen*), which shares the Larynx feature with but differs in the principal-articulator features from the allophone /eh/ of Figure 2b (considering feature spread from left only). The state-transition graph of this new allophone constructed from its feature-overlap pattern is shown in Figure 3. Note that five out of seven nodes for the new allophone, C, D, E, F, and G, have the identical feature compositions as the corresponding nodes in the graph of Figure 2b. This is due to the identical context for all the feature components from right and for partial feature components from left (all except the Lips and TB features). In particular, the common Larynx feature in left is responsible for the common node C, which corresponds to the steady-state voiceless portions of the /eh/ segments after the influence of the principal articulator (induced by TB for Figure 2b and by Lips for Figure 3) diminishes. The differing Lips and TB features, on the other hand, produce two distinct nodes A and B in Figure 3, one for the voiceless (aspiration) portion of the /p/-to-/eh/ labial-like transition (node A) followed by the other for the ensuing voiced portion of the transition (node B).

A further example of sharing nodes of different allophones is given in Figure 4 for /eh/ in /d eh n/ (*den*). This allophone shares the principal-articulator (TB) feature with but differs in the Larynx feature from the allophone in Figure 2b. We note first that there is only one node (B) which results from the feature spreading from left, as opposed to the three nodes in Figure 2b or in Figure 3. This is not surprising as the Larynx feature agrees between the home segment /eh/ and the neighboring segment /d/ and hence there is virtually no feature spreading in this dimension. Further, all the nodes in Figure 4 can find their identities in Figure 2b. In particular, node B identifies itself in both allophones as the voiced portion of the alveolar-like transition of /eh/ and node D as the voiced and steady-state portion of /eh/.

To give an example demonstrating validity of the concept of sharing model states, arising from sharing articulatory features, for real speech data, we contrast the spectrograms (Figure 5a vs. Figure 5b) and the associated Mel-frequency cepstral coefficients (MFCCs) (Davis and Mermelstein, 1980) (Figure 6 for MFCCs up to the fourth order) for the diphone /t eh/ (in word *tenant*) versus the diphone /d eh/ (in word *academic*) from TIMIT data. The values of the shared features are (0 L(1) 9 1 1), corresponding to node B in Figure 2b for diphone /t eh/ and node B in Figure 4 for diphone /d eh/. The acoustic correlates of this shared feature bundle are: formant transition induced from alveolar place in Left (TB feature L(1)) and towards the target formant of /eh/ (TD feature 9), non-nasalization (Velum feature 1), and voicing (Larynx feature 1). The acoustic evidence supporting the commonality of these acoustic correlates examined from spectrograms can be found in the region right to the arrows in Figures 5a and 5b. The commonality of the acoustic correlates examined from the MFCCs (computed every 3 ms) is shown in the regions right to the vertical dashed lines in Figure 6 (the vertical lines marking the the time origin are aligned with the time points indicated by arrows in Figures 5a and 5b). Such commonality is revealed in contrast to the significantly greater differences in the MFCC values outside the temporal regions where the articulatory features and associated acoustic correlates share across the two diphones.

The sharing mechanism for the nodes in the state-transition graphs and consequently for acoustic data demonstrated above can be shown to be operative not only among the same phones over different phonetic contexts as in the example of Figures 5 and 6, but also across different phones within and even across different words. The sharing is automatically established as long as the feature compositions are constructed to be identical. The essence of this sharing mechanism, which naturally takes into account contextual dependence among phonetic segments, is that it enables decomposition of the generally complex influence of a contextual factor into its individual, phonologically meaningful feature constituents. Because the features used are so designed that they can be interpreted as actions of the articulatory structure, the decomposition of contextual factors allows us to tease apart the various causes of the acoustic variability at the level of speech production. As a result, the context dependent aspect of acoustic variability in speech can be modeled more parsimoniously than other earlier enumeration-based approaches to modeling context dependence (e.g. Deng et al., 1990; Lee and Hon, 1990; Schwartz et al., 1984) [10]

Before we conclude this most important subsection in this paper, we add a description on the convention used to handle diphthongs, which, unlike the usual quasi-phonemic segments described above, are characterized by the two associated target vowels. We use the example of diphthong /ow/ in context /A/./ow/./B/ to illustrate the process of constructing its state-transition graph. First, /ow/ is decomposed into /ao/ + /uh/. Then, the state-transition graphs for the two new phones-in-context — /A/./ao/./uh/ and /ao/./uh/./B/ — are constructed in the same way as for the monophthongs described above. Finally, concatenation of the two resulting graphs is used as the state-transition graph for /A/./ow/./B/.

In summary, our model for lexical representation consists of two parts. First, given the phonemic transcription of any arbitrary utterance and given the feature specification system, an abstract lexical representation is established in a matrix form. Second, this abstract representation is further compiled into a state-transition graph through the processes of probabilistically and fractionally overlapping the features and of constructing the state-transition graph from the obtained statistical pattern for the feature overlaps. We should emphasize that it is this latter compilation process which endows the model with the ability to parsimoniously share and organize the acoustic data in speech, which is highly desirable for large-vocabulary speech recognition applications. This ability is closely related to the observation that the fractional feature-overlap process, as a key component of our lexical-representation scheme, allows for a succinct description of apparently complex context dependent phenomena and phonological variations, which we will illustrate below using examples.

E. Examples of coarticulation and phonological variations accounted for by fractional feature overlaps [11]

Two practical criteria to evaluate a lexical-representation scheme are its generality and parsimony. As part of our design goal and part of our observations, a significant number of contextual (or co-articulatory) factors and phonological variations which have been identified and documented in the literature (e.g. Lehiste, 1967; Zue, 1991) can be concisely described by a single mechanism — fractional feature overlaps. The examples are provided in this subsection, which are arranged into various major categories of the phenomena.

First, homorganic stop insertion in nasal-fricative consonant clusters, such as insertion of /t/ after /n/ in *prince*, and insertion of /d/ after /n/ in *tens*. The insertion is implemented according to our feature-overlap model either by spreading the TB feature of /n/ (value "1" indicating alveolar gesture) into a fractional segment of /z/, or by spreading the Velum feature of /z/ (value "1" indicating raised velum) into a fractional segment of /n/. Readers are invited to draw the feature-overlap pattern for the phrase *tens* in the format of Figure 1, and convince themselves that it is an effortless exercise to find a time slice, either in the segment of /n/ or of /z/, which contains the feature composition (0 1 0 1 1) that is exactly the feature specification for /d/ listed in Table 1. We emphasize that the feature overlap described here is "fractional", because complete overlap of either the TB or the Velum feature would allow identification of neither feature composition of (0 1 0 2 1) /n/ nor of (0 3 0 1 1) /z/ in the feature-overlap pattern, which would amount to a substitution rather than an insertion process.

Second, "gesture" hiding and syllable reduction, such as hiding of /t/ in *must be* and in *perfect memory* and hiding of schwa in *beret*, all in casual speech. In careful speech, however, the hidings manifest themselves to a decreasing degree. Our fractional feature-overlap model accounts for this continuum of behavior simply by allowing for an increasing degree of temporal overlaps of features from non-overlap for slow speech to near-complete overlap for fast speech. A similar type of feature overlap is accountable for homorganic stop deletion such as deletion of /t/ in *mostly*.

Third, non-release of a stop consonant in a word-final position while a gesture in a principal articulator is formed. Near-complete feature overlaps readily account for the non-release part. But further to the hiding example, the anticipatory co-articulation manifested as formant transitions in the preceding vowel due to the formation of the gesture can be simultaneously accounted for by fractionally spreading the principal-articulator feature left into the vowel segment.

Fourth, vowel nasalization, with the example of /eh/ nasalization in Figure 1, is accounted for by fractionally spreading the Velum feature from a nasal-consonant segment into an adjacent vowel segment. (As in English anticipatory co-articulation tends to prevail over carry-over one, it would be reasonable to impose the constraint in our feature-overlap model that the average depth or fraction in the spread of the Velum feature (value "2") is greater to left than to right.)

Fifth, dental consonant nasalization, such as nasalization of /dh/ in *in the room*, is accounted for by spreading in right the Velum feature from /n/ into /dh/ [12].

Sixth, distinction between fronted velar stop vs. back velar stop, with examples in Figure 1, is accounted for by either fractional or complete spread of the TD feature of adjacent vowels into the home stop segment. (Left or anticipatory spread would have a stronger effect than right spread.)

Seventh, acoustic transitions in vowels induced by an adjacent consonant, either left or right, with several examples shown in Figure 1, are accounted for by fractional spread of one of the principal-articulator features from a consonant into an adjacent vowel.

Eighth, consonant lip rounding, such as the rounding in *strawberry* throughout the word-initial consonant cluster. This is accounted for by complete spread of the Lips feature (value "3") left from the /ao/ segment all the way into the /s/ segment.

Nineth, voicing assimilation, such as devoicing (possibly partially) of /z/ in word *his* in the phrase *his shoes*. This is accounted for by (fractional) spread of the Larynx feature left from /sh/ into /s/.

Tenth, assimilation of final alveolar consonants, such as assimilation of /t/ to /p/ in *note pad*, /s/ to /sh/ in *gas shortage* and in *miss you*, /t/ to /k/ in *late call*, and /n/ to /ng/ in *ten cats*, is accounted for by (fractional) spread of the TD or Lips feature from either a palatal/velar or a labial consonant left into the alveolar consonant *and* simultaneously by assuming that the TB value ("1") occupies only a fractional segment of the alveolar home segment (see in Figure 1 the example of fractional assimilation of /n/ to /ng/.) A similar type of the feature spread is accountable for assimilation of /d y/ to /dh/ in *did you*. For the assimilation of final alveolar, the constraint can be strictly imposed that the TD spread to right is forbidden.

IV. Implementation and Evaluation of the Feature-Based Recognizer

The model for the feature-based lexical representation and the design of the HMM speech recognizer using such a representation described up to now are intended for unlimited English vocabularies containing exhaustive contextual variations at the acoustic level. As a careful but feasible step in evaluating our model and the recognizer, a realistic but relatively simple task was chosen with the goal that from the performance result we will be able to further refine the model and improve the recognizer design. In this section, we present experimental results for the task of phonetic classification using TIMIT database (with vocabulary size on the order of 6,500), along with technical detail for implementing the recognizer evaluated.

A. Task and corpus

The experiments described in this section are aimed at classifying the 61 quasi-phonemic labels (called phones) defined in the TIMIT database. However, in keeping with the CMU/MIT standards for determining classification accuracy , we folded 22 phone labels into the remaining 39 ones [13]. After the folding, all the phones in the effectively 39 separate categories are: {uw ux}, {uh}, {ah ax ax-h}, {aa ao}, {ae}, {eh}, {ih ix}, {ey}, {iy}, {y}, {ay}, {ow}, {aw}, {oy}, {er axr}, {r}, {l el}, {w}, {m em}, {n en nx}, {ng eng}, {dx}, {v}, {th}, {dh}, {hh hv}, {z}, {s}, {sh zh}, {f}, {jh}, {ch}, {b}, {d}, {g}, {p}, {t}, {k}, and {cl dcl gcl kcl pcl tcl epi ps q pau}. The confusions within the same categories are not counted in calculating classification accuracy. Consistent with the finding documented in Lee and Hon (1989), in some limited experiments we found that many confusions were made within the 10-member closure category and so folding closure was necessary for good performance. But the effect of folding other categories was found to be rather minor.

As one of the goals in evaluating our recognizer's performance is to examine the sensitivity of phonetic classification accuracy to the training size, we initially used only 40 speakers' speech data in the training set, which was then gradually expanded to include 120 and 240 speakers. The test set consisted of 20 speakers (5774 test tokens in total) [14]. In these data sets, only "sx" and "si" sentences were used. So the test set consisted of 160 sentences, and the training sets consisted of 320, 960, and 1920 sentences, respectively, for the three separate sets of experiments to be described later in this section.

B. Signal processing

The raw speech data in TIMIT was in the form of waveforms with a sampling rate of 16 kHz. Although one major long-term goal of this research is to design the signal representation of speech based on the knowledge of the acoustic and auditory correlates of the articulatory features, we restrict the scope of this study to only a relatively simplistic signal representation — MFCC. In computing the MFCCs, a Hamming window of duration 25.6 ms was applied every 10 ms. Within each window, a vector consisting of the MFCCs from the first order to the seventh order was calculated. The seven-dimensional vector was then extended to a 15-dimensional one (frame n) by appending the differences between the MFCCs at frame $n+2$ and those at frame $n-2$, and by further appending the differential log power value.

C. Training and recognition algorithms

According to our feature-based lexical-representation model described in Section III, each phone-in-context is linked to a distinct state-transition graph (or a topological structure of a Markov chain) and the utterance is represented as a sequential concatenation of the graphs corresponding to these phones-in-context. As outlined in Section II, the feature-to-acoustics mapping in the model is accomplished statistically by associating each state, which represents a unique feature composition, with an "output" distribution function. (So the Markov chain becomes a hidden process and hence hidden Markov model.) Again, as a feasible step at present in implementing and evaluating our model, such a function was chosen in this study to be a rather simplistic one — stationary mixture Gaussian distribution.

The training of the recognizer concerns determination of the mixture weights, the Gaussian means, and of the covariance matrices for each state in the HMM. Again, for implementation simplicity, we assume that all the covariance matrices are diagonal. Further, all the transition probabilities of the Markov chain were fixed a priori and were not trained further[15].

The training process begins with model parameter initialization. In principle, the initialization should be accomplished straightforwardly by manually assigning microsegmental frames from a small portion of the training data (to be used for bootstrapping) to the relevant HMM states according to the acoustic interpretation of the state's feature components[16]. In practice, however, we used a somewhat more automatic approach to the initialization problem for all the experiments to be reported in this section.

The initialization procedure assumes that segmentation information at the quasi-phonemic level is available for the seed training data. The initialization proceeds as follows:

1. Read in a phone token (including its label, contexts and segment boundaries).

2. Search for the phone's state-transition graph according to its contexts, and determine the *depth* of the graph [17].

3. Partition the segment of the input phone token into D approximately equally spaced intervals. Assign all the frames within the i-th interval to all the states of depth i.

4. If the input token is the last phone in the seed training set go to 5; else go to 1.

5. For each distinct state in the ensemble of the state-transition graph for the lexicon, find the centroid, x_0, of all the frames assigned to the state (denoted by n, the total number of such frames). Then arrange the n data frames, $x_1, x_2, ..., x_n$, according to their distances [18]from x_0 such that: $d(x_1, x_0) \leq d(x_2, x_0) \leq \cdots \leq d(x_n, x_0)$. Finally, from this ordered list of frames, those from the rank $1 + (m-1)n/M$ to mn/M are used to initialize the mean and the variance parameters associated with the k-th mixture component according to their sample mean and sample variance, where $k = 1, \cdots, M$ with M being the number of Gaussian mixtures. Mixture weights are initialized uniformly as $1/M$.

After the HMM model parameters are initialized, the Baum-Welch algorithm (Baum, 1972) or the segmental k-means algorithm (Rabiner, 1989) can be used to improve the model using a given set of training data. Likewise, there are also two standard recognition algorithms available: the forward algorithm or the Viterbi algorithm (Rabiner, 1989). In our experimentation with the feature-based speech recognizer, we found that the two training algorithms had little effect on the recognizer's performance, which is consistent with many previously reported findings (e.g. Deng, 1991). However, somewhat unexpectedly we found that use of the forward algorithm in the feature-based speech recognizer produced consistently better performance than use of the Viterbi algorithm All the experimental results reported in this section were obtained using the forward algorithm in recognition.

D. Phonetic classification results

We have constructed four different versions of the feature-based HMM recognizer according to four different ways of constraining feature overlaps in constructing the feature-overlap patterns and the associated state-transition graphs. We have experimented with several primitive rules restricting the feature spreading. They are: 1) Feature spreading does not extend beyond the immediately adjacent phones; 2) Closures are not susceptible to feature spread from adjacent phones; 3) Closures and /h/ do not spread their features into adjacent phones; 4) Flaps are not susceptible to feature spread from adjacent phones; 5) The Larynx feature of a phone does not spread into adjacent phones; and 6) The principal-articulator features (Lips, TB, TD) for nasals & fricatives are not susceptible to feature spread from adjacent phones. Use of various combinations of these rules allows us to reduce the impractically large sized "compiled" TIMIT lexicon (over 20,000 HMM states) to the ones which become manageable in implementing the recognizer. Four ways of combining these rules produce the following four versions of the recognizer: Version I with a total of 3097 states; Version II with a total of 2556 states; Version III with a total of 1992 states; and Version IV with a total of 1143 states.

The results of speaker-independent phonetic classification experiments, with details of the task, recognizer implementation, and four versions of the feature-based HMM recognizer described above, are plotted in Figure 7. The classification performance is measured as the percent correct for top one choice. Shown as each data point in Figure 7 is the performance measure conditioned on three factors: 1) training size (40, 120, or 240 training speakers); 2) number of Gaussian mixtures (the M value, 1, 5, or 16) in the output distributions; and 3) version of the feature HMM (I, II, III, or IV). For comparison purposes, we also list the performance measures of the context-independent HMMs for the identical task and on the identical experimental conditions (labeled by "5"s). During the classification phase, prior phone probabilities were used to combine the acoustic likelihood scores with the weights of the prior set equal to the number of frames in each test token. For all the four versions of the feature-based recognizer, the left-phone and the right-phone contexts were provided explicitly as side information [19].

The results shown in Figure 7 allow us to draw several preliminary conclusions. First, the feature-based approach, which is intended to model phonetic context dependency using a unified feature-overlap mechanism and which allows sharing and organization of acoustic data in a parsimonious manner, is superior to the context-independent approach by significant margins with the error rate reduction ranging from 15% to 27%. This may not sound surprising considering many successful attempts to develop methods for context-dependent modeling in the past (e.g. Schwartz et al., 1984; Lee and Hon, 1990; Deng et al., 1990). However, these more and less standard methods, usually of the type based on enumerating contexts, all demanded as large amounts of training data as they are available. Perhaps of particular interest among the results shown in Figure 7 is that the superiority of the feature-based approach is extended to as low as 40 speakers in training data, which is about ten percent of the training data used in other context-dependent modeling methods such as those in (Lee and Hon, 1989) for a similar task[20]. Second, the performance of the feature-based recognizer is rather insensitive to the size of the training data (note similar accuracy between use of 120 speakers and of 240 speakers in training). This again should not be surprising despite the large degree of such sensitivity that has been reported in the literature (e.g. Deng et al., 1991; Deng et al., 1992) [21]. As the principal motivation of the feature-based recognizer's design is to extract common acoustic information in the entire acoustic space, a relatively small amount of training data may have already provided sufficient information for phonetic classification purposes. Third, different constraints on the feature overlaps, reflected in the four versions of the feature HMM shown in Figure 7, produced the performance measures varying to a significant degree. Such interesting dependence is, however, rather difficult to rationalize at present. It appears to be the result of complex interactions among such factors as quality of the feature-overlap rules, complexity of the resulting HMM topology, total size of the compiled lexicons, and the training size. Finally, a reasonably small number of mixtures (M=5) suffices for the peak performance produced by all versions of the feature-based recognizer. This again contrasts the conventional recognizers which require considerably higher number of mixtures [22]. Such a low mixture number is expected as our feature-based recognition scheme is designed to capture the context-dependent behaviors in a principled way and subsequently to reduce acoustic variabilities associated with the HMM states. We have analyzed the classification results in detail and found a common and highly regular confusion pattern; that is, the confusions are dominated by the phone pairs which differ by only one single feature within the feature-specification system of Tables 1 and 2.

Despite the above preliminary but promising results, we note that the most recent results reported in (Leung et al., 1993) on the same phonetic classification demonstrated 74% accuracy using conventional pattern classification approaches and 78% with connectionist approaches, outperforming our best result using features — 73% (Figure 7). These superior results were obtained using a much larger amount of training data than ours and, perhaps most importantly, using a comprehensive evaluation procedure to optimize nearly every single processing step to compute the speech signal representation. Although it is not clear whether application of a similar optimization procedure will improve the classification rate in our feature-based system to the same degree as reported in Leung et al. (1993), the results of Leung et al. (1993) suggest to us that there is a vast amount of discriminative information which is lost in our generic signal processor producing uniform MFCCs. How to capitalize on such information within our currently developed feature-based framework will be our future research direction.

V. Discussions and Future Directions

In summary, the main contributions of the work presented in this paper can be viewed from two angles. For one, we have developed a general statistical framework in which the highly variable speech data in the acoustic surface form can be structurally organized and shared across a potentially unlimited number of lexical items. This process of organizing and sharing speech data over the entire acoustic space of speech produces the statistically "compiled" lexicon so designed that it is amenable to statistical learning and hence becomes directly applicable to automatic speech recognition. In the "compiled" lexicon, the atomic speech units, or the minimal speech segment sharable within the entire acoustic space of speech, are constructed specifically at the consistent and uniform microsegment level — feature constellation or composition being assembled automatically from a phonological process called feature overlapping. We emphasize that our scheme of sharing acoustic data across the units of the feature-defined microsegments is more flexible and powerful than the conventional ways of data sharing designed at the considerably grosser level — phonetic segment. In this sense, the current work can be regarded as a principled way of extending our earlier studies on the speech units (Deng and Erler, 1992; Deng, Lennig, and Mermelstein, 1990). Specifically, the extension has been made from the data-sharing scheme applicable to only a small subset of English sounds to that applicable to all classes of English sounds.

¿From the second, more theoretical angle, our current study can also be viewed as generalization of the theory of the phonological features in that our feature-overlap model embeds the critical notions of "fractional overlap" and of "probabilistic overlap". The significance of such generalization lies in the creation of a natural framework in which an overwhelming majority of context-dependent speech phenomena at the acoustic level can be described in relatively tidy terms of a unifying feature-overlapping mechanism at the articulatory level. It is this nature of parsimony in modeling the context-dependent behaviors which provides the "compiled" lexicon with the many strongly desirable properties of speech-data sharing, among which is a relatively small number of atomic speech units in total for characterizing the entire acoustic space in speech. This goal of structural and parametric modeling of context-dependent behaviors (as opposed to the conventional modeling methodology of an enumerative type) has also been pursued in several other studies (e.g. Bakis, 1991; Kenny et al., 1991).

One distinguishing characteristic of the recognizer design described in this paper is its integrated nature in the utilizations of the hard-wired speech knowledge and of automatic learning procedures. In the design of the recognizer, the whole process of lexicon compilation, which determines the overall topology of the HMM for an arbitrary utterance, is hard-wired. The sub-processes of the compilation, including establishment of the feature-specification system, rules imposed on the constraints of feature overlaps, and the mapping from a feature-overlap pattern to the state-transition graph, are all designed in a deterministic manner, with inspirations from the theory of phonological features and from notions in articulatory phonology. On the other hand, the mapping from the symbolic feature-defined speech units to acoustic observations, as the surface-component of the recognizer, is made stochastic and is largely learned from (a small amount of) training data [23]. There is an interesting intermix between the lexical-representation component (hard-wired) and the acoustic-mapping component (stochastic) of the recognizer: The result of the learning designed to seek the "optimal" statistical distribution parameters also simultaneously determines the "optimal" depth or fraction of the overlap for each individual articulatory feature.

Another aspect of the integrated nature of our feature-based recognizer design is that the lexical-representation and acoustic-mapping components are combined into a unified Bayesian framework for lexical access and it is neither necessary nor desirable to explicitly extract the features as an intermediate representational form from the acoustic signal [24]. In one sense, within this integrated statistical framework, the lexical access is accomplished essentially by an analysis-by-synthesis strategy. That is, a lexical item which, via its representation by the feature-overlap pattern (or equivalently the feature-defined state-transition graph) and via the subsequent statistical sampling from the proper probability distributions on speech data, generates the acoustic data most closely matching the input speech data is decoded as the output word by the recognizer.

The current design and implementation of the feature-based speech recognizer as presented in this paper have been limited in several aspects. First, the rules we used for constraining the feature overlaps as described in Section IV.D are rather simplistic and contain several unrealistic aspects. The first step in improving the current speech recognition framework is to devise a systematic and comprehensive set of constraining rules at the same level of the articulatory structure as that denoted by the feature set. While limiting the total number of distinct statistical distributions characterizing the overall acoustic space of speech, the carefully designed constraining rules will be key to reducing overfitting the model to data and hence be essential for improving modeling accuracy and for relaxing the training data requirement. Second, as mentioned in Section II, the current recognizer design treats all the feature assignments as the symbolical labels only and does not explore the clear relationships among the assigned feature values. Our next step in furthering the current work will be to devise a systematic methodology to incorporate such relationships, possibly in conjunction with adoption of more desirable sets of phonetic labels than the TIMIT labeling system currently in use. This will provide additional powerful constraints on the degree of freedom in the parameters of the distribution functions [25]. Investigation into the quantal nature of speech, especially the possible quantal properties expressed in terms of the multi-dimensional articulatory structure generalized from the limited single-dimensional case as in the study of Stevens (1989), will provide the main knowledge source for this direction of the future work. This work will be based primarily on simulations on some specialized speech production models for exploring quantitative aspects of the non-monotonic relations between articulatory configurations and their acoustic and auditory consequences. Third, the current framework has yet to be further developed to equip itself with the capability of incorporating the articulator-free features and the hierarchical feature structure induced by them [26]. We envision that these key concepts and their deduced consequence of "acoustic landmarks", which play critical roles in human speech production and perception, can be naturally embedded into our current rather general statistical framework after a mathematically rigorous formalism for non-uniform signal sampling becomes established from our future research efforts. Finally, choice of the uniform and generic acoustic measurements (MFCCs) currently in use takes no advantages of the rich information and vast opportunity, which is offered by the lexical compilation process, that each state (feature constellation) in the HMM has its explicit interpretation in terms of the underlying multi-dimensional articulatory structure. Our future goal will be to develop a mathematically rigorous framework for state-conditioned speech representation and signal modeling, making proper use of detailed knowledge of the acoustic and auditory correlates of the articulatory features. A large amount of this knowledge has been accumulated over the past 40 years by speech and hearing scientists and more important ones will be discovered in the future. Such scientific knowledge can be expected to contribute directly to the technologically challenging field of automatic speech recognition once the statistical framework outlined above is established from our future work on the basis of the approach presented in this paper.

Acknowledgements

Discussions with Ken Stevens on several key aspects of this work, his encouragement, comments, and suggestions, as well as those of Victor Zue, Louis Braida, and Jonathan Allen are gratefully acknowledged. We also wish to thank Mike Phillips, Jim Glass and other members of the SLS group of MIT for suggestions and discussions on the evaluation tasks described in Section IV. Two anonymous reviewers provided constructive comments which improve the quality of the paper. A major portion of this study was carried out while the first author was on sabbatical leave from University of Waterloo and he thanks Victor Zue for the invitation and for the support. This work is supported by Natural Sciences and

Engineering Research Council of Canada and by DARPA grant (N00014-89-J-1332).

[1]We are grateful to Professor Ken Stevens who offered suggestions which improve this feature specification system.

[2]In contrast, the Lips and TD features for /t/ are unmarked as given by value "0"; the actual active Lips and TD features associated with a /t/ segment are determined solely by their contexts.

[3]However, glottal stop and stop flap are treated as separate phones unrelated to /t/ in this study. This choice follows the TIMIT convention for labeling and was made for convenience purposes in evaluating the recognizer's performance.

[4]Professor Louis Braida pointed out to us that psychoacoustic experiments can be designed to determine the actual feature-overlap depths in human speech perception and to examine the degree to which the perceptual measurement and the computational result agree with each other.

[5]Professor Ken Stevens pointed out this interesting interpretation to us.

[6]As this example shows, in the current feature-overlap model there is no need to distinguish the front from the back velar stop in the feature specification system (Table 1) since such context dependence is automatically captured by the feature overlapping mechanism.

[7]The highly parameterized models such as the task-dynamic model describing coordination of the speech articulators (Saltzman and Munhall, 1989) appear to be well suited for this application.

[8]However, due to implementation complexity, we limit this study to the conventional stationary Gaussian distributions, rather than the more desirable distributions such as those described in Deng (1992) for first-order nonstationarity and those described in Deng (1993) for second-order nonstationarity, in the HMM states; but the Gaussian distribution parameters were made distinct between the states containing the *intrinsic* value and the *induced* value.

[9]Recall that in an HMM the parameters in the output distribution functions are forced to be identical (tying) for the same states, no matter where the states are in the global topology of the HMM. Hence, tying the HMM states implies sharing the acoustic data.

[10]It is recognized that a full degree of parsimony offered by the feature-based framework has not been realized by the current implementation discussed in this section; for instance, the current implementation fails to achieve HMM parameter sharing when four out of five features are the same. Two ways of overcoming this limitation, numerical treatment of feature values and state-conditioned speech representation, will be discussed in Section V.

[11]Several examples in this subsection were drawn from various publications of Browman and Goldstein, Browman and Goldstein (1992) in particular; some others were generated from discussions with Ken Stevens and with Victor Zue.

[12]However, it appears that common voicing features between /n/ and /dh/ is a precondition for such nasal spread. For instance, the Velum feature does not spread from /n/ into /f/ in *in fact*.

[13]This folding scheme is identical to the one that the SUMMIT system is currently using for performance evaluation for the same classification task. We thank M. Phillips for providing the phone folding table to us.

[14]These training and test sets are a random subset of those used for performance evaluation in the current SUMMIT system.

[15]Therefore, the statistical variation in the depth of the feature spread from one phonetic segment to

another is characterized solely by the collection of the output distribution functions and by their topological connections. The most probable feature-overlap pattern for any particular utterance token can be determined by an optimal Viterbi-like search based on a sequence of spectral matches.

[16]Figure 2b serves a good example to illustrate this point. For instance, the aspirated frames within /eh/ which are in transition from /t/ should be assigned to state A. The frames which can be identified as nasalized and transitional should be assigned to state G and those non-nasalized be assigned to state F, etc.

[17]The depth of a state-transition graph D is defined as the depth of the terminating state in the graph, where the *depth* of a state in the graph is defined as the maximal number of states to be traversed through in order to reach the current state from the starting state. The construction of a state-transition graph ensures that all the states with the same depth are not connected to each other.

[18]We used the most simple distance measure, the Euclidean distance, for this purpose in our experiments.

[19]This requirement for explicit contextual labels should not be construed as limitation of the current framework for applications in (large vocabulary) continuous speech recognition since the current framework is designed such that recognizer should be most efficiently deployed after the N-best hypothesized sentences (with a negligible rejection rate) are obtained using a reasonably good fast-match algorithm. The low rejection rate can be flexibly traded with a large number of N and the computation required to re-score these large N candidates using the feature-based recognizer we have developed is fully parallel in nature.

[20]The task in Lee and Hon (1989) is phonetic recognition on the same database as ours. Taking into account the facts that the phonetic recognition is a more difficult task than the classification task but that insertion errors were not counted in the reported result, the calibrated performance reported in Lee and Hon (1989) for the context-independent HMMs is comparable to ours. It is interesting to note that the performance improvements via use of contexts are very similar between the two studies but in our study the improvement is achieved with use of a considerably smaller amount of training data.

[21]For the identical training and test data sets, we have run a version of the SUMMIT system, which also demonstrated a relatively large degree of the performance sensitivity to the training size. We thank M. Phillips for providing us with the SUMMIT softwares.

[22]For instance, the SUMMIT system uses M= 16, and the system reported in Deng et al. (1991) that is context-independent (hence more variability) but speaker-dependent (hence less variability) uses M=25 for the peak performance.

[23]As an informal parallel, many aspects of the processes of speech production and perception in human must be learned; but a few somewhat deeper properties of the processes cannot be attributed to learning and have to be genetically “hard-wired” (Halle and Stevens, 1991).

[24]Doing so, in fact, would result in “sub-optimal” lexical-access performance according to the Bayesian theory. Experimentally, we found that the forward algorithm in our feature-based recognizer consistently outperformed the Viterbi algorithm; the Viterbi score is derived from a single state sequence in the HMM and thus explicitly determines the exact decoded features for each speech frame.

[25]Extreme cases of such constraints in a rudimentary form have, in fact, been implemented in the framework described in this paper, which is implied in the feature-specification system summarized in Tables 1 and 2.

For instance, we assigned the identical symbolic value "1" of the Lips feature to all the labial consonants, identical symbolic value "1" of the TB feature to all the alveolar consonants, and identical value "2" in the Velum feature to all the nasal consonants, etc.

[26]We are grateful to Ken Stevens for pointing out this important direction to us.

References

[1] Bakis, R. (1991). "Coarticulation modeling with continuous-state HMMs," *Proc. 1991 IEEE Workshop on Automatic Speech Recognition,* Arden House, Harriman, New York, pp. 20-21.

[2] Baum, L. E. (1972). "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, Vol. 3, pp. 1–8.

[3] Browman, C. and Goldstein, L. (1992). "Articulatory phonology: An overview," *Phonetica,* Vol.49, pp. 155-180.

[4] Browman, C. and Goldstein, L. (1990). "Gestural specification using dynamically-defined articulatory structures," *J. Phonetics,* Vol.18, pp. 299-320.

[5] Browman, C. and Goldstein, L. (1986). "Towards an articulatory phonology," *Phonology Yearbook,* Vol.3, pp. 219-252.

[6] Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*, New York: Harper and Row.

[7] Clements, G. N. (1985). "The geometry of phonological features," *Phonology Yearbook 2*, pp. 225-252.

[8] Davis, S. and Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 4, pp.357–365.

[9] Deng, L. (1993). "A stochastic model of speech incorporating hierarchical nonstationarity," *IEEE Transactions on Speech and Audio Processing*, Vol. 1, No. 4, pp. 471–474.

[10] Deng, L. (1992). "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Processing*, Vol.27, No.1, pp. 65-78.

[11] Deng, L. (1991). "The semi-relaxed algorithm for parameter estimation of hidden Markov models," *Computer Speech and Language*, Vol. 5, No.3, pp. 231–236.

[12] Deng, L. and Erler, K. (1992). "Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech units," *J. Acoust. Soc. Am.*, Vol.92, No.6, December, 1992, pp.3058–3067.

[13] Deng, L., Kenny, P. Lennig, M., Gupta, V., Seitz, F., and Mermelstein, P. (1991). "Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. 39, No. 7, pp. 1677–1681.

[14] Deng, L., Kenny, P., Lennig, M., and Mermelstein, P. (1992). "Modeling acoustic transitions in speech by state-interpolation hidden Markov models," *IEEE Transactions on Signal Processing*, Vol.40, No.2, pp.265-272.

[15] Deng, L., Lennig, M., and Mermelstein, P. (1989). "Use of vowel duration information in a large vocabulary word recognizer," *J. Acoust. Soc. Am.*, Vol. 86, pp. 540-548.

[16] Deng, L., Lennig, M., and Mermelstein, P. (1990). "Modeling microsegments of stop consonants in a hidden Markov model based word recognizer," *J. Acoust. Soc. Am.*, Vol. 87, pp. 2738-2747.

[17] Deng, L., Lennig, M., Seitz, F., and Mermelstein, P. (1990). "Large vocabulary word recognition using context-dependent allophonic hidden Markov models," *Computer Speech and Language*, Vol. 4, pp. 345-357.

[18] Erler, K. and Deng, L. (1993). "Hidden Markov model representation of quantized articulatory features for speech recognition," *Computer Speech and Language*, Vol. 7, No. 3, pp. 101-118.

[19] Fant, G. (1960). *Acoustic Theory of Speech Production*, Mouton, The Hague.

[20] Halle, M. and Stevens, K. (1991). "Knowledge of language and the sounds of speech," In J. Sundberg, L. Nord, and R. Carlson (eds.) *Music, Language, Speech, and Brain,* London: MacMillan, pp. 1-19.

[21] Hwang, M., Huang, X., and Alleva, F. (1993). "Predicting unseen triphones with senones," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pp.311-315.

[22] Kenny, P., Zhao, R., Gupta, V., Lennig, M., and O'Shaughnessy, D. (1991). "Articulatory Markov models," *Proc. 1991 IEEE Workshop on Automatic Speech Recognition,* Arden House, Harriman, New York, pp. 22-23.

[23] Lee, K. and Hon, H. (1990). "Context-dependent phonetic hidden Markov models for continuous speech recognition," *IEEE Transactions on Signal Processing*, Vol.38, No.4, pp.599-609.

[24] Lee, K. and Hon, H. (1989). "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Signal Processing*, Vol.37, No.11, pp.1641–1648.

[25] Lehiste, I. (1967). *Readings in Acoustic Phonetics*, MIT Press, Cambridge, MA.

[26] Leung, H., Chigier, B., and Glass, J. (1993). "A comparative study of signal representations and classification techniques for speech recognition," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 680-683.

[27] McCarthy, J. (1988). "Feature geometry and dependency: a review," *Phonetica*, Vol. 43, pp. 84–108.

[28] Mermelstein, P. (1973). "Articulatory model for the study of speech production," *J. Acoust. Soc. Am.*, Vol. 53, pp. 1070–1082.

[29] Phillips, M., Glass, J., and, Zue, V. (1992). "Automatic discovery of acoustic measurements for phonetic classification," *Proc. Int'l Conf. Spoken Language Processing*, Vol. 1, pp. 795-798.

[30] Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. of the IEEE*, Vol. 77, pp. 257–285.

[31] Saltzman, E. (1986). "Task dynamic coordination of the speech articulators: A preliminary model," In *Generation and modulation of action patterns,* (H. Heuer & C. Fromm, editors), pp. 129-144, New York: Springer-Verlag.

[32] Saltzman, E. and Munhall, K. (1989). "A dynamical approach to gestural patterning in speech production," *Ecological Psychology*, Vol. 1, No. 4, pp. 333-382.

[33] Schwartz, R., Chow, Y., Roucos, S., Krasner,, M., and Makhoul, J. (1984). "Improved hidden Markov modeling of phonemes for continuous speech recognition," *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, pp. 35.6.1–35.6.4.

[34] Stevens, K. (1986). "Models of phonetic recognition II: An approach to feature-based recognition," In P. Mermelstein (ed.) *Symposium on Units and Their Representation in Speech Recognition,* 12th Int'l Congress on Acoustics, Montreal, pp. 67–69.

[35] Stevens, K. (1989). "On the quantal nature of speech," *J. Phonetics,* Vol.17, pp. 3–45.

[36] Stevens, K, Manuel, S., Shattuck-Hufnagel, S., and Liu, S. (1992). "Implementation of a model for lexical access based on features," *Proc. Intĺ Conf. Spoken Language Processing*, Vol. 1, pp. 499-502.

[37] Zue, V. (1985). "The use of speech knowledge in automatic speech recognition," *Proc. of the IEEE*, Vol. 73, pp. 1602–1550.

[38] Zue, V. (1991). "Notes on Speech Spectrogram Reading", course notes, M.I.T., Cambridge, MA.

[39] Zue, V., Glass, J., Goodine, D., Phillips, M., and Seneff, S. (1990). "The SUMMIT speech recognition system: Phonological modeling and lexical access," *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. 49-52.

[40] Zue, V. and Sia, E. (1982). "Nasal articulation in homorganic clusters in American English," *Working Papers in MIT Speech Communications Group*, Vol. 1, pp. 9–17.

Table and Figure Captions

Table 1: Multi-valued feature specification for consonantal segments (TIMIT labels).

Table 2: Multi-valued feature specification for non-consonantal segments (TIMIT labels).

Figure 1: Schematic illustration of a result of the feature-overlapping process for an example phrase /t eh n k ae t s/ (*ten cats*). **S1, S2, S3,** and **S4** are four sampling points in the utterance which possess explicit interpretations in terms of the overlapping articulatory features.

Figure 2: State-transition graphs for the first two phones-in-context (a)-(b) in the example phrase illustrated in Figure 1.

Figure 3: State-transition graph of the allophone /eh/ in context /p eh n/ (*pen*).

Figure 4: State-transition graph of the allophone /eh/ in context /d eh n/ (*den*).

Figure 5: Spectrograms of a TIMIT utterance *tenant* containing diphone /t eh/ (a) and of a TIMIT utterance *academic* containing diphone /d eh/ (b). Some shared articulatory features across these two utterances are responsible for common acoustic properties located to the right of the regions indicated by the arrows (see text for detail).

Figure 6: Similarity of acoustic properties, within a local temporal region to the right of vertical dashed lines, across the two utterances in Figure 5 expressed in terms of the MFCC parameters (see text for detail).

Figure 7: Phonetic classification accuracy for 20 test speaker data in TIMIT measured as percent correct for top one choice for four versions of Feature HMM recognizer and for context-independent phonetic HMM recognizer.

type	symbol	example	Lips	TB	TD	Velum	Larynx
Closure:	bcl,dcl,gcl		0	0	0	1	1
	pcl,tcl,kcl		0	0	0	1	2
Stops:	b	bee	1	0	0	1	1
	d	day	0	1	0	1	1
	g	geese	0	0	1	1	1
	p	pea	1	0	0	1	2
	t	tea	0	1	0	1	2
	k	key	0	0	1	1	2
	dx	muddy,dirty	0	2	0	1	1
	q	at, bat	0	0	0	1	3
Fricatives:	s	sea	0	3	0	1	2
	sh	she	0	3	2	1	2
	z	zone	0	3	0	1	1
	zh	azure	0	3	2	1	1
	f	fin	2	0	0	1	2
	th	thin	0	4	0	1	2
	v	van	2	0	0	1	1
	dh	then	0	4	0	1	1
Nasals:	m	mom	1	0	0	2	1
	n	noon	0	1	0	2	1
	ng	sing	0	0	1	2	1
	em	bottom	1	0	17	2	1
	en	button	0	1	17	2	1
	eng	Washington	0	0	20	2	1
	nx	winner	0	2	0	2	1
Affricates:	jh	joke	dcl + d + zh				
	ch	choke	tcl + t + sh				

Table 1:

type	symbol	example	Lips	TB	TD	Velum	Larynx
Semivowels:	l	lay	0	5	3	1	1
	r	ray	5	6	4	1	1
	w	way	3	0	5	1	1
	y	yacht	0	7	6	1	1
Vowels:	iy	beet	0	0	7	1	1
	ih	bit	0	0	8	1	1
	eh	bet	0	0	9	1	1
	ey	bait	0	0	9/7	1	1
	ae	bat	0	0	10	1	1
	aa	bott	0	0	11	1	1
	aw	bout	0/3	0	11/15	1	1
	ay	bite	0	0	11/7	1	1
	ah	but	0	0	12	1	1
	ao	bought	4	0	13	1	1
	oy	boy	4/0	0	13/7	1	1
	ow	boat	4/3	0	13/15	1	1
	uh	book	3	0	14	1	1
	uw	boot	3	0	15	1	1
	ux	toot	0/3	0	8/15	1	1
	el	bottle	0	5	17	1	1
	er	bird	5	6	16	1	1
	ax	about	0	0	17	1	1
	ix	debit	0	0	18	1	1
	axr	butter	0	0	19	1	1
	ax-h	suspect	0	0	19	1	2
Aspirations:	hh	hay	0	0	0	1	2
	hv	ahead	0	0	0	1	1

Table 2: