Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition

Aren Jansen and Partha Niyogi[a)]
Department of Computer Science, University of Chicago, 1100 East 58th Street, Chicago, Illinois 60637

(Received 13 September 2007; revised 10 June 2008; accepted 11 June 2008)

This paper elaborates on a computational model for speech recognition that is inspired by several interrelated strands of research in phonology, acoustic phonetics, speech perception, and neuroscience. The goals are twofold: (i) to explore frameworks for recognition that may provide a viable alternative to the current hidden Markov model (HMM) based speech recognition systems and (ii) to provide a computational platform that will facilitate engaging, quantifying, and testing various theories in the scientific traditions in phonetics, psychology, and neuroscience. This motivation leads to an approach that constructs a hierarchically structured point process representation based on distinctive feature landmark detectors and probabilistically integrates the firing patterns of these detectors to decode a phonological sequence. The accuracy of a broad class recognizer based on this framework is competitive with equivalent HMM-based systems. Various avenues for future development of the presented methodology are outlined.
 [DOI: 10.1121/1.2956472]

PACS number(s): 43.72.Ne [DOS] Pages: 1739–1758

I. INTRODUCTION

We are interested in the problem of *pure speech recognition*—the ability of humans to interpret (decode) the speech wave form reaching their ears in terms of a sequence of phonological units without invoking any higher level (syntactic, semantic, or pragmatic) linguistic knowledge. This ability to perform pure acoustic–phonetic decoding is manifest in our ability to recognize streams of nonsense words, to detect familiar words in unfamiliar streams,[1] and in a host of other such basic phonological perception processes. Furthermore, this ability must underpin all higher level language learning; after all, every time we hear a new word, it is effectively a nonsense word for us. We are able to perform such acoustic–phonetic decoding in the absence of higher level linguistic knowledge, but we do need knowledge of phonetic, phonotactic, and phonological regularities of the language at hand. Correspondingly, our method rests heavily on accurate models of such regularities.

Our approach is based on the following principles:

(1) In our system, *distinctive features* are the atomic units in terms of which all higher level phonological units such as broad phonological classes, phonemes, syllables, and the like are composed. This is in contrast to most traditional speech recognition systems that use phone-based units (e.g., triphones). The distinctive features are symbolic (discrete), define natural classes, have articulatory and acoustic correlates, are hierarchically structured, and may overlap in time as suggested by autosegmental phonology.[2–4] They also provide a compact representation for phonological rules to capture coarticulatory effects that result in pronunciation variability.
(2) Distinctive features are associated with articulatory gestures that have natural acoustic consequences in the speech signal. The acoustic properties that distinguish different sounds from each other and different distinctive features from each other reside at different scales in time and frequency. Consequently, rather than having a "one size fits all" representation as is common in much of modern speech recognition, we will have multiple representations tuned for different distinctions.
(3) Correspondingly, we build *feature detectors* for the patterns and characteristic signatures for the distinctive features. Of particular significance in our system is the points in time when feature detectors fire. These points in time are associated with important events or *landmarks*, which correspond to maxima, minima, and inflection points of specialized acoustic properties. As a result, one obtains a *sparse, point process* representation of the speech signal. Such a representation is reminiscent of spike train patterns observed in the behavior of selective neurons, particularly in parts of the auditory cortex of a variety of animals (see Refs. 5 and 6). Such a spike-based representation is in contrast to the vector time series (i.e., frame-based) representation used in nearly all modern recognition systems.
(4) The decoding of the signal proceeds by integrating the firings of the individual feature detectors in a hierarchical way, where high-level decisions trigger off further context-dependent processing at lower levels. At the root of the hierarchy is the sonorant-obstruent feature, which is the most basic and perceptually salient distinction among speech sounds. Vowels correspond to peaks of the sonority profile and provide anchor points that define syllable-sized analysis units. Integration of detector outputs occurs at such *syllabic* time scales on the rationale that this is the smallest perceptually robust unit. The information content of the signal within each analysis unit is coded in the temporal statistics of the point process

[a)]Electronic mail: niyogi@cs.uchicago.edu

representation. This is in the spirit of *temporal coding* in neural systems and allows us to model durational patterns in a novel way.

Our goal in this paper is to give computational expression to these principles. In recognition of the inherent variability in the acoustic correlates of distinctive features, we pay particular attention to what might be natural and coherent statistical frameworks in which to model different aspects of the signal. In the process, we end up with a system that is quite unlike any other built so far, though it is clearly closest in spirit to those inspired by acoustic phonetics, distinctive features, and landmarks (e.g., Refs. 7–10; see Sec. II E for discussion).

One may ask, what are the benefits of an approach such as ours? First, our motivations may be traced more directly to scientific understanding of related phenomena in linguistics, psychology, and neuroscience. Consequently, our system may provide a computational platform to test specific hypotheses about speech perception in these fields. Second, the simplicity of our modular design may aid diagnostics and portability to new languages and environments. Third, the hierarchical approach leads to fewer parameters, allows reuse of training data for different distinctions, and makes possible the efficient training of our system from limited amounts of training data. Finally, the system design, with its specialized detectors and temporal coding, provides a new way to characterize the statistics of speech signals and reason about issues of invariance and robustness.

As an intermediate step toward our long-term goal, we have chosen in this paper to concentrate on the task of broad class recognition. This is the simplest nontrivial sequence recognition task, requiring not only the capacity to distinguish between phonological classes (i.e., classification), but also manage possible insertion and deletion errors. For this reason, broad class recognition has often been an intermediate step in evaluating new approaches (see Refs. 8 and 11). It is worth noting that broad class recognition is not without practical merit in its own right. As one example, Huttenlocher and Zue (see Ref. 12) proposed an approach to lexical access based on partial phonetic information (i.e., broad class information only). A method of incorporating a broad class recognizer in noisy environments to improve robustness has also been proposed (see Ref. 13). Finally, certain small vocabulary tasks can be performed with a broad class recognizer. One straightforward example is spoken digit recognition, as each digit has a unique broad class sequence (see Ref. 14). Another possibility would be key-word spotting, where partial phonetic information could be used for a low specificity search.

In Sec. II, we outline the overall architecture of the system. In Secs. III and IV, we elaborate on the computational details and experimental performance of each of the modules. It is worth noting that the approach of this paper is only a particular instantiation of the general principles outlined earlier and we discuss variants at each stage. Consequently, in Sec. V we discuss future directions and in particular the challenges to work up to a complete recognition system.

II. OVERALL SYSTEM ARCHITECTURE

Here we elaborate on the principles outlined in Sec. I, motivate them, describe how they are instantiated in our overall architecture, and make connections to related research at appropriate points.

A. Distinctive feature representation

Our goal is to build a machine that takes as input the speech signal and produces as an output a symbolic (linguistic) representation. It is overwhelmingly the case that modern speech recognition systems based on hidden Markov models (HMMs) use some kind of phone-based representation (typically triphones in large scale applications) with an HMM mediating the mapping between phone sequences and an acoustic vector time series.

In contrast, we will represent the phonological units in terms of distinctive features. Since this may be a confusing term to the speech engineer (in the terminology of the engineer or statistician, a feature is typically a property of the speech signal), it is worthwhile to elaborate a little bit on the phonological notion of distinctive feature, its justifications, and its implications for our approach.

Although antecedents of the feature may be found in the sivasutras of Panini, the modern notion is usually traced to Trubetzkoy, Jakobson, and others in the early part of the 20th century. While we recognize intuitively that the objects of language may be hierarchically composed of smaller objects (at a rough cut, sentences are made up of phrases, phrases of words, words of syllables, and syllables of phonemes), it was the insight of Jakobson and others that phonemes were not the primitive, atomic units or building blocks of language but rather each phoneme may be usefully considered to be a complex of distinctive features. These features typically emphasized binary oppositions among groups of phonemes that share some common characteristic. Thus a collection $f_1, \ldots, f_k$ of binary-valued features define a possible set of at most 2^k different phonemes in a language.[15]

Distinctive features are motivated by several different considerations. First, distinctive features provide one way to group the phonemes into *natural classes*. The primary argument comes from linguistics, where it is a curious fact of language that some groups of phonemes seem to behave as an equivalence class as they participate in phonological processes. For example, in English, to pluralize a noun, one takes the root word (stem) and modifies it as follows:

If the word ends in /s z tʃ dʒ ʃ ʒ/ add /ɪz/,
  Examples: places, porches, cabbages, ambushes;
else if the word ends in /p t k f θ/ add /s/,
  Examples: lips, lists, maniacs, telegraphs;
else add /z/,
  Examples: clubs, herds, fangs, holes, pies
(after Ref. 16).

Thus, the sounds /p t k f θ/ behave as a class with respect to this rule. When one examines the class of sounds /p t k f θ/ and asks what they share in common that distinguishes them from the subclass /b d g v ð/, one arrives at the understanding that the former consists of unvoiced sounds while the latter consists of their voiced counterparts. Indeed,

TABLE X. Confusion matrix for $S/N=+12$ db and frequency response of 200–1200 cps.

	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	165	46	31	3	1				1				1			
t	91	83	68	4	1	2			1			2				
k	48	55	147	2	3							1				
f	16	4	3	146	60	3	2	11			1	2				
θ	4	3		109	76	17	2	12	1			2	1	1		
s	2	1	1	43	83	83	11	3		1	1	7				
ʃ	1	6	2	12	41	86	90		6	4		4				
b				14	5			223	4		5	1				
d	1				1	3	4	4	173	37		2	1	2		
g	1					1			102	107	1	2	7	7		
v	2	2		2	1			23	1	2	163	62	14	3	1	
ð				1		3	2	27	6	32	87	107	36	7		
z	1							4	12	48	10	15	114	39		1
ʒ							1		3	35	1	16	60	134	2	
m	1									1					229	9
n															5	247

FIG. 1. Consonant confusion matrix from Ref. 19.

voicing is the feature distinguishing the minimal pairs (/b p/, /t d/, /k g/, /f v/, /θ ð/) with phonemes /p t k f θ/ being [−voiced] and phonemes /b d g v ð/ being [+voiced]. Indeed, the English plural rule may be summarized by the following: If the word ends in [+coronal +strident], add /ɪz/; if it ends in [−voiced +stiff vocal cords] add /s/; add /z/ otherwise. A large number of similar examples in English may be found in Ref. 17, while subsequent works have elaborated and refined the analysis for many different languages of the world (see Ref. 18 for some details).

A second argument comes from studies of phonetic perception in psychology. For example, the classic work of Miller and Nicely[19] tested various subjects on their ability to discriminate between consonantal phonemes at different noise levels. Confusion matrices were constructed and shown in Fig. 1 is an example of one such confusion matrix at 12 dB signal-to-noise ratio. Notice the block diagonal structure of the confusion matrix with a natural grouping of phonemes into perceptually similar classes. These coincide with natural classes organized by distinctive features.

A third argument comes from articulatory and acoustic correlates of the phoneme classes. Following principles of speech production, phonemes produced with similar articulatory gestures will have similar acoustic properties. Since all linguistically relevant sounds (phonemes) are produced by the manipulation of a small number of articulators (the tongue, lips, glottis, nasal coupling, etc.), one may group phonemes into classes based on similarity of articulatory configurations. Thus, the distinction between the groups of vowels /i ɪ u/ and /a ɑ æ/ can be based on feature of tongue height (high versus low); the distinction between /p b m/ and /t d n/ can be based on place of constriction (labial versus coronal) in the vocal tract when producing the consonants. Reference 20 is an elaboration of this aspect of speech analysis.

The last few decades have seen a convergence of these kinds of arguments into different kinds of feature systems that differ from each other in detail based on their application purpose but have the same natural coherence. Thus, feature systems are rooted in phonology but have natural articulatory interpretations and corresponding acoustic and perceptual correlates.

More recently, work by Goldsmith,[2] Sagey,[3] McCarthy,[4] and others suggests that features have an internal organization with a hierarchical relationship with respect to each other. For our purposes, we will adopt the hierarchy of Fig. 2 leading to the major classes of vowels, approximants, nasals, stops, fricatives, and affricates. This hierarchy involves the following distinctive features:

(1) [son]: distinguishes sonorant sounds ([+son]) that are made with an open vocal tract (such as vowels, approximants, and nasals) from obstruent sounds ([−son]) such as stops and fricatives.
(2) [cons]: distinguishes consonantal sounds from vowel sounds.
(3) [cont]: distinguishes stop consonants [−cont] from everything else. [−cont] sounds are produced with a complete closure of the airway at some point during the articulation.
(4) [nasal]: distinguishes those sounds that couple the nasal cavity /n m ŋ/ from the rest.

These features are closely related to manner features,[21] stricture features,[18] and articulator-free features.[20]

What is the significance of the hierarchy in our context? The hierarchy has a justification in our minds from several perspectives.

(1) Nodes higher up in the tree correspond to features that are somehow more basic or fundamental and whose acoustic correlates are less context dependent. As a result, these features may be derived from the speech signal in a more robust and reliable way.
(2) Some features are irrelevant based on values of certain other features. For example, if a sound is [−son], then

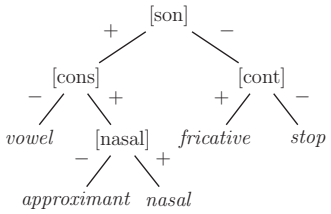

FIG. 2. The hierarchy of distinctive features leading to the broad (manner) classes.

the sound is automatically [−nasal]. Similarly, if the sound is [+son], then the sound is automatically [+cont]. Further features come into play once the broad class distinctions are made.

(3) The grouping of the features indicate a further restriction in the set of natural classes that are relevant for phonological processes.

As a result of the above-presented observations, we see that a featural notational system allows a more compact description of phonological processes such as assimilation and coarticulatory dynamics. Thus, pronunciation networks expressed featurally are likely to be more streamlined than those that are expressed in phonemic units.

So far, we have discussed how distinctive features allow one to describe the set of phonemes without regard to the fact that linguistic utterances are composed of words that are realized as linear sequences of segmental units or phonemes. A traditional view within generative phonology regards each phoneme to be a bundle of distinctive features and therefore each sequence is a sequence of feature vectors (bundles). This suggests an idealized view of a phonological utterance as a sequence of segments with the distinctive feature values being synchronized in time. However, we will adopt a representation that is more in line with the developments of autosegmental phonology, where associated with each node of the hierarchy we have a timing tier showing how the particular feature is distributed in time. Since the distinctive feature is a binary-valued, phonological variable, it will need to be estimated from the acoustic properties of the speech signal. Associated with these features are natural articulatory gestures with their corresponding acoustic consequences. These acoustic consequences will define points in time that are naturally related to the realization of the feature in time.

B. The sonority profile and vowels

The most obvious and perceptually salient aspect of the speech signal is the sonority profile of the signal distinguishing the sonorant sounds from the obstruent; recent psychological studies support this claim.[22] At the same time, perhaps the most basic distinction between classes of phonemes in their role in phonological processes is that between the consonants and the vowels. These two are related through the notion of a sonority hierarchy[23] with vowels occupying the top of the sonority hierarchy as the most open, full throated sonorant class of sounds. As a result, vowels correspond to local maxima of the sonority profile, occupy the nucleus position of syllables, and are the stress-bearing elements in the intonational contours of the speech stream.

Consequently, our entry point into the interpretation of the signal for further analysis is to

(1) Segment[24] the signal into sonorant and obstruent regions.
(2) Identify the vowel landmarks as the peaks of the sonority profile.

It is worthwhile to make a few remarks about the sonorant–obstruent distinction and its differences from the consonant–vowel distinction. Intuitively, sonorant sounds are produced with a relatively open vocal tract, with no significant pressure buildup anywhere in the tract, and with the vocal cords vibrating. Consequently, such sounds are always voiced; more significantly, from our point of view, because these sounds are associated with vibrating vocal cords, they are always periodic. The lack of significant pressure buildup and the relatively open vocal tract results in these sounds having marked formant structure in the lower frequency bands and relatively high energy. Thus sonorant sounds are easy to distinguish acoustically from obstruent sounds: This is perhaps the easiest distinction to make and as a result, we place this feature at the top of our hierarchy.

While all obstruent sounds are consonantal, not all sonorant sounds are vocalic. The approximants and nasals are consonantal sounds that are sonorant. Acoustically they are very similar to vowels, yet, phonologically, they play the role of consonants. The distinction between consonants and vowels is tied intimately with the syllable structure of a phonological sequence. The number of syllables in an utterance is equal to the number of vowels in that utterance with each vowel occupying the nucleus of a syllable.[25] Following the sonority hierarchy, and noting that vowels are at the top of the hierarchy, one finds that syllabic nuclei coincide with the local maxima of the sonority profile of the utterance.

Since we begin by identifying the sonorant and obstruent regions and the vowel landmarks,[26] the task of modeling the utterance reduces to modeling the intervocalic segments. The intervocalic segments correspond to continuous sequences of consonants that lie between two vowels. These are made up of sequences of sonorant consonants and obstruent consonants that are modeled separately. By doing this, we model portions of the signal at a syllabic scale. This potentially allows us to capture coarticulatory effects. Following the arguments of Ref. 27, there seems to be some justification in this choice from perceptual considerations as well.

C. Feature detectors for subsequent processing

The distinctive features corresponding to the lower tiers of Fig. 2 make further distinctions. Associated with each such feature, we build a *feature detector* that ideally should detect the presence or absence of that feature. One has to come to terms with the basic fact of speech that while the distinctive feature is categorical, the acoustic correlates are continuous and gradient. Consequently, the output of the feature detector is a real-valued function of time whose magnitude at each time may sometimes be interpreted as the degree (probability) to which the distinctive feature is present at that time.

There are three key ideas involved in the construction of a feature detector. First, for each feature of interest, a specialized acoustic representation is constructed in which that feature best expresses itself, i.e., a representation that captures the acoustic correlates of the feature that help distinguish positive from negative instantiations. Second, in this representational space, a classifier is built to separate positive examples from negative ones.[28] Third, the real-valued output of the classifier is further processed to provide a sparse rep-

resentation in time as a *point process*. Thus, maxima, minima, and inflection points of the output of the classifier correspond to points in time when "interesting" phonological events take place. Following the general philosophy of Stevens[7] and others, we refer to these points in time as *landmarks*.

Let us describe briefly four detectors that may be associated with leaf nodes of the distinctive feature tree of Fig. 2. These effectively serve as broad (manner) class detectors at each point in time.

(1) **Stop detector:** This is a detector for the plosive sounds /p t k b d g/. These sounds are [−cont] and are produced when there is a complete closure of the vocal tract followed by a sudden release. This leads to the acoustic consequence of having a period of extremely low energy followed by a noisy, wideband burst spectrum. Thus one is looking for a transition from closure to burst in the signal. Details surrounding the construction of a stop detector are provided in Refs. 29 and 30. We implement a variant of such a detector in our system.
(2) **Fricative detector:** This is a detector for the class of fricated sounds /s z ʃ ʒ f v θ ð/. These sounds are produced with a partial closure of the vocal tract so that there is turbulent pressure generated, leading to a noisy signal whose broad spectral profile is governed by the shape of the vocal tract. One may try to make a detector that fires during the fricated regions of the signal, defining landmark locations that correspond to local maxima of the continuous classifier output.
(3) **Nasal detector:** This is a detector for the nasal sounds /n m ŋ/ that are produced with a coupling of the nasal cavity. This nasal coupling has some characteristic acoustic consequences: The total energy is reduced substantially from the neighboring vowel, there is a drop-off in energy above 500 Hz, and the first formant is around 300 Hz. Here, a detector may be constructed to locate landmarks when nasal coupling is maximal.
(4) **Approximant detector:** This detector for liquids /l r/ and glides /h j w/ is the most difficult to implement because their acoustic characteristics are not completely distinct from adjacent vowels. Formant transitions and energy profiles in appropriately chosen bands may provide some discriminatory power.

Construction of each of the above-presented detectors constitute research projects in their own right (see Refs. 31–33 for examples). Our stop detector is constructed with an acoustic representation motivated directly by speech production. We construct the other detectors by computing canonical short-time (windowed) mel frequency cepstral coefficients (MFCCs) with window lengths, frame rates, and frequency ranges that are appropriate for each of the broad classes in question. Specifics regarding detector construction are provided in Secs. III B and IV B. Some further remarks are worthwhile:

(1) Sonorant regions are fundamentally different from obstruent regions of the signal. In sonorant regions, the signal is periodic and formant structure is evident. Most of the information content is at low frequencies and measures of spectral gravity, formants, and the like make sense. Consequently, the construction of nasal and approximant detectors and all further context-dependent processing may need to be sensitive to such considerations. We leave such refinements to future work.
(2) We do not pay attention to the output of the feature detector at all points in time. This leads to a sparse representation that may have some computational advantages. The specific points in time may be interpreted as the most relevant from perceptual or articulatory perspectives. These points may be related to the notion of landmarks in the theory of Stevens.[7] The duration between these points in time is naturally correlated with the timing of articulatory and perceptual events and may be modeled directly in our framework. Thus, rather than model the detailed value of the feature detector output at all points in time, we shift the burden to the modeling of the durational statistics.
(3) Our detectors may also be viewed as complex feature detectors that may themselves be trained on the output of more basic detectors along the lines of Ref. 34. It is also worthwhile to reflect on various neurophysiological findings that suggest the presence of neurons that fire selectively when certain complex acoustic attributes are present in the input stimulus. Our detectors may be analogized to such selective neurons.[5] A further connection to neurobiologically motivated speech recognition models may be found in Ref. 35, where spike pattern recognition models were also suggested.

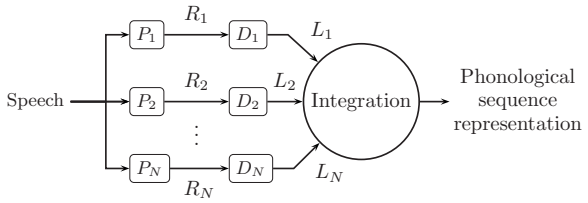

FIG. 3. Architecture of our landmark-based system. Here, $\{P_i\}$ are signal processors that output representation $\{R_i\}$, which are input into a set of feature detectors, $\{D_i\}$. The detectors output a set of candidate landmarks, $\{L_i\}$, which are probabilistically integrated to predict a phonological sequence.

At this point, the architecture of the system looks like that shown in Fig. 3. The speech signal is processed by several different signal processing transformations $(P_1,\ldots,P_N)$ to give rise to multiple representations $(R_1,\ldots,R_N)$. In each representational space, a feature classifier acts producing a real-valued output (first stage of $D_1,\ldots,D_N$). The last step is picking maxima, minima, or inflection points in the output of the classifier (second stage of $D_1,\ldots,D_N$) giving rise to points in time where the feature is most acoustically or perceptually prominent. These points in time may be naturally associated with events or landmarks $(L_1,\ldots,L_N)$, around which further context-dependent processing may be conducted.

Because the distinctive features have an internal hierarchical structure, we end up with a representation of speech as a collection of marked point processes associated with the

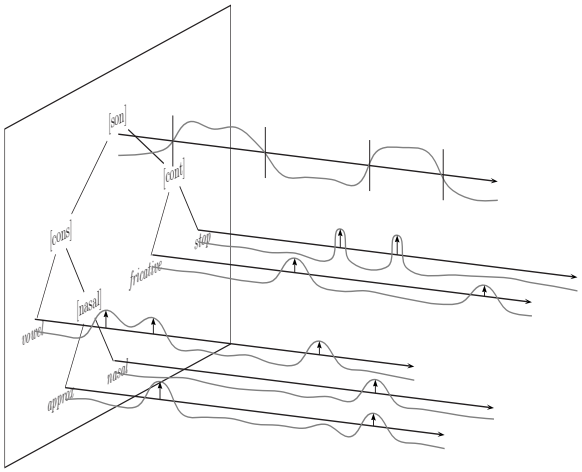

FIG. 4. Schematic diagram of the hierarchical timing tier representation. Landmarks are indicated by vertical arrows.

hierarchically structured timing tiers as shown in Fig. 4. A few aspects of this representation are worth noting. First, we have access not just to the points in time when the detector output peaks but also to the strengths of the detectors at that point, resulting in a natural marked point process. Second, the representation is very sparse. This is in stark contrast to typical representations of speech where one models the local spectral content at every time instant leading to a vector time series. Third, the information in the representation is now coded in the temporal dynamics of these spikes and it is natural to expect that the statistics of interspike times will be correlated with the durations between articulatory events and ultimately the durations of various linguistic segments. Finally, because of potentially different processing at different tiers, the time scales at which the different feature detectors fire may be quite different. This representation thus reflects our intuition that events in speech occur at multiple time scales that need to be decoupled from each other.

D. Integration in time

The challenge now is to integrate the firings of all these detectors in a coherent way to decode the sequence of broad classes that underlie the speech stream. In other words, we want a consistent way to map the hierarchical point process representation into a linear sequence of phonological units that are most likely to represent the underlying message in the speaker's mind.

The key idea is that we first segment the utterance into sonorant and obstruent regions and identify all the vowels in the utterance. The vowels are the nuclei of the syllables in the utterance and now each intervowel region is mapped into its most likely broad class sequence. If our feature detectors worked perfectly, this task would be trivial: Simply read off the output of the feature detectors to obtain the corresponding broad class sequence. On the other hand, we have to deal with the fact that our detectors have nonzero error rates. Thus, if between two vowel landmarks, the underlying sequence was /r k s/, we expect the stop detector fires exactly once at an appropriate point in time in between the two vowel landmarks and in an obstruent region of the signal.

TABLE I. Definition of broad (manner) classes used in our system, in order of descending sonority.

Broad Class	Abbreviation	Sonorant	Phones (Arpabet)
Vowels	V	Yes	iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, ux, er, ax, ix, axr, ax-h
Approximants	A	Yes	1, r, w, y, hh, hv, el
Nasals	N	Yes	m, n, ng, em, en, eng, nx
Fricatives	F	No	s, sh, z, zh, f, th, v, dh, jh, ch
Stops (Plosives)	P	No	b, d, g, p, t, k, dx, q
Silence	sil	No	bcl, dcl, gel, pci, tcl, kcl, pau, epi, h#

However, the stop detector may fire twice (for example, once correctly at the closure–burst transition and once incorrectly at some other point).

Therefore, we need to model the statistical distribution of the pattern of firings associated with each underlying sequence and choose the most likely sequence given the observed pattern of firings. Let

$$O = \{O^{\text{seg}}, O^{\text{v}}, O^{\text{sc}}, O^{\text{obs}}\}$$

be the set of observables associated with the timing tier representation, where

(1) $O^{\text{seg}} = \{0, t_1^{\text{seg}}, t_2^{\text{seg}}, \ldots, T\}$ contains the sonority segmentation boundaries,
(2) $O^{\text{v}} = \{(t_i^{\text{v}}, f_i^{\text{v}})\}$ contains the vowel landmarks,
(3) $O^{\text{sc}} = \{(t_i^{\text{sc}}, f_i^{\text{sc}})\}$ contains the sonorant consonant (approximant and nasal) landmarks, and
(4) $O^{\text{obs}} = \{(t_i^{\text{obs}}, f_i^{\text{obs}})\}$ contains the obstruent (fricative, stop, and silence) landmarks.

Here, each landmark is described by an ordered pair of the form (t, f), where t is the instance in time it occurred and f is the strength of the detection.

Given the timing tier observation variables O, our goal is to choose the most likely broad class sequence, $B \in \Sigma^*$ for $\Sigma = \{\text{V}, \text{N}, \text{A}, \text{F}, \text{P}, \text{sil}\}$ (see Table I for definitions of symbols), according to the maximum *a posteriori* (MAP) rule,

$$\begin{aligned} B_{\text{opt}} &= \arg \max_{B \in \Sigma^*} P(B|O) \\ &= \arg \max_{B \in \Sigma^*} P(B|O^{\text{seg}}, O^{\text{v}}, O^{\text{obs}}, O^{\text{sc}}). \end{aligned} \tag{1}$$

Applying Bayes' rule, and noting that $P(O^{\text{obs}}, O^{\text{sc}})$ is a constant in the optimization, we can write

$$B_{\text{opt}} = \arg \max_{B \in \Sigma^*} P(B|O^{\text{seg}}, O^{\text{v}}) P(O^{\text{obs}}, O^{\text{sc}}|B, O^{\text{seg}}, O^{\text{v}}).$$

According to the feature hierarchy, we model the terms $P(B|O^{\text{seg}}, O^{\text{v}})$ and $P(O^{\text{obs}}, O^{\text{sc}}|B, O^{\text{seg}}, O^{\text{v}})$ as follows (see Fig. 5).

(1) *Definition of sonorant and obstruent segments*: The sonority transition times O^{seg} define a set of interleaved obstruent and sonorant segments, denoted as

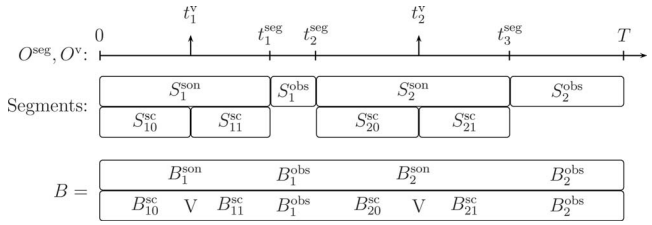

FIG. 5. Schematic representation of the segments and broad class subsequences used in the model derivation. Here we assume a sonorant segment comes first.

$$\{S_1^{\text{obs}}, \ldots, S_{N^{\text{obs}}}^{\text{obs}}\} \text{ and } \{S_1^{\text{son}}, \ldots, S_{N^{\text{son}}}^{\text{son}}\},$$

respectively, where each S_i^{obs} and S_j^{son} is a time interval. Without loss of generality, assume S_1^{son} occurs before S_1^{obs} and $N^{\text{son}}=N^{\text{obs}}$. Given a candidate sequence B, we declare it to be consistent with O^{seg} if there exist mappings $S_i^{\text{obs}} \rightarrow B_i^{\text{obs}} \in \{\text{F},\text{P},\text{sil}\}^*$ and $S_j^{\text{son}} \rightarrow B_j^{\text{son}} \in \{\text{V},\text{A},\text{N}\}^*$ such that

$$B = B_1^{\text{obs}} B_1^{\text{son}} B_2^{\text{obs}} B_2^{\text{son}} \cdots B_{N^{\text{obs}}}^{\text{obs}} B_{N^{\text{son}}}^{\text{son}}.$$

(2) *Definition of sonorant intervocalic segments*: For each sonorant region, S_i^{son}, the contained vowel landmarks,

$$O_i^{\text{v}} = \{(t,f) \in O^{\text{v}} | t \in S_i^{\text{son}}\},$$

further partition the interval into a set of sonorant intervocalic regions. That is, if $N_i^{\text{v}} \equiv |O_i^{\text{v}}|$, then each interval S_i^{son} can be partitioned about the vowel landmarks it contains into $N_i^{\text{v}}+1$ sonorant intervocalic segments, denoted $\{S_{i0}^{\text{sc}}, \ldots, S_{iN_i^{\text{v}}}^{\text{sc}}\}$, where each S_{ij}^{sc} is an interval in time. A candidate sequence B is consistent with O^{v} if, for every $i \in \{1, \ldots, N^{\text{son}}\}$, the number of vowel ($V$) tokens in B_i^{son} is equal to N_i^{v}. Thus, for B consistent with O^{v}, we may write

$$B_i^{\text{son}} = B_{i0}^{\text{sc}} V B_{i1}^{\text{sc}} V \cdots V B_{iN_i^{\text{v}}}^{\text{sc}},$$

where the $B_{ij}^{\text{sc}} \in \{\text{A},\text{N}\}^*$ is the jth vowel-separated sonorant consonant sequence contained in B_i^{son}.

(3) *Model for* $P(B|O^{\text{seg}},O^{\text{v}})$: In general, the term $P(B|O^{\text{seg}},O^{\text{v}})$ can accommodate a probabilistic segmentation strategy. This amounts to considering candidate sequences B that may not be consistent with O^{seg} and O^{v} in the manner defined earlier. However, in this paper we consider only a hard segmentation. Thus, given the above-presented definitions and assuming independence of sonorant and obstruent regions according to the distinctive feature hierarchy, we may write

$$P(B|O^{\text{seg}},O^{\text{v}}) = \prod_{i=1}^{N^{\text{obs}}} P(B_i^{\text{obs}}) \prod_{j=1}^{N^{\text{son}}} P(B_j^{\text{son}})$$

if B is consistent with O^{seg} and equal to 0 otherwise. Furthermore, assuming independence of sonorant intervocalic regions, we may write

$$P(B_j^{\text{son}}) = \prod_{k=0}^{N_j^{\text{v}}} P(B_{jk}^{\text{sc}})$$

if B is consistent with O^{v} and equal to 0 otherwise. Putting it together, we have

$$P(B|O^{\text{seg}},O^{\text{v}}) = \prod_{i=1}^{N^{\text{obs}}} P(B_i^{\text{obs}}) \prod_{j=1}^{N^{\text{son}}} \prod_{k=0}^{N_j^{\text{v}}} P(B_{jk}^{\text{sc}}), \tag{2}$$

where we need only consider candidate sequences B that are consistent with the segmentation defined by O^{seg} and O^{v}.

(4) *Model for* $P(O^{\text{obs}},O^{\text{sc}}|B,O^{\text{seg}},O^{\text{v}})$: Since the broad class detectors at the leaf nodes of the hierarchy are dominated by the sonorant–obstruent distinction at the root node, we assume conditional independence of the form

$$P(O^{\text{sc}},O^{\text{obs}}|O^{\text{v}},O^{\text{seg}},B) = P(O^{\text{sc}}|O^{\text{v}},O^{\text{seg}},B) P(O^{\text{obs}}|O^{\text{v}},O^{\text{seg}},B).$$

Note that we only need to evaluate $P(O^{\text{sc}},O^{\text{obs}}|O^{\text{v}},O^{\text{seg}},B)$ for those B where $P(B|O^{\text{seg}},O^{\text{v}})$ is nonzero. Detector firings that lie in different segments are assumed to be independent as well. Therefore,

$$P(O^{\text{obs}}|O^{\text{v}},O^{\text{seg}},B) = \prod_{i=1}^{N^{\text{obs}}} P(O_i^{\text{obs}}|B_i^{\text{obs}})$$

and

$$P(O^{\text{sc}}|O^{\text{v}},O^{\text{seg}},B) = \prod_{i=1}^{N^{\text{son}}} P(O_i^{\text{sc}}|O^{\text{v}},B_i^{\text{son}}),$$

where $O_i^{\text{obs}} = \{(t,f) \in O^{\text{obs}} | t \in S_i^{\text{obs}}\}$ and $O_i^{\text{sc}} = \{(t,f) \in O^{\text{sc}} | t \in S_i^{\text{son}}\}$. Likewise, assuming independence between the firings in separate sonorant intervocalic regions, we may write

$$P(O_i^{\text{sc}}|O^{\text{v}},B_i^{\text{son}}) = \prod_{j=0}^{N_i^{\text{v}}} P(O_{ij}^{\text{sc}}|B_{ij}^{\text{sc}}),$$

where $O_{ij}^{\text{sc}} = \{(t,f) \in O_i^{\text{sc}} | t \in S_{ij}^{\text{sc}}\}$. Note that the preceding probability factorizations assume the pattern of detector firings in a given segment are generated solely by the corresponding subsequence of B. Collecting terms, we have

$$P(O^{\text{sc}},O^{\text{obs}}|O^{\text{v}},O^{\text{seg}},B) = \prod_{i=1}^{N^{\text{obs}}} P(O_i^{\text{obs}}|B_i^{\text{obs}}) \prod_{j=1}^{N^{\text{son}}} \prod_{k=0}^{N_j^{\text{v}}} P(O_{jk}^{\text{sc}}|B_{jk}^{\text{sc}}). \tag{3}$$

Given the form of Eqs. (2) and (3), we can recast the optimization problem of Eq. (1) as

$$B_{\text{opt}} = \arg\max_{B\in\Sigma^*} \prod_i P(O_i^{\text{obs}}|B_i^{\text{obs}})P(B_i^{\text{obs}}) \times \prod_{jk} P(O_{jk}^{\text{sc}}|B_{jk}^{\text{obs}})P(B_{jk}^{\text{sc}}) = \arg\max_{B\in\Sigma^*} \prod_{i=1}^{N^{\text{obs}}} P(B_i^{\text{obs}}|O_i^{\text{obs}}) \prod_{j=1}^{N^{\text{son}}} \prod_{k=0}^{N_j^{\text{v}}} P(B_{jk}^{\text{sc}}|O_{jk}^{\text{sc}}), \quad (4)$$

where we consider only B consistent with and O^{seg} and O^{v} as described earlier. Therefore, the global MAP optimization problem reduces to a set of segment-level optimizations which are performed independently. Furthermore, we assume that some fixed set of obstruent and sonorant intervocalic prior distributions generate all possible observations. We will present separate MAP models for the terms $P(B_i^{\text{obs}}|O_i^{\text{obs}})$ and $P(B_{jk}^{\text{sc}}|O_{jk}^{\text{sc}})$ in Sec. III C.

E. Connections to previous work

Many of the ideas going into our system may be traced to the work of Stevens and colleagues.[7] There are, however, significant challenges in translating the philosophy of distinctive features and acoustic landmarks into a viable computational strategy. In particular, coping with the demands of the immense variability in the speech signal makes it essential to find a statistical framework in which those ideas can be embedded.

The distinctive feature aspect of our approach is most closely related to the event-based system (EBS) developed by Juneja and Espy-Wilson,[36] which arrives at a broad class segmentation using a Viterbi-style decoding of frame-level distinctive feature probabilities. (This approach is also used as a component in the systems developed in Ref. 9.) Landmarks for further processing are derived from the transition points of this broad class segmentation. Therefore, the representation for each feature must run on a common frame rate to allow for the frame-level comparison, effectively making it a frame-based dynamic model.

Our approach deviates from EBS in two fundamental ways: (i) We immediately divide the utterance into a series of syllable-sized analysis units using a sonority profile and (ii) we immediately adopt a sparse point process representation composed of landmarks in time, which are probabilistically integrated to arrive at a broad class sequence. Since our broad class decoding procedure is not performed on the frame level, we have the freedom to vary the frame rate of the individual feature representations. Furthermore, the phonological dynamics are modeled entirely on the point process representation, completely distinguishing our approach from frame-based methods.

Still, others have deviated from frame-based approaches and created probabilistic landmark models. In particular, various versions of the SUMMIT system[10] model acoustic observations made on a phonetic segment level, as well as at landmarks coinciding with segment transitions. However, our approach distinguishes itself from SUMMIT in two important ways. First, since we employ a distinctive feature hierarchy, we reduce our recognition problem into independent syllable-sized chunks, minimizing the complexity of the model and exploiting the robust context dependence the distinctive feature hierarchy provides. Second, within syllable-sized chunks, we explicitly model the relative temporal dynamics of the landmarks, which in this syllable-centric setting are highly constrained.

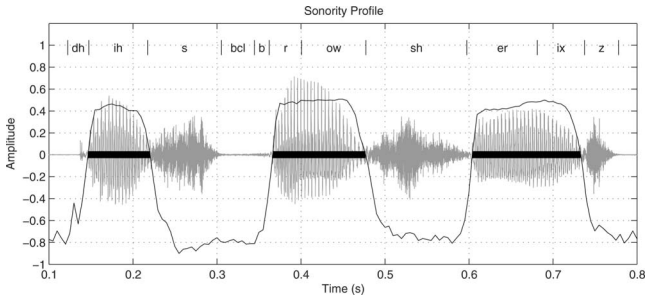

FIG. 6. A speech signal for the phrase "this brochure is" with the computed sonority profile overlaid. Sonorant segments are identified by horizontal bars.

III. TIMING TIER CONSTRUCTION AND INTEGRATION

Our timing tier representation requires the construction of a sonority segmenter and six broad class landmark detectors. The resulting landmark set must then be probabilistically integrated to perform a phonological decoding. In the following, we outline the theoretical and algorithmic details involved with our computational approach to these problems. Experimental details and results will be covered separately in Sec. IV.

A. Sonority segmentation

Computationally, a sonority segmentation may be accomplished using any available machine learning method. In our implementation, we employ support vector machines (SVMs). This popular machine learning technique involves solving the optimization problem

$$f^* = \min_{f\in\mathcal{H}_K} C\sum_{i=1}^{l} \max(0, 1 - y_i f(x_i)) + \|f\|_K^2,$$

for a decision surface f^* restricted to a representing kernel Hilbert space (RKHS) $\mathcal{H}_K$ for some kernel function K. Here, $\{(x_i, y_i)\}_{i=1}^l$ are the labeled training data and $\|f\|_K$ indicates the RKHS norm. Using the method of Lagrange multipliers, this convex optimization problem can be solved using a quadratic programming solver.

The SVM hinge-loss weight parameter C must be chosen appropriately regardless of the kernel used. Furthermore, we employ the radial basis function (RBF) kernel, $K(x,y) = \exp(-\gamma\|x-y\|^2)$, which introduces a second parameter, γ, the Gaussian argument coefficient. These parameters are chosen using holdout validation with the training set data. We employ a 39-dimensional mel frequency cepstral coefficient feature set.[37] However, any reasonable feature set may be used here, including possibly more robust acoustic parameters.[31–33,38,39]

Once the SVM is trained, to determine the segmentation we simply threshold the sonority SVM output, as shown in Fig. 6. In the subsequent stages of the system, *sonorant* re-

gions will refer to the segments above this threshold, while *obstruent* regions will refer the segments that fall below. Ideally, sonorant regions will contain sequences of vowels, approximants, and nasals; obstruent regions will contain sequences of silences, stops, and fricatives. In practice, the sonority segmentation is imperfect, and incorrect broad class content can be present in a particular region.

B. Broad class landmark detectors

The construction of detectors involves first training appropriate SVM classifiers for each broad class in Table I and then converting their real output into sparse sets of landmarks.[40]

1. Constructing the classifiers

If our sonority segmenter performed without error, there would be no presence of sonorant phonemes in predicted obstruent regions, and vice versa. In this ideal setting, the nasal classifier, for example, would only need to be constructed to differentiate between nasals and other sonorant classes. However, with an imperfect sonority segmentation, the classifiers need to be proficient at discriminating against all other possible phonemic content. For this reason, each one-versus-all classifier is trained with examples across the entire phonetic space, not just those in its sonority class. However, when integrating detectors, we only consider the firings of broad class detectors that are consistent with the initial segmentation. Fortunately, the robustness of the sonorant–obstruent distinction results in small differences between the performance of SVMs trained within their sonority class versus those trained across all phones. In the language of SVMs, the vast majority of the support vectors discriminate within the given sonority class. Our broad class SVMs also employ the RBF kernel, so we must again determine optimal C and γ parameters for each classifier via hold-out validation.

Since each classifier processes the signal independently, their construction can be specialized according to the individual broad class content. While we choose 39-dimensional MFCC features for the silence, fricative, nasal, approximant, and vowel SVMs, the framing parameters and frequency ranges used for each vary. Furthermore, our stop classifier employs energy and Wiener entropy parameters shown to be successful in this setting.[29] The modularity of this independent detector approach provides maximal flexibility for future development of our framework.

2. From classifier to detector

The output of each SVM is a real number for each frame of the signal. In general, after thresholding this series, we define the landmark time as the position of any local maximum of the SVM output and the landmark strength as the corresponding maximal values. The one exception made to this landmark picking strategy is for the vowel detector. It is common for the output of the broad class classifiers to experience multiple local maxima within a single phone as a result of acoustic variation arising from coarticulation. This does not pose a problem when landmarks are subsequently processed by our probabilistic integration scheme, as resulting degenerate (multiple correct within a single phone) detections are accommodated (see Sec. III C). However, in the positioning of vowel landmarks, degenerate detections will necessarily result in vowel insertions.

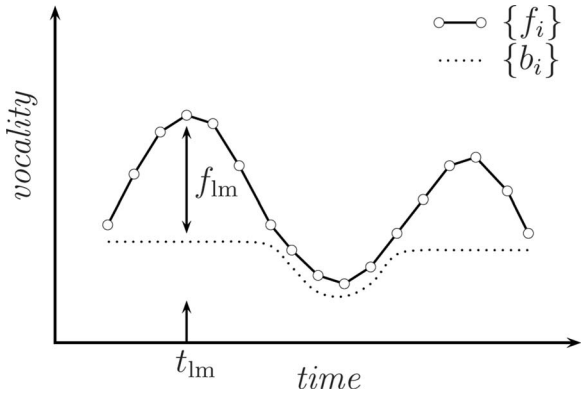

FIG. 7. A schematic plot of a series $\{f_i\}$ with its corresponding dynamic baseline $\{b_i\}$ for the given iteration. The corresponding landmark time t_{lm} and strength f_{lm} for this iteration are also shown.

To address this complication, we require a more sophisticated approach to choosing vowel landmarks given the continuous output of the vowel classifier. For this, we employ an adaptation of the "convex-hull" approach presented in Ref. 39 to recursively compute a time-dependent baseline. The input of the algorithm is the vowel classifier output time series, $\{f_i\}_{i=1}^{N}$, and proceeds as follows (see Fig. 7):

(1) Let $k=\arg\min_i\{f_i\}$ be the index of the absolute minimum of $\{f_i\}$. We can then compute the dynamic baseline, $\{b_i\}$,

$$b_i=\begin{cases}\min\{b_{i-1},f_i\} & \text{for } i=2,\ldots,k-1\\ \min\{b_{i+1},f_i\} & \text{for } i=k+1,\ldots,N-1\\ f_i & \text{for } \quad i=1,k,N.\end{cases}$$

This baseline is monotonically decreasing to the absolute minimum and monotonically increasing afterward.

(2) Create a new series $d_i=f_i-b_i$, equal to the difference between the original series and the dynamic baseline.

(3) Define a new landmark with $t_{\mathrm{lm}}=\arg\max_i\{d_i\}$ and $f_{\mathrm{lm}}=\max_i\{d_i\}$. Split $\{d_i\}$ about t_{lm} into two series, $\{l_i\}$ and $\{r_i\}$.

(4) Repeat all steps for both $\{f_i\}=\{l_i\}$ and $\{f_i\}=\{r_i\}$ while preserving absolute time positions, until a recursion depth is reached that overgenerates the number of candidate landmarks. This candidate set is then pruned by thresholding on the values of f_{lm}. This threshold is chosen empirically on a validation set.

Using this scheme, the strength of the landmark is determined relative to the baseline computed for the given iteration. While the amplitude of the local maxima may be large, nearby candidates compete with respect to a baseline computed in the local region. Therefore, small variations of the detector output that would otherwise result in degenerate landmarks are rejected with an appropriate choice of threshold on the difference series, $\{d_i\}$.

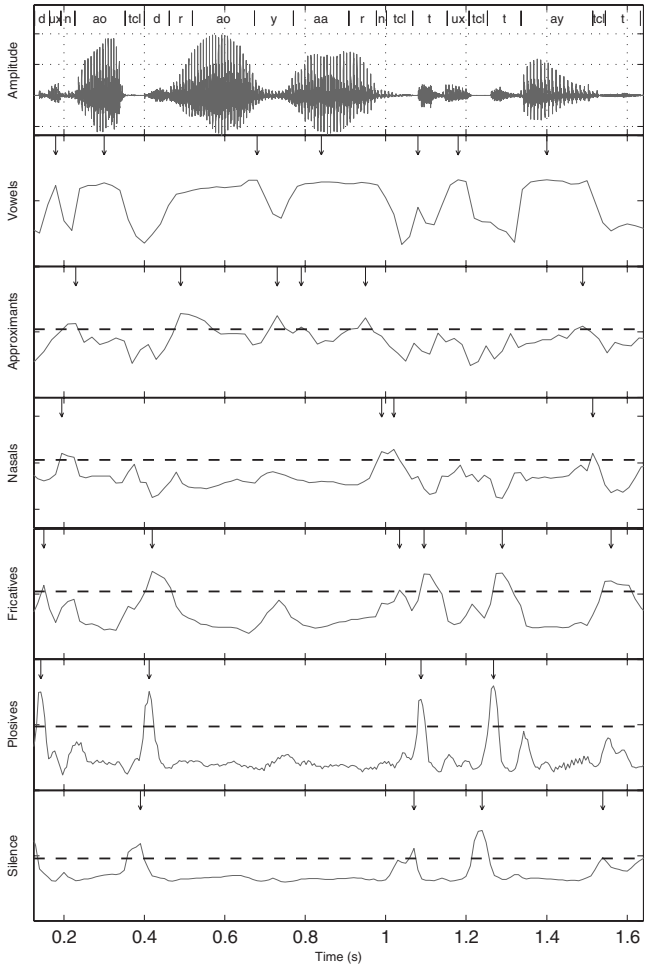

FIG. 8. The signal, broad class detector outputs and thresholds (dashed lines), and resulting landmarks (arrows) for the sentence "Do not draw yarn too tight."

Figure 8 shows the speech signal for the sentence "Do not draw yarn too tight," along with the output of the six classifiers, the thresholds used,[41] and the corresponding landmark positions. In Fig. 8, we see the vowel classifier reaches a local maximum at about 1.5 s. However, since it is lower than the adjacent local maximum at 1.4 s, the dynamic baseline algorithm prevents the insertion of a degenerate landmark within the /ay/ phoneme.

C. Probabilistic landmark integration

We now return to the task of probabilistically integrating the landmarks within each obstruent and sonorant intervocalic segment. This is akin to modeling $P(B_i^{\text{obs}}|O_i^{\text{obs}})$ and $P(B_{jk}^{\text{sc}}|O_{jk}^{\text{sc}})$ of Eq. (4) (see Sec. II D). To simplify the following discussion, we reset the above-used global notation, where (B,O) now refers to either $(B_i^{\text{obs}},O_i^{\text{obs}})$ or $(B_{jk}^{\text{sc}},O_{jk}^{\text{sc}})$, depending on the segment type we are discussing.

Now, if each broad class detector operated flawlessly, we could simply chronologically sort the detections in each sonorant or obstruent region, resulting in a perfect transcription of the content. However, in the presence of false positives and negatives, modest individual detector mistakes can combine to drastically reduce performance. For example, if each detector operated at an admirable 10% precision-recall equal error rate, this naive method of decoding would have a combined 50% insertion rate. Clearly, a more sophisticated integration strategy is required to clean up misfires.

To address the complication of spurious and missed landmarks, we have developed a framework for probabilistic segment decoding (PSD) based on a MAP estimate of the broad class landmark sequence in a given obstruent or sonorant intervocalic region. To see how this works, consider an interval of the speech signal (T_1,T_2) of duration $T=T_2-T_1$ that when combined with the activity of N broad class detectors, defines a set of observables

$$O=\{T,O_{X_1},\ldots,O_{X_N}\},$$

where each O_{X_i} denotes the observables for the class X_i detector. These consist of L_{X_i} time–strength paris (one per detection) which we denote as

$$O_{X_i}=\{(t_1^{X_i},f_1^{X_i}),\ldots,(t_{L_{X_i}}^{X_i},f_{L_{X_i}}^{X_i})\},$$

where we have converted the absolute landmark times to the fraction of the segment that passes before the landmark occurs. That is, if t is an absolute landmark time, the corresponding observable is $t^{X_i}=(t-T_1)/T$.

At this point we can immediately write down a simple MAP estimate of the segment broad class sequence, $B_{\text{opt}}=\max_B P(B|O)$. However, in the context of our hierarchical landmark-based system, we would like our model to also estimate which landmarks within the region were correct and which were misfires. With this information, we can proceed with transcription refinement at true landmarks. To address this, we can define a set of indicator variables,

$$H=\{H_{X_1},\ldots,H_{X_N}\},\quad H_{X_i}=\{h_1^{X_i},\ldots,h_{L_{X_i}}^{X_i}\},$$

where $h_k^{X_i}=1$ if the kth detection of class X_i is a true positive, and 0 otherwise.

Our goal is to determine the most likely broad class sequence, B_{opt}, and which landmarks construct it, H_{opt}. Given the above-presented nomenclature, this can be accomplished by computing the MAP estimate,

$$\begin{aligned}(B_{\text{opt}},H_{\text{opt}}) &= \arg\max_{B,H} P(B,H|O)\\ &= \arg\max_{B,H} P(O|B,H)P(H|B)P(B).\end{aligned}\tag{5}$$

Notice each probability term in this optimization problem is estimable by application of the sonority segmentation and broad class detectors to a corpus of transcribed training data. Furthermore, since our approach is to first segment the utterance down to short analysis units consisting of a limited number of phonemes, we can accomplish optimization by simply calculating the likelihood for all possibilities. If we attempted the same exhaustive approach for word or sentence-long reconstruction units, this combinatorial problem would become prohibitively cumbersome.

1. Decoding obstruent regions

Given an obstruent region determined by the sonority segmentation and the set of observables determined by the obstruent landmark detectors, it remains to apply our proba-

bilistic segment decoding algorithm introduced earlier. [This is equivalent to modeling $P(B_i^{\text{obs}}|O_i^{\text{obs}})$ of Eq. (4) in Sec. II D.] The goal is to determine a transcription of stops, fricatives, and silences present in each obstruent region (i.e., $B \in \{\text{P},\text{F},\text{sil}\}^*$). Note that silence landmarks are included in both the provided set of observables and target sequence search space. However, their inclusion in the final broad class transcription is not necessary for typical applications, and thus will be ignored in our performance evaluations.

Specializing for the task of obstruent segment decoding, we can further simplify the general MAP estimation problem of Eq. (5) with several independence assumptions (in the following expressions, $\mathcal{C}_{\text{obs}}=\{\text{sil},\text{P},\text{F}\}$):

(1) The behaviors of the broad class detectors are independent of each other and the obstruent segment duration,

$$P(O|B,H) = P(T|B) \prod_{X \in \mathcal{C}_{\text{obs}}} P(O_X|B,H).$$

(2) The behavior of each broad class detector depends only on its own indicator variables and not those of other detectors,

$$P(O_X|B,H) = P(O_X|B,H_X).$$

(3) The detection correctness pattern for one broad class detector is independent of that of the others,

$$P(H|B) = \prod_{X \in \mathcal{C}_{\text{obs}}} P(H_X|B).$$

(4) The times of the detections for a particular class are independent of each other. That is, detection times depend only on the broad class sequence encountered and whether the detection is a true positive,

$$P(t_1^X, \ldots, t_{L_X}^X|B,H_X) = \prod_{i=1}^{L_X} P(t_i^X|B,h_i^X).$$

(5) The strengths of the detections for a particular class are independent of each other and the broad class sequence encountered. That is, detection strengths depend only on whether the detection is a true positive,

$$P(f_1^X, \ldots, f_{L_X}^X|B,H_X) = \prod_{i=1}^{L_X} P(f_i^X|h_i^X).$$

(6) Detector strengths and times are independent,

$$P(O_X|B,H_X) = \prod_{i=1}^{L_X} P(t_i^X|B,h_i^X)P(f_i^X|h_i^X).$$

While the extent of the validity of these assumptions has not been rigorously established, their inclusion in the formulation vastly reduces the number of training sentences required to estimate the component distributions. Under these independence assumptions, the optimization problem of Eq. (5) reduces to

$$(B_{\text{opt}},H_{\text{opt}}) = \arg\max_{B,H} P(T|B)P(B) \prod_{X \in \mathcal{C}_{\text{obs}}} P(H_X|B) \times \prod_{i=1}^{L_X} P(t_i^X|B,h_i^X)P(f_i^X|h_i^X). \quad (6)$$

2. Decoding sonorant regions

The decoding of sonorant regions involves a significant complication over the obstruent task. Phonotactic constraints of English and the limitations of the human vocal apparatus make the production of long obstruent sequences extremely unlikely. In fact, in TIMIT all obstruent sequences have a length of four phones or less. This is not the case for sonorant regions; for example, the sentences "We were away all year." and "When will you hear me?" are entirely sonorant. Therefore, in order to limit the combinatorial complexity of sonorant region decoding, we must further segment the signal into smaller, more easily analyzable units. Here, the natural choice is to again turn to the syllable to set the meter of analysis. Since syllables are tied to the vowels present in the sonorant regions, the logical points of separation are vowel landmarks produced by the vowel landmark detector.

Within a given sonorant region, L vowel landmarks determine a series of $L+1$ intervocalic regions that ideally contain sequences of approximants and nasals. Note that the first and last of these regions are bounded by adjacent obstruent regions as determined by the sonority segmentation.[42] As done for obstruent regions, we can apply the probabilistic segment decoding approach to reconstruct the approximant and nasal content of each intervocalic region. [This is equivalent to modeling $P(B_{jk}^{\text{sc}}|O_{jk}^{\text{sc}})$ of Eq. (4) in Sec. II D.] Now, in the case of obstruent regions, T was simply the duration of the entire segment. For intervocalic regions, T is defined as the time elapsed either between adjacent vowel landmarks, between a landmark and an adjacent sonorant region boundary, or, if there are no vowel landmarks, the entire length of the sonorant region. The nasal and approximant landmarks (time and strength) round out the set of observables.

We employ the same set of observable independence assumptions for intervocalic decoding as listed in Sec. III C 1 for obstruent regions. The optimization problem again takes the form of Eq. (6), where the obstruent class set $\mathcal{C}_{\text{obs}}$ is replaced with $\mathcal{C}_{\text{iv}}=\{\text{A},\text{N}\}$ and the possible sequences are $B \in \{\text{A},\text{N}\}^*$.

3. Estimating the probability distributions

Our probabilistic framework, under the independence assumptions described earlier, requires the measurement of several prior distributions. These include:

(1) $P(B)$ for each possible intervocalic and obstruent segment broad class sequence, B.
(2) $P(T|B)$ for each possible broad class sequence, B.
(3) $P(H_X|B)$ for each possible broad class sequence/detector pair, (B,X).

(4) $P(t^X|B,h)$ for each possible broad class sequence/detector/indicator variable value triplet, (B,X,h) for $h \in \{0,1\}$.
(5) $P(f^X|h)$ for each broad class detector/indicator variable value pair, (X,h) for $h \in \{0,1\}$.

Given segmented data, i.e., speech wave forms along with their transcriptions, estimating each of these distributions is fairly straightforward. Thus, $P(B)$, is simply a distribution on a finite set of sequences that occur in obstruent and intervocalic regions, respectively. For each segment S, if we knew the values of H, B, O_{seg}, O_V, O_{sc}, and O_{obs} (see Sec. II D), then we could estimate $P(T|B)$, $P(t^X|B,h)$, and $P(f^X|h)$ as distributions on the real line.

The first step in this process involves running the sonority segmenter on each training sentence. This will result in several sonorant and obstruent regions per sentence for analysis and eventual contribution to the distributions. Likewise, vowel landmarks are computed for the sonorant regions, resulting in multiple intervocalic regions per sonorant region for contribution to the distribution estimation. In general, we may arrive at the "true" value of B for each segment by force aligning the phonetic transcription with the sequence of obstruent and sonorant intervocalic segments determined by the sonority segmenter and vowel landmarks. Since our segmentation may insert or delete segments, we discard any elements of each segment's B that are inconsistent with the segment type. Thus, we ensure the true B is always consistent with the segmentation.

We set T to be the measured duration according to the segmentation and not the actual transcription interval. The detectors result in a collection of landmarks, each consisting of a time–strength pair, (t_i^X, f_i^X). The corresponding indicator variables, $\{h_i^X\}$, can be determined by checking time position against the transcription. This indicator value will determine whether that time–strength pair will be logged in the $h=0$ or $h=1$ distributions. In addition, the individual indicator variables will be combined to form sequences $H_X=(h_1^X,\ldots,h_{L_X}^X)$ for the $P(H_X|B)$ distributions.

There are several possible ways of handling the estimation of these distributions. The most computationally straightforward is the histogram method, which involves simply maintaining a list of all values encountered for each. Using this list, a probability $P(X|Y)$ is calculated by

$$P(X|Y) = \frac{\text{No. cases of } X \text{ in } Y \text{ list}}{\text{length of } Y \text{ list}}.$$

For the discrete distributions, $P(B)$ and $P(H_X|B)$, this histogram prescription is adequate. However, for scalar variable distributions in f, t, and T, we instead implement uniform (i.e., rectangular) kernel density estimation, for which we much choose an appropriate kernel width. This leads to the introduction of three kernel width parameters into the model: Δf, Δt, and ΔT. For example, given a list for the distribution $P(f^X|h=1)$, the probability at a given strength f_0 is estimated by

$$P(f_0|h=1) = \frac{\text{No. cases of } f^X \in [f_0-\Delta f, f_0+\Delta f] \text{ in } h=1 \text{ list}}{\Delta f \times \text{length of } h=1 \text{ list}}.$$

In practice, to limit the model parameters, we choose one set of kernel widths for obstruent region decoding and one set for intervocalic region decoding. However, separate sets could be assigned for the observables for each broad class.

These approaches require a significant amount of training data to provide good distribution estimates. To circumvent this problem, more sophisticated techniques of distribution estimation may also be used. We tried applying Gaussian mixture models, but they resulted in inferior performance. For example, parametric modeling or nonuniform kernel smoothing may improve estimates in the face of limited examples. Exploring these methods lies outside the scope of this paper.

IV. EXPERIMENTAL DETAILS AND RESULTS

In the following, we present the performance results for each of the components of our landmark-based recognition system, as well as the end-to-end broad class transcription performance. When applicable, we evaluate a given component assuming ideal performance of the other elements of the system. This allows us to isolate the merits and shortcomings of each component to aid in future development of integration strategies.

A. Sonority segmentation performance

The support vector machine for the sonority segmenter was trained on a total of 100 "sx" (phonetically compact) and "si" (phonetically diverse) sentences chosen at random from the training section of the TIMIT database. For each of these sentences, 39-dimensional mel frequency (40 spectral bands) cepstral coefficients spanning the full frequency range (0–8 kHz) were computed in 10 ms windows every 5 ms. For the 100 sentences, this translates into approximately 60 000 39-dimensional training vectors, nearly evenly distributed between sonorant and obstruent regions. We employed the SVM^{light} software package[43] to construct the SVM. An operating threshold of 0.1 and the SVM parameters $C=0.0816$ and $\gamma=5\times10^{-4}$ (see Sec. III A) were chosen via holdout validation. This resulted in a frame-level training error of 6.12%.

Likewise, we tested the sonority segmentation performance on 100 randomly chosen sx/i test sentences. Evaluating performance at this initial stage of the overall system is not entirely straightforward. The frame-level test error of 6.44% provides a useful first approximation of SVM performance, but we are primarily interested in evaluating the performance in the context of our overall system architecture. Since our initial sonority segmentation is rigid, we must get this initial classification correct on a phone level to have a chance of correct transcription at later stages. (We will return to this shortcoming in Sec. V.)

This argument points to two phone-level performance metrics:

TABLE II. Performance of the sonority segmentation.

$F_{\min}$	C_{son} (%)	C_{obs} (%)
0.10	98.5	95.4
0.33	96.6	92.9
0.50	95.0	89.3
0.67	93.4	85.8
0.90	82.1	68.7

(1) C_{son}=percentage of the individual sonorant phones for which at least a fraction $F_{\min}$ of its duration falls in a single sonorant segment as determined by the sonority segmenter.
(2) C_{obs}=percentage of the individual obstruent phones for which at least a fraction $F_{\min}$ of its duration falls in a single obstruent segment.

Since the SVM output is effectively smoothed by the windowing parameters, a perfect segmentation down to the sampling interval (1/16 000 s) is impossible. Inaccurate TIMIT transcription time boundaries further complicate the matter. Therefore, setting the minimum overlap $F_{\min}$ to one or even close to one is unreasonable. Further, since the entire phone need not be present in a given region for successful decoding, segmentation performance for low $F_{\min}$ values can still be a good indicator for success in later stages. It is also important to note that segment insertions do not necessarily preclude correct decoding at later stages, as spurious sonorant or obstruent regions may be decoded to be empty. However, each phone must have positive overlap with a proper sonority segment to have a chance to be decoded.

The values of the performance measures as evaluated on our test set are summarized in Table II for various minimum overlap requirements. As expected, the correctness percentages drop as we require larger fractional overlaps. However, the rate of performance decline is higher for obstruent phones. This is largely a result of their shorter average duration, for which systematic errors caused by the windowing parameters and transcription inaccuracies constitute a larger relative portion.

Figure 9 shows the phonetic breakdown of the segmentation errors. More precisely, it displays the number of instances that the durational majority of each phone was placed into an incorrect segment; the sonorant and obstruent phone errors correspond to the 5.0% and 10.7% error rates of the $F_{\min}$=0.5 line of Table II, respectively. For sonorant phones, the largest error contributors are the approximant [hh] and the nasals [m] and [n]. The phone [hh] is a glottal transition whose sonority status is not always well defined (varying definitions may be found in the literature). Therefore, given arbitrary context, mistakes either way are to be expected. Similarly, the nasals are closest to the sonorant–obstruent boundary, so it is not surprising that mistakes occur.

The most prominent mistakes for obstruent phones are for the glottal stop [q] (allophone of /t/) and flap [dx] (allophone of /d/ or /t/). Both allophones are typically of extremely short duration and are surrounded by vowels. Therefore, the sonority profile tends to peak for a minority of the duration or not at all, resulting in a significantly diminished chance of their successful decoding later on. Unfortunately, these allophones are somewhat common. From the error breakdown, we also find that voiced fricatives and stops tend to contribute a higher error rate than their unvoiced counterparts of the same place (e.g. [d] vs [t] or [v] vs [f]). Again, this is not surprising as sonority is largely a measure of the periodicity introduced by glottal voicing and so the SVM training follows suit.

B. Landmark detector performance

Constructing the six landmark detectors required the construction of six support vector machines trained to recognize phones of the target class. We worked with a set of 100 randomly chosen TIMIT sx/i training sentences, though not all frames of all sentences were used for every detector. With full flexibility in constructing each detector, we worked with several representations. For the vowel, approximant, nasal, fricative, and silence detectors, we use 39-dimensional mel frequency (40 bands) cepstral coefficients, but the window size (T_{win}), step size (T_{step}), and frequency range (F_{range}) parameters varied according to Table III. The SVM C and γ

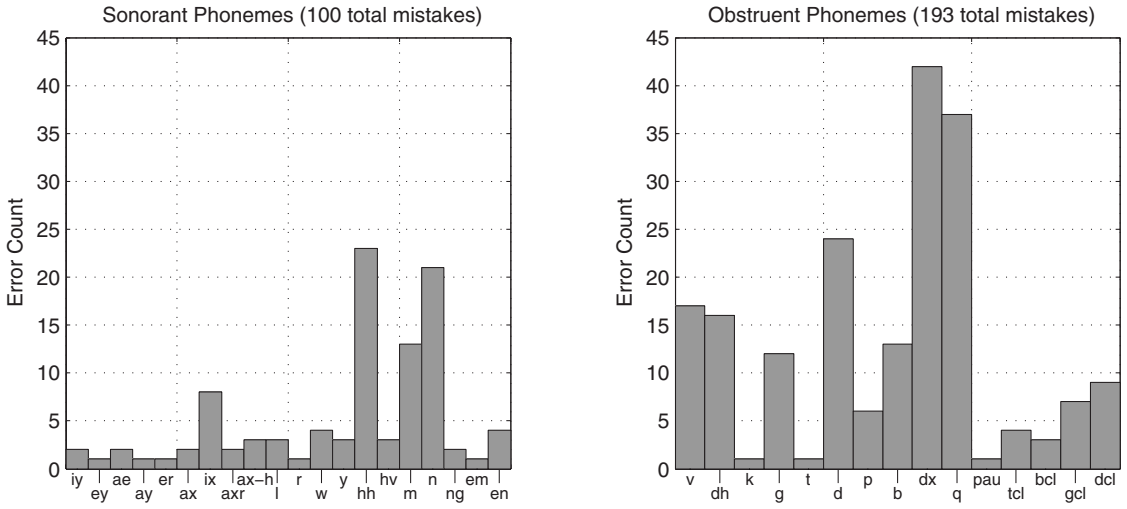

FIG. 9. Phonetic composition of sonority segmentation errors for $F_{\min}$=0.5. Phones for which no mistakes were made are not shown.

TABLE III. Representation and training parameters for the landmark detectors.

Detector	T_{win} (ms)	T_{step} (ms)	F_{range} (kHz)	C	γ	E_{train} (%)
Vowel	40	20	0–4	0.600	0.000 65	10.3
Approximant	20	20	0–8	0.632	0.000 56	19.2
Nasal	30	15	0–8	0.469	0.000 80	6.0
Fricative	30	15	0–8	0.236	0.000 71	6.9
Stop	35	5	N/A	0.064	0.016	6.2
Silence	20	10	0–8	0.017	0.000 53	6.0

parameters, both found via holdout validation, as well as the frame-level training errors (E_{train}) are also listed. For training these five SVMs, all frames centered within the desired phoneme boundaries were considered positive examples.

For the stop detector, we used the acoustic parameter prescription provided by Ref. 29 as an alternative to the MFCC representation, though we modified the frame rate to reduce computational costs. There are three quantities involved in this prescription: total energy, energy above 3 kHz, and the Wiener entropy, $\Sigma_i \log S_i - \log(\Sigma_i S_i)$, where $\{S_i\}$ is the discrete Fourier amplitude spectrum. Each of these parameters was computed in nonoverlapping 5 ms windows. Each 21-dimensional training vector consisted of seven consecutive frames of this type, spanning an effective window width of 35 ms. Given this representation, the SVM has implicit access to the differences in the three parameters over the seven component frames, allowing for the identification of inflection points associated with the stop–burst transition. Furthermore, the 5 ms temporal resolution provides adequate precision for the detection of this transition. The training setup and parameters for the stop detector are also provided in Table III. For training, only the frame centered closest to the stop–burst transition was considered a positive example. Therefore, there were only as many positive examples per sentence as there were stop phones present. For this reason, we limited the number of negative frames to a random sample that is five times the number of positive frames.

Figure 10 shows the phoneme-level landmark performance curves for each detector as a function of threshold. Here, the false positive (FP) rate is defined as the number of spurious detections divided by the number of negative (i.e., not target) phones; the false negative (FN) rate is defined as the number of missed detections of the target class divided by the number of phones of the target class present. Since these rates are computed on a per landmark basis, arbitrarily decreasing the threshold does not necessarily mean the false negative rate goes to zero, distinguishing the curves from the traditional receiver operating characteristic (ROC) variety. However, the operating thresholds are chosen at the point of equal FN and FP rates, as overlaid on each plot. For the vowel detector curve, we include degenerate (extra correct detections within a single phone) vowel detections in the FP count, as they will necessarily result in insertions. However, for the other five detectors, degenerate landmark detections are ignored, as the probabilistic integration model can accommodate such mistakes.

Unfortunately, the stop detector prescription we use was not designed to identify the above-mentioned allophones [q] and [dx]. However, for our purposes we cannot ignore these somewhat common sounds, so our detector phoneme-level error rate is seven points higher than the performance quoted in Ref. 29. Ideally, these allophones would require the construction of their own detector, which can either function on its own or as part of a composite stop detector. We will return to this issue in our discussion of future research directions in Sec. V.

Table IV shows the broad class breakdown of insertion errors made by each detector, along with the number of correct, degenerate, and deleted landmarks. There are several points to note from this breakdown. First, we find that a significant majority of errors are made between broad classes

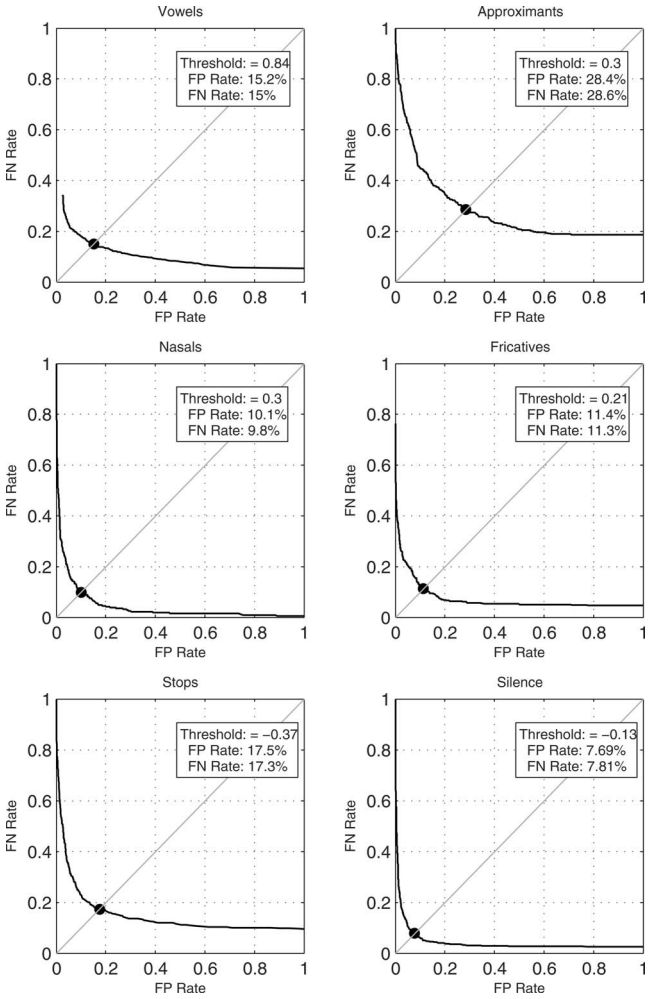

FIG. 10. Phoneme-level ROC curves for the six detectors. The operating thresholds taken at the equal error points are listed on the plots.

TABLE IV. Insertion composition by broad class, number correct, number deleted, and number of degenerate detections for each landmark detector.

Detector	V	A	N	F	P	sil	Degen	Del
V	1036	84	26	96	112	62	21	174
A	606	329	159	28	103	30	51	122
N	125	82	276	42	32	72	81	30
F	7	8	12	400	229	128	237	50
P	160	93	47	182	473	74	63	98
sil	28	18	46	60	24	722	764	42

of the same sonority superclass. For example, only 7% of the fricative detector misfires occur within sonorant phones. This trend validates our choice of the sonorant feature as an appropriate initial point of speech segmentation. Still, we find that mistakes across the sonorant–obstruent distinction do occur. However, so long as the sonority segmentation is correct in the region, such spurious landmarks will be discarded.

Next, we find that the vowel detector results in the lowest degenerate landmark rate (only 21 degenerate vowel landmarks in 1210 vowel phones). This is a direct result of the dynamic baseline algorithm presented in Sec. III B 2. While using a fixed threshold would result in a slightly higher detection rate, the massive resulting increase in degenerate detections would greatly decrease the accuracy. However, the higher detection rate that a fixed threshold provides is beneficial for the other detectors, since the landmark integration strategy is not significantly harmed by degeneracy. As we can see from Table IV, degeneracy rates can be quite high for the longer fricative phones and silence regions.

The weakest link by far in the set is the approximant landmark detector. While the intervocalic decoding model can clean up these mistakes to some extent, poor detector performance definitely translates into unstructured probability distributions and ultimately poor integration performance. However, the low relative approximant incidence rate in typical speech still allows for good overall performance. We will return to the topic of detector-level inadequacies in Sec. V.

C. Obstruent region decoding performance

To separate the performance of the obstruent decoding model from that of the sonority segmenter, we conducted experiments using the actual sonority segmentations provided by the TIMIT transcription for both training and testing.

Prior distribution data were collected from 1000 randomly chosen sx/i training sentences. According to the prior $P(B)$ estimated from TIMIT training data, there are 42 possible sequences of stops, silences, fricatives that may lie in any obstruent region. Using an additional 100 training sentences, we determined optimal-accuracy kernel width parameters of $\Delta f=0.03$, $\Delta t=0.03$ (i.e., 3% of the obstruent region), and $\Delta T=10$ ms. Note that optimal kernel width parameters differ when estimated sonority segmentations are used.

We tested on all 1344 sx/i sentences contained in the TIMIT test set. For each obstruent region, the decoding results in a predicted transcription. We evaluate this prediction relative to the actual sequence present using minimum string edit distance alignment. For N phones encountered, this evaluation will result in the number of correctly identified phones (C), number of deleted phones (D), number of substituted phones (R), and number inserted phones (I). The accuracy can then be computed as $A=1-(D+R+I)/N=(C-I)/N$.

Table V shows the transcription performance for several variations of the decode procedure. The first is a naive measure of baseline performance without using the probabilistic model. Here, we simply chronologically sort the landmarks above the appropriate operating threshold in each obstruent region. The predicted sequence is simply the corresponding broad classes of these landmarks. The second method is the standard implementation of the probabilistic decoding method outlined in this paper. Finally, the two "Rank$\leqslant N$" methods assume we have an oracle that can identify the true obstruent region sequence if it is one of the N most probable sequences. (The standard decode is equivalent to "Rank$\leqslant 1$.") In all of these variations, we ignore silence landmarks in the performance analysis since their presence is not necessary in the final transcription.

The poor naive baseline performance illustrates the main problem with integrating multiple error prone detectors: correctness rates average, but insertion rates add together. However, our probabilistic decoding approach effectively cleans up false detections, admitting an insertion rate of only 6% while maintaining the correctness rate of the baseline. As we consider more than just the top sequence, prediction accuracy quickly improves further. These Rank$\leqslant N$ methods provide a ceiling estimate of expected model performance when higher-level linguistic constraints are incorporated, such as a language model. In such a setting, multiple high-probability predictions can be considered in the context of a word or phrase, effectively providing an approximation to an oracle function. While our model attains a 77% phoneme recognition accuracy in the top choice, it is exceedingly good at

TABLE V. Obstruent region decoding performance on 17 525 phones (fricative and stops) contained in 15 766 obstruent regions.

Method	Accuracy	% correct	% ins	% del	% repl
Baseline	42.0	79.2	37.2	14.9	5.9
Std. deccode	77.0	83.0	6.0	6.0	11.1
Rank$\leqslant 2$	89.2	92.1	2.9	3.8	4.2
Rank$\leqslant 3$	93.8	94.7	0.9	3.0	2.4

TABLE VI. Intervocalic region decoding performance in 12 915 phones (approximants and nasals) contained in 36 255 intervocalic regions.

Method	Accuracy	% correct	% ins	% del	% repl
Baseline	25.5	54.0	28.5	42.5	3.5
Std. decode	53.0	69.9	16.9	23.2	6.9
Rank ≤ 2	85.1	90.4	5.3	6.7	2.9
Rank ≤ 3	95.1	96.8	1.7	2.1	1.0

paring down the original 42 possibilities to a few candidate sequences, with close to 90% accuracy in the top two ranking candidates alone.

D. Intervocalic region decoding performance

For the performance evaluation of intervocalic decoding, we again determine region boundaries using the actual TIMIT transcription for both training and testing. The transcription center point of each vowel is used in place of detected vowel landmarks. This procedure allows us to isolate the performance of the intervocalic decoding from that of both the sonority segmenter and the vowel landmark detector. We estimate the prior distributions using the same 1000 randomly chosen TIMIT training sentences used for the obstruent region model. We again accomplish alignment using minimum string edit distance and use holdout validation to estimate the optimal-accuracy kernel width parameters of $\Delta f = 0.1$, $\Delta t = 0.1$ (i.e., 10% of the intervocalic region), and $\Delta T = 10$ ms.

We again test on all 1344 sentences contained in the TIMIT test set. According to the prior $P(B)$ estimated from the TIMIT training data, there are now only 12 possible sequences of nasals and approximants that may lie in any intervocalic region (including the empty sequence). Table VI shows the transcription performance for the same methods studied for obstruent region decoding. We find significantly lower performance of both the naive baseline and the standard decode relative to obstruent region decoding. This is largely due to the exceptionally poor performance of the approximant detector, as discussed in Sec. IV B. In particular, the high insertion rate of approximant landmarks in nasal phones is especially detrimental to the reliability of probabilistic predictions. Still, our decoding method significantly reduces the insertion rate while increasing the correctness percentage over the baseline performance, resulting in more than twice the accuracy. The ranking methods result in an even more striking performance improvement. When considering just the two most likely sequences, the accuracy improves 32% (absolute) over the standard decoding method. This portends great improvements in this module of the system when higher-level linguistic constraints are imposed.

E. Overall performance

We now turn to the overall performance of our landmark-based broad class recognizer, implementing the sonority segmenter, landmark detectors, and probabilistic segment decoding for both obstruent and intervocalic regions. Because we are now using estimated sonority boundaries and vowel landmarks, the optimal kernel width parameters have been retuned using holdout validation, as listed in Table VII. We find that the intervocalic parameters and the obstruent region Δf parameter are the same as when we used the actual TIMIT transcription boundaries. However, optimal values for the two timing parameters (Δt and ΔT) for obstruent regions have increased. This is largely a result of an imprecise sonority segmentation, resulting in a significant smear of the probability distributions. In this degraded setting, the model performs better when the distributions are more highly smoothed.

TABLE VII. Optimal kernel width parameters for probability distribution estimation.

Region type	Δt	ΔT (ms)	Δf
Obstruent	0.2	25	0.03
Interocalic	0.1	10	0.1

We tested our system and four continuous CMU Sphinx-3 HMM variations,[44] using both context-independent (CI) and dependent (CD) decoding with either broad class (BC) or individual phone (Ph) three-state models. Each HMM was trained on all 3696 TIMIT sx/i training sentences, using standard 39-dimensional mel frequency cepstral coefficients (Sphinx feature set 1s_c_d_dd), eight-mixture Gaussian observation densities, no skip transition. Furthermore, in our HMM experiments, no language model is applied (including no transition probability rescaling), restricting the study to the domain of pure speech recognition.

In our system, probabilistic segment decoding is a context-dependent approach, though the sonority segmentation and vowel landmark detection methods are context independent. Therefore, our composite system is only partially context dependent. The context-dependent HMM system uses triphone models in all cases, so its use of contextual information is more complete. The complexity of our system is closest to the HMMs using broad class models, as we only implement one detector per broad class. For HMMs using phoneme models, the resulting individual phoneme decoding is collapsed into a broad class transcription. Minor variations among the individual phones within each broad class may be captured in these otherwise redundant models, and the performance gain is significant.

Minimum string edit distance alignment was performed for all five systems. Table VIII summarizes the broad class transcription performance on all 1344 sx/i test sentences.[45] Our system accuracy exceeds that of both CI and CD HMMs using broad class models, which are of most similar complexity. We also exceed the accuracy of the context-independent HMM using phone models, though our system is using contextual constraints. Our system accuracy falls two points short of the context-dependent HMM using phone models. However, as we will discuss in the next section, there is vast room for improvement of our approach.

We find that the context-dependent HMMs attain significantly higher correctness rates relative to their context-independent counterparts. This, however, is at the expense of a significantly higher insertion rate, predominantly a result of not applying a language model for cleanup. Our landmark-

TABLE VIII. Broad class transcription performance for our landmark-based system vs various HMM approaches. HMMs result in high insertion rates in the pure speech recognition domain.

System	Accuracy	% correct	% ins	% del	% repl
Our System	70.3	76.0	5.7	11.3	12.7
HMM, CI/BC	65.5	68.4	2.9	17.6	13.9
HMM, CD/BC	65.1	90.5	25.4	1.4	8.1
HMM, CI/Ph	69.0	79.7	10.7	6.0	14.2
HMM, CD/Ph	72.2	91.5	19.3	1.6	6.9

based system is a conservative guesser, resulting in a low insertion rate similar to the context-independent HMMs. This is largely a result of landmark thresholding before decoding. Theoretically, the probabilistic decoding requires no thresholding, as low strength landmarks will have a correspondingly low $P(f|h=1)$. Decreasing or even removing the operating thresholds do increase correctness rates, but at the expense of insertions. We find that the accuracy levels are slightly higher after detector thresholding.

V. DIRECTIONS FOR FUTURE RESEARCH

In the preceding sections we have presented a landmark-based speech recognition frame-work fashioned on the principles outlined in Sec. I. Our broad class recognizer based on these ideas is competitive with equivalent HMM-based systems. Still, our implementation involves several design choices that are not necessarily optimal, leaving significant room for improvement of our computational approach. The immediate focus of our future research is to translate the ideas contained in this paper to practical speech recognition tasks, of which full phonetic recognition (discussed in detail in the following) is one example. However, history demonstrates that for all successful new approaches to the recognition problem, state-of-the-art performance was eventually attained by incremental advances in the various modules of the initial framework. Thus, in the following, we also discuss in some detail several areas for improvement, along with possible solutions, each addressing a limitation of our current implementation.

A. Full phonetic recognition

There are two distinct strategies to accomplish full phonetic recognition using the ideas developed in this paper.

1. Modeling phonetic sequences

This paper outlines a possible approach to map firing patterns of detectors into sequences over a symbolic inventory. Therefore, there are two classes of objects that are worth distinguishing. The first is $\mathcal{D}=\{d_f(t)|f\in\mathcal{F}\}$, a collection of feature detections. Here $\mathcal{F}$ is a family of feature types and for each feature type, $d_f(t)$ is a detector firing pattern associated with that type. The second object is a finite symbol set Σ and sequences over this symbol set, i.e., elements of Σ^*. In the system we have implemented, we chose $\Sigma=\{\mathrm{V},\mathrm{N},\mathrm{A},\mathrm{F},\mathrm{P},\mathrm{sil}\}$ (i.e., the set of broad classes), while $\mathcal{F}$ corresponds to the small number of leaf nodes in the distinctive feature hierarchy of Fig. 2.

In order to transition to full phonetic recognition, we need to let Σ be the set of phonetic units and, correspondingly, we will need to consider a more exhaustive set $\mathcal{F}$ of features. As long as the detectors for the set $\mathcal{F}$ and their firing patterns have sufficient discriminative power to separate all the phonemic classes from each other, our approach is in principle applicable to full phonetic recognition. There are two challenges. The first is to find an adequate set of feature detectors. The second is to deal with the possible combinatorial explosion one might anticipate if one were searching over all phonetic sequences rather than broad sequences alone.

Fortunately, neither challenge is insurmountable. In particular, our recent work[46] demonstrates the feasibility of scaling up to full phonetic recognition. In the implementations we have experimented with, we chose $\mathcal{F}$ to be a set of phonetic units and trained the corresponding phone detectors using methods from statistical learning. This larger set is sufficient to achieve phonetic recognition accuracy rates competitive with basic HMM systems.

Regarding the combinatorial challenge, it turns out that given the syllable-sized analysis units proposed in this paper, one only has to search over highly constrained phone sequences. Note that an important part of our strategy is to chunk the signal into syllabic nuclei, intervocalic sonorant sequences, and intervocalic obstruent sequences. Thus, for broad class recognition, we only need to consider possible broad class sonorant sequences and broad class obstruent sequences that can possibly lie between two adjacent vowels. These numbers are very small (12 and 42 in TIMIT, respectively, including silences). Our phone recognition experiments on TIMIT have shown that when one considers intervocalic sonorant and obstruent *phone-level* sequences, the numbers that occur are surprisingly limited (61 and 385, respectively, not including silences).

Thus, while there are two broad sonorant consonantal classes (nasal and approximant) that may combine to $32=2^5$ possible strings of length less than or equal to 4, only 12 actually occur in TIMIT. Similarly, though there are ten sonorant consonants and therefore 10^5 possible strings at the phonetic level, only 61 actually occur. These numbers illustrate that phonotactic and phonological constraints of the language dramatically reduce the set of possible consonantal sequences one needs to consider given the high-level chunking into syllable-sized units. Note that the statistics collected from TIMIT provide only an approximation to the distribution of phonological sequences in natural speech. However,

it seems evident that naturally occurring sequences will remain (with high probability) an exceedingly small subset of logically possible sequences.

In light of these experiments, it is clear that the principles outlined in this paper can be adapted to yield a viable phonetic recognition strategy. On the other hand, the approach explored in Ref. 46 is by no means the optimal execution of our high-level principles. For instance, a phone detector set could be replaced with a more general set of acoustic property detectors. Similarly, the search through possible sequences could be managed through various pruning or indexing strategies.

2. Transcription refinement

A second strategy is to perform transcription refinement of the broad class transcription provided by the system described in this paper. This involves expanding the distinctive feature hierarchy of Fig. 2 to include children of the current leaves that make distinctions between the individual phonemes within each broad class in a manner similar that suggested by Ref. 36. For example, adding place and voicing nodes under the stop leaf is adequate to distinguish between p [+labial, −voiced], b [+labial, +voiced], t [+alveolar, −voiced], d [+alveolar, +voiced], k [+velar,−voiced], and g [+velar,+voiced]. Determining additional feature values for transcription refinement can be accomplished using SVMs or any appropriate machine learning method.

In this new setting, the role of the current system is to provide landmarks around which these further features may be determined in a context-dependent way. Recall that the H variables of Sec. III C were introduced precisely for this reason. However, a complication arises from our probabilistic segment decoding method, which is capable of predicting a sequence for which there were degenerate detections (i.e., multiple correct candidates) or even no detections at all (this is rare). In these cases, it is not entirely clear where the true point for further analysis lies. When there are degenerate detections, an average of the landmark times weighted by their strength seems a reasonable choice. When a landmark is missing, we could take the maximal point of the prior time distribution, which amounts to the model's best estimate for the predicted context. Still, it may be the case that given a postulated broad class sequence with insufficient landmark information, more significant analysis must be performed. However, we believe improvements discussed earlier will not only improve broad class transcription, but also the selection of correct landmarks. Such an improvement will minimize this type of complication involved in transcription refinement.

B. Probabilistic sonority segmentation and vowel landmarks

Our method of probabilistic segment decoding (PSD) provides a means to accommodate error-making detectors to produce a list of likely transcription sequences for each obstruent or intervocalic segment. In contrast, the first two stages of our architecture, sonority segmentation and vowel landmark detection, are hard decisions. Therefore, mistakes made by these modules cannot be recovered. Furthermore, since they determine the obstruent and intervocalic regions for PSD, their errors propagate through later stages. This results in two significant performance bottlenecks.

The simplest approach to minimizing these bottlenecks is to improve the hard decisions they make. In the current scheme, this means improving the SVM performance, possibly through alternative signal representations, or even using alternative machine learning techniques. For example, neural networks have also been shown to be useful in this domain (see Ref. 47). Still, a more robust approach to solving the bottleneck problem is to transition to probabilistic sonority segmentations and vowel landmarks. Using such an approach, we could consider multiple candidate segmentations and vowel landmark sets for a given utterance.

This effectively amounts to performing the optimization of Eq. (1) over multiple candidate O^{seg} and O^{v}, in addition to B (see Sec. II D). Under this scheme, a less likely, but more accurate candidate segmentation can lead to better PSD performance and, ultimately, a more accurate transcription. A possible approach to implementing probabilistic vowel landmark detections arises from the dynamic baseline algorithm presented in Sec. III B 2. The number of landmarks chosen increases with the recursion depth d of the algorithm, resulting in one set V_d for each depth (i.e., $V_{d-1} \subset V_d$). The probability $P(V_d)$ of each set can be computed by

$$P(V_d) = \prod_{v \in V_d} P(h_v = 1 | f_v),$$

where $P(h_v=1|f_v)$ is the probability that the vowel landmark v is correct $(h_v=1)$ given its strength f_v. An approach for probabilistic sonority segmentation is not as immediate, though a scheme using variable thresholds may be useful.

C. Language model incorporation

Clearly, the above-presented broad recognition accuracies for HMMs are not the numbers we are used to seeing for state-of-the-art systems. In our experiments, we did not implement a phone- or word-level language model to clean up the phonetic transcription, resulting in lower broad class performance than is normally associated with HMM systems. The question remains of whether our framework would also admit comparable gains when a language model is incorporated.

We found in Secs. IV C and IV D that our probabilistic segment decoding method, while not always successful at choosing the correct sequence, provides exceedingly accurate N-best estimates. This means that if we impose higher-level linguistic constraints, we could have a superior chance at recovering the actual sequence from multiple top choices. Phoneme-, syllable-, or word-level n-gram models are the common choice for HMMs and would easily lend themselves to application within our probabilistic framework. Transitioning to probabilistic sonority segmentations and vowel landmark sets would result in even deeper language model benefits. However, it remains to be seen if the language model benefit for our system will exceed that of HMMs.

D. Landmark detector improvement

The individual landmark detectors are a major area for improvement that would lead to immediate overall performance gains. Reducing the number of detector insertions and deletions would put less burden on the integration procedure. Furthermore, more accurate landmarks would also sharpen the prior distributions, increasing the reliability of the posterior estimates.

One possible approach would be to implement acoustic parameters (APs) as an alternative to MFCCs. In our current implementation, the stop detector APs resulted in superior performance and computational efficiency. There is a significant body of existing research detailing the merits of APs over MFCCs for the broad classes (for examples, see Refs. 29, 31–33, 38, and 39). These parameters have been shown to increase robustness, exhibiting higher noise and speaker invariance.

Another approach is to individually address specific phoneme-level detector errors. One example is our stop detector's poor performance in [q] and [dx] detection. In this case, a possible remedy would be to create separate detectors for these problematic phones, resulting in multiple classifiers for each broad class. These classifiers could be integrated into a single broad class detector, using a logical OR of the subdetectors. This logic could be extended to a separate subdetector for every phoneme, resulting in complexity similar to HMMs using phoneme models.

There is also room for redesign of the detector set itself. In particular, we could implement broad class transition detectors, one for each ordered broad class pair. This would result in an augmented set of observables, but the probabilistic segment decoding formulation would remain exactly the same. A possible benefit of such an approach would be the sharpening of the prior time distributions. As it stands, a fricative detection, for example, can occur anywhere within the phoneme. Transition detectors, on the other hand, would be contained in much smaller regions, resulting in more pronounced distribution structure. However, it is unclear how the error rates of such detectors would fare to those currently in place.

The last approach for detector improvement would be to implement a different machine learning approach. SVMs have the nice property of providing a maximal separation between the two data classes, which helps reduce generalization error. However, the typically high number of support vectors involved in evaluation can be computationally taxing. Since the landmarks' detectors are independent modules of the system, we could implement any combination of machine learning methods here, so long as at the end of the day a series of landmarks are output. A study of various methods in this context will be required to determine suitable alternatives.

E. Alternative integration models

A final direction for system improvement is the probabilistic integration model. We have already touched on the possibility of alternative prior distribution estimation techniques. Another avenue is to explore the consequences of limiting the number of independence assumptions made in our probability model. More sophisticated detector time normalization methods may also be studied in order to sharpen the prior time distributions. The current approach of using the fraction of the obstruent or intervocalic region elapsed before a landmark is certainly better than using absolute times. However, for a stress-timed language such as English, there is still significant variation in the relative timing of phones in different contexts (see Ref. 48 for a discussion of the issues involved here).

The MAP approach developed in this paper is by no means the only statistical framework that can be employed (see Refs. 6, 49, and 50 for other examples). One possible alternative is to model the detector firings as Poisson processes with rate parameters dependent on the broad class sequence present and the region of that sequence you are in. For example, for the sequence "P F" the Poisson rate parameter for the stop detector would be high in the first half of the region and low in the second. While it is unclear if this approach would yield better results, it fits nicely with the point process representation provided by the detector hierarchy. It may also provide computational expression of neural coding theories of auditory processing.

[1]An example of this is our ability to learn new words, especially in a foreign language. In situations like this, we clearly have little or no access to morphological, syntactic, or higher level constraints, so our ability to recognize words in this case is an instantiation of pure speech recognition.

[2]J. Goldsmith, *Autosegmental Phonology* (Garland Press, New York, 1979).

[3]E. Sagey, "The representation of features and relations in non-linear phonology," Ph.D. thesis, MIT, Cambridge, MA, 1986.

[4]J. J. McCarthy, "Feature geometry and dependency: A review," Phonetica **43**, 84–108 (1988).

[5]K. H. Esser, C. J. Condon, N. Suga, and J. S. Kanwal, "Syntax processing by auditory cortical neurons in the FM-FM area of the mustached bat pteronotus parnellii," Proc. Natl. Acad. Sci. U.S.A. **94**, 14019–14024 (1997).

[6]Z. Chi, W. Wu, and Z. Haga, "Template-based spike pattern identification with linear convolution and dynamic time warping," J. Neurophysiol. **97**, 1221–1235 (2007).

[7]K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," J. Acoust. Soc. Am. **111**, 1872–1891 (2002).

[8]A. Juneja and C. Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," in *Proceedings of IJCNN*, International Joint Conference on Neural Networks, Portland, OR, July 20–24, 2003.

[9]M. Hasegawa-Johnson *et al.*, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins Summer Workshop," in *Proceedings of ICASSP*, International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA, March 18–23, 2005.

[10]J. R. Glass, "A probabilistic framework for segment-based speech recognition," Speech Commun. **17**, 137–152 (2003).

[11]S. Riis and A. Krogh, "Hidden neural networks: A framework for HMM/NN hybrids," in *Proceedings of the ICASSP*, International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, April 21–24, 1997.

[12]D. Huttenlocher and V. Zue, "A model of lexical access from partial phonetic information," in *Proceedings of the ICASSP*, International Conference on Acoustics, Speech and Signal Processing, San Diego, CA, March 1984.

[13]C. Demiroglu and D. V. Anderson, "Broad phoneme class recognition in noisy environments using the GEMS," in *Proceedings of the ACSSC*, Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, November 7–10, 2004.

[14]C. Espy-Wilson, T. Pruthi, A. Juneja, and O. Deshmukh, "Landmark-based approach to speech recognition: An alternative to HMMs," in *Proceedings of Interspeech*, Antwerp, Belgium, August 27–31, 2007.

[15]Not all 2^k possibilities may be instantiated as phonemes in a particular language. However, every phoneme in a natural language is one of these 2^k elements. In this sense, they constitute a universal inventory. Furthermore, the distinctive features allow one to characterize the set of natural classes and see why they are in fact a very small number of the logically possible subsets of 2^k phonemes. According to this theory, the natural classes are defined as all possible partial assignments to a k-dimensional boolean vector, admitting $O(3^k=\Sigma_{i=1}^{k}\binom{k}{i}2^i)$ possibilities. In contrast, without this boolean vector structure, one could potentially have $O(2^{2^k}=\Sigma_{i=1}^{2^k}\binom{2^k}{i})$ possible classes.
[16]M. Halle, "Phonology," in *Language*, edited by D. Osherson and H. Lasnik (MIT, Cambridge, MA, 1990), Vol. **1**, pp. 43–68.
[17]N. Chomsky and M. Halle, *The Sound Pattern of English* (Harper & Row, New York, 1968).
[18]M. Kenstowicz, *Phonology in Generative Grammar* (Blackwell, Oxford, 1994).
[19]G. A. Miller and P. Nicely, "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352 (1955).
[20]K. N. Stevens, *Acoustic Phoenetics* (MIT, Cambridge, MA, 1998).
[21]R. Jakobson, G. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features* (MIT, Cambridge, MA, 1952).
[22]S. G. Parker, "Quantifying the sonority hierarchy," Ph.D. thesis, University of Massachusetts-Amherst, Amherst, MA, 2002.
[23]The sonority hierarchy arranges classes of sounds into equivalence classes arranged in decreasing order of sonority. This leads to the hierarchy Vowels > Approximants > Nasals > Fricatives > Affricates > Stops. Making a split at the Nasal/Fricative boundary realizes the binary distinctive feature [±son].
[24]One need not have a hard segmentation. One could postulate many candidate segmentations with a probability associated with each candidate segmentation. The associated probability may then be one of several considerations in settling upon an ultimate segmentation based on bottom-up and top-down analysis.
[25]The exceptions to this rule are syllabic nasals and liquids. These sounds are not protypical vowels in the acoustic sense, but function as syllabic nuclei nonetheless. In our implementation, we choose to regard them as sonorant consonants with the intent of improving vowel recognition performance at the expense of recognizing sonorant consonants, which compose a smaller fraction of the phonetic content of English.
[26]A vowel landmark is a point in the vowel at which its "vowelness" is most strongly exhibited. This corresponds to the peak of the sonority profile.
[27]S. Greenberg, H. Carvey, L. Hitchcock, and S. Chang, "Temporal properties of spontaneous speech—A syllable-centric perspective," J. Phonetics **31**, 465–485 (2003).
[28]Our use of the word classifier may lead to some confusion. Most machine learning methods produce a real-valued function that is thresholded to yield a binary-valued output. We use "classifier" to denote the real-valued function.
[29]P. Niyogi and M. M. Sondhi, "Detecting stop consonants in continuous speech," J. Acoust. Soc. Am. **111**, 1063–1076 (2002).
[30]C. Burges and P. Niyogi, "Detecting and interpreting acoustic features with support vector machines," Technical Report No. TR-2002-02, Computer Science Department, University of Chicago, Chicago, IL, 2002.
[31]C. Espy-Wilson, "Acoustic measures for linguistic features distinguishing the semivowels /wjrl/ in American English," J. Acoust. Soc. Am. **92**, 736–757 (1992).
[32]A. Salomon, C. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," J. Acoust. Soc. Am. **115**, 1296–1305 (2004).
[33]T. Pruthi and C. Espy-Wilson, "Acoustic parameters for automatic detection of nasal manner," Speech Commun. **43**, 225–239 (2004).
[34]Y. Amit, A. Koloydenko, and P. Niyogi, "Robust acoustic object detection," J. Acoust. Soc. Am. **118**, 2634–2648 (2005).
[35]D. W. Tank and J. J. Hopfield, "Neural computation by concentrating information in time," Proc. Natl. Acad. Sci. U.S.A. **84**, 1896–1900 (1987).
[36]A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. thesis, University of Maryland, College Park, MD, 2004.
[37]This includes one energy and 12 cepstral coefficients, along with their delta and acceleration (double-delta) coefficients.
[38]O. Deshmukh, C. Espy-Wilson, and A. Juneja, "Acoustic-phonetic speech parameters for speaker-independent speech recognition," in *Proceedings of the ICASSP*, International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, May 13–17, 2002.
[39]Z. Xie and P. Niyogi, "Robust acoustic-based syllable detection," in *Proceedings of the ICSLP*, International Conference on Spoken Language Processing, Pittsburgh, PA, September 17–21, 2006.
[40]We will use Arpabet phonetic notation for the remainder of the paper to facilitate connections with experimental result figures.
[41]Note that the threshold for the vowel detector is not shown as it is not applied to the raw classifier output shown in the figure.
[42]In this sense, the use of the word "intervocalic" is technically a misnomer. However, to simplify discussion, we use it to refer to both segments bounded by two vowels and segments bounded by one or two sonorant–obstruent boundaries.
[43]T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*, edited by B. Schlkopf, C. Burges, and A. Smola (MIT, Cambridge, MA, 1999).
[44]K. Seymore *et al.*, "CMU Sphinx-3 English broadcast news transcription system," in *Proceedings of the DARPA Speech Recognition Workshop*, DARPA Broadcast News Transcription and Understanding Workshop, February, 1998.
[45]The reader might notice that the HMM broad class recognition accuracies quoted in Table VIII are in the ballpark of published phone recognition accuracies on the TIMIT database. However, such high phonetic accuracy quotes are the result of applying word-level language models or transition probability rescaling. To substantiate this claim, we reproduced a recent CI/Ph HMM experiment performed in Ref. 51 on TIMIT data that quotes a *phonetic* recognition accuracy of 67% when applying transition probability rescaling. Our implementation of their system led to similar phonetic recognition performance and translated into 81% *broad class* recognition accuracy, which is significantly higher than the numbers we quote for Sphinx-3. However, when the transition probability rescaling is removed from their decoding procedure, the broad class accuracy falls to 70%, which is consistent with our Sphinx-3 results.
[46]A. Jansen and P. Niyogi, "Point process models for event-based speech recognition," Technical Report No. TR-2008-04, Computer Science Department, University of Chicago, 2008.
[47]S. Chang, L. Shastri, and S. Greenberg, "Automatic phonetic transcription of spontaneous speech (American English)," in *Proceedings of the ICSLP*, International Conference on Spoken Language Processing, Beijing, China, October 16–20, 2000.
[48]E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin, 2003), Vol. **1**, pp. 515–546.
[49]W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects," J. Neurophysiol. **93**, 1074–1089 (2005).
[50]N. Morgan, H. Bourlard, S. Greenberg, and H. Hermansky, "Stochastic perceptual auditory-event-based models for speech recognition," in *Proceedings of the of ICSLP*, International Conference on Spoken Language Processing, Yokohama, Japan, September 18–22, 1994.
[51]F. Sha and L. K. Saul, "Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models," in *Proceedings of the ICASSP*, International Conference on Acoustics, Speech, and Signal Processing, Honolulu, HI, April 15–20, 2007.