Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 51 (2009) 1155–1168

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Point process models for event-based speech recognition

Aren Jansen *, Partha Niyogi

University of Chicago, Department of Computer Science, 1100 E 58th Street, Chicago, IL 60637, United States

Received 27 February 2008; received in revised form 22 May 2009; accepted 24 May 2009

Abstract

Several strands of research in the fields of linguistics, speech perception, and neuroethology suggest that modelling the temporal dynamics of an acoustic event landmark-based representation is a scientifically plausible approach to the automatic speech recognition (ASR) problem. Adopting a point process representation of the speech signal opens up ASR to a large class of statistical models that have seen wide application in the neuroscience community. In this paper, we formulate several point process models for application to speech recognition, designed to operate on sparse detector-based representations of the speech signal. We find that even with a noisy and extremely sparse phone-based point process representation, obstruent phones can be decoded at accuracy levels comparable to a basic hidden Markov model baseline and with improved robustness. We conclude by outlining various avenues for future development of our methodology.

PACS: 43.72.Ne; 43.72.−p

Keywords: Event-based speech recognition; Speech processing; Point process models

1. Introduction

In this paper, we investigate statistical point process models in the context of automatic speech recognition. Such models arise naturally if one wishes to explicitly engage the following facts regarding speech production and perception:

(1) Speech is generated by the movement of independent articulators that produce acoustic signatures at specific points in time. Some examples are the point of greatest sonority at the center of a syllabic nucleus, the points of closure and release associated with various articulatory movements such as closure-burst transitions for stop consonants; obstruent–sonorant transitions; and onsets and offsets of nasal coupling, frication, or voicing. Phonetic information is coded both in terms of which events occur as well as the durations between these events (e.g. voice onset time). Stevens (2002) refers to such points in time as acoustic event landmarks and assigns them a central status in lexical decoding.

(2) Perceptual and neurophysiological studies of speech perception (see Poeppel et al. (2007) for an account) suggest that there are two fundamental time scales at which information is processed. The first is the time scale at which various segmental and subsegmental units occur (25–80 ms). The second is the time scale at which suprasegmental or syllabic integration occurs (150–300 ms). This suggests that phonetic information is integrated at syllabic timescales and syllable sized units are perceptual primitives that are central to phonetic decoding (see Greenberg et al. (2003) for a related treatment).

(3) A series of neuroethological studies has identified neurons that fire selectively when a certain constellation of acoustic properties are present in the stimulus. For example, the existence of such combination-sensitive

* Corresponding author.
E-mail addresses: aren@cs.uchicago.edu (A. Jansen), niyogi@cs.uchicago.edu (P. Niyogi).

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.05.008

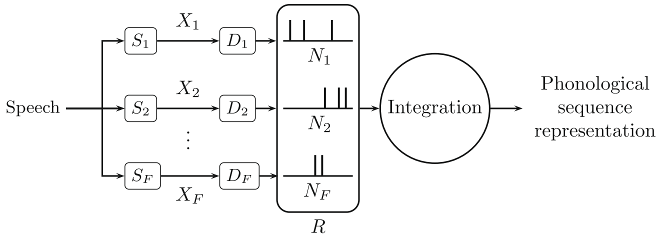

Fig. 1. Architecture of our event-based framework. In general, we construct one signal processor S_i for each acoustic property of interest ($F = |\mathscr{F}|$), which produces a specialized representation X_i. Each representation is input to a detector D_i for the property, producing a point pattern N_i. The combined set of point patterns for all of the detectors (R) is probabilistically integrated to predict a phonological sequence.

neurons in the auditory cortex of several animal species has been demonstrated (birds by Margoliash and Fortune (1992), bats by Esser et al. (1997), and frogs by Fuzessery and Feng (1983)). These findings led to the formulation of the detector hypothesis (see Suga, 2006), which states that a biologically important acoustic signal is represented by the excitation of detector (or, more generally, information-bearing parameter filter) neurons selectively responsive to its presence. The related synchronization hypothesis suggests that auditory information is further encoded in the temporal pattern of such neural activity, i.e., temporal coding. There is evidence that such principles are instantiated in auditory systems more generally (Suga, 2006).

Taken together, these observations suggest that speech may be (i) adequately represented as an asynchronous collection of acoustic or perceptual events that *need not be tied to a common clock or constant frame rate*, and (ii) decoded according to the *temporal statistics* of such events. The need therefore arises to formulate and evaluate recognition strategies that can operate on representations based on the firing patterns of nonlinear detectors specialized for various acoustic events or properties.

Thus we consider a sparse detector-based representation of the speech signal that should efficiently encode the underlying linguistic content. In general, the detector set may include detectors for any set $\mathscr{F}$ of linguistic properties (e.g. phones or distinctive features) or acoustic signatures (e.g. band energy inflection points or periodicity maxima).[1] The linguistic information is a sequence over some alphabet $\mathscr{P}$, which may, for example, be the set of phones, broad classes, distinctive features, articulatory variables, or even syllables or words. Fig. 1 shows a schematic of our architecture.

In this paper, we assume one has a detector for each phonological unit $p \in \mathscr{P}$ (i.e. $\mathscr{F} = \mathscr{P}$), each producing a point pattern $N_p = \{t_1, \ldots, t_{n_p}\}$, where each $t_i \in \mathbb{R}^+$. Arrivals of each process, which may be viewed as acoustic event landmarks, should ideally occur when and only when the corresponding phonological unit is maximally articulated and/or most perceptually salient. Furthermore, asynchronous detectors imply that the quantization of arrivals of each phonological unit's point process may vary. In practice, creating an ideal detector is of course unachievable, so we may generalize this notion to marked point processes, $\{N_p, M_p\}$, where the marks $M_p = \{f_1, \ldots, f_{n_p}\}$ are interpreted as the strengths (e.g. probabilities) of the corresponding landmarks.

It is worthwhile to note that there has been significant recent interest in what has been termed *detection-based* speech recognition (Lee, 2007), especially in the context of the Automatic Speech Attribute Transcription collaboration (Ma et al., 2006). These approaches have some motivational and architectural similarities to our proposed framework, including (i) accommodation of asynchronous and/or overlapping distinctive and articulatory featural representations; (ii) modular parallel detectors for the features of interest; and (iii) a statistical integration module that combines the feature detectors in order to decode the linguistic content. However, it is important to emphasize that the approach proposed in this paper is a continuation of the earlier detection-based philosophy of Stevens and others (see Stevens et al., 1992; Stevens, 2002; Niyogi et al., 1998). In particular, we are not interested in detectors that produce frame-level estimates for a feature, but instead in detectors that identify the points in time that relevant speech events occur (i.e. landmarks). Accordingly, we consider dynamic models of the temporal statistics of such events rather than frame-based modelling of feature probabilities. It is also in this way that our approach differentiates itself from other graphical-model-based proposals for incorporating distinctive and articulatory feature-based phonological systems in speech recognizers (e.g. Deng and Sun, 1994; Sun and Deng, 2002; Livescu and Glass, 2004).

In Section 2, we consider several statistical models that are natural choices when presented with such a marked point process representation of the speech signal. In order to evaluate the potential merits of each model, we consider the problem of phonetic recognition in obstruent regions, a speech recognition subtask that is consistent with the multiscale analysis hypothesis of Poeppel et al. (2007). In particular, this subtask comprises one module in our previous hierarchical approach to recognition in which one first chunks the signal into sonorant and obstruent regions and decodes each separately (see Jansen and Niyogi, 2008). While decoding these constrained-length obstruent sequences may be viewed as a large multi-class classification

[1] The design of a suitable family of detectors is itself the subject of an interesting program of research (see Stevens and Blumstein, 1981; Stevens, 2002; Niyogi and Sondhi, 2002; Hasegawa-Johnson, 2002; Pruthi and Espy-Wilson, 2004; Li and Lee, 2005; Amit et al., 2005; Xie and Niyogi, 2006)). However, we will not explore this question in any detail here. Rather, we will assume that a detector-based representation is made available to us and models for recognition will have to be constructed from such representations. In our own experiments in this paper, we choose a simple phone-based detector set, which we define in Section 3.1.

task, we evaluate performance in the context of a recognition problem, tabulating phone-level insertion, deletion, and substitution errors.

Given the linguistic and neuroscientific motivation described above, we view the investigation of point process models for speech recognition as a natural research question. Yet to the best of our knowledge, there has been no prior study of the potential use of such models in automatic speech recognition. For related investigations in the context of neuroscience, see Brown (2005), Chi et al. (2007), Truccolo et al. (2005), and references therein. From our experiments, we find that by adopting a suitable statistical model, it is possible to recover the linguistic content of the speech signal from an extremely sparse point process representation. In addition to the information-theoretic efficiency that such sparse coding provides, we believe that sparse representations are more invariant and thus may lead to greater robustness in the resulting recognition systems. While this assertion has not been previously explored for speech, it certainly has merit in context of visual processing (see Olhausen, 2003; Geiger et al., 1999; Serre et al., 2007).

2. Statistical models

In this section, we present several statistical models to recover the phonological sequence generating a predefined supersegment[2] (in this paper, regions of constant sonority) of the speech signal given the point process representation defined above. The approaches fall into two categories: global and supersegment-level.

The naive and hidden Markov model-based approaches are global in that they are applied to continuous speech of arbitrary length; the phonological sequence prediction for the predefined supersegments are subsequently extracted from the global decode. In particular, the hidden Markov model-based approaches accomplish this by applying a dynamic programming algorithm (global Viterbi decode) on the entire utterance; likewise, the naive approach involves no probabilistic modelling and thus global processing is achieved trivially.

Contrastingly, the explicit time-mark and Poisson process models are supersegment-level models; that is, we first process the utterance into supersegments whose space of possible underlying sequences must be limited by phonological constraints. This leads us to a supersegment-level maximum *a posteriori* estimation strategy reminiscent of standard stochastic (phonetic) segment models (see Ostendorf et al., 1992; Ostendorf, 1996); however, our motivations lead us to a fundamentally distinct approach to the recognition problem.

2.1. Naive approach

The simplest method of converting a set of point process patterns $\{N_p\}$ to a label sequence S is to sort the landmarks and read off the labels. Formally, given a set of landmarks $\{t_i^{p_i}\}$ over phonological units $p_i \in \mathscr{P}$ where $t_i^{p_i} < t_j^{p_j}$ for $i < j$, the global prediction is determined by

$$S_{\text{global}} = p_1 p_2 \dots p_N. \tag{1}$$

As indicated above, the prediction for any predefined supersegment of the speech signal may be extracted from this global sequence using the landmark times.

The obvious problem with this approach is that integrating insertion-prone detectors in this manner quickly leads to a significant deterioration in performance. For example, integrating 20 detectors, each with a mere 5% false positive rate, could theoretically combine to a 100% overall insertion rate. It follows that successful decoding of a noisy point process representation will require a probabilistic detector integration strategy.

2.2. Hidden Markov model of the point process representation (HMM-PP)

Next, we consider the application of a standard hidden Markov model (HMM) to our point process representation $R = \{N_p, M_p\}_{p\in\mathscr{P}}$. If we limit ourselves to point patterns that are synchronous[3] (i.e. for all $t_p \in N_p$ and $t_{p'} \in N_{p'}$, there exists $n, m \in \mathbb{Z}^+$ and $\Delta t \in \mathbb{R}^+$ such that $t_p = n\Delta t$ and $t_{p'} = m\Delta t$), we may construct a sparse vector time series representation $X = \vec{x}_1\vec{x}_2 \dots \vec{x}_T$ defined by

$$\vec{x}_l[j] = \begin{cases} f_k \in M_{p_j} & \text{if } \exists k \text{ s.t. } t_k \in N_{p_j} \text{ and } t_k = l\Delta t \\ 0 & \text{o/w} \end{cases} \tag{2}$$

We can then proceed to apply an HMM model to recover the hidden state sequence, $Q = q_1q_2\dots q_T$ for $q_t \in \mathscr{Q}$, by maximizing the joint likelihood over Q and X. Here, $\mathscr{Q}$ represents the state space, which contains exactly one state for each element in $\mathscr{P}$. Under the Markov assumption, the joint likelihood takes the form

$$\log P(X, Q) = \sum_{t=1}^{T} [\log P(\vec{x}_t|q_t) + \log P(q_t|q_{t-1})]. \tag{3}$$

[2] For lack of existing terminology, we will use the term *supersegment* to refer to a region of the speech signal where any phonological feature is held constant, including regions of constant sonority. A supersegment may contain one or more phonetic segments. This is not to be confused with (though, can be related to) the term *suprasegmental*, which refers to vocal effects that extend over more than one phonetic segment in an utterance (typically pitch or tone).

[3] Several recently-proposed, multiple stream HMM-based methods could be implemented to accommodate an asynchronous point process representation (for examples, see Mak and Tam, 2000; Zhang et al., 2003; Nock and Ostendorf, 2003). We leave a study involving such methods applied to asynchronous detector sets for future work.

Given the lattice of emit probabilities and matrix of state transition probabilities, the standard Viterbi algorithm[4] is applied to determine the state sequence Q^* for the entire utterance that maximizes this joint likelihood. The global phonological sequence prediction S_{global} may then be determined directly from Q^* by collapsing repeated states, and the phonological sequence prediction for any predefined supersegment may be extracted from S_{global} accordingly.

The transition probabilities $P(q_t|q_{t-1})$ are determined by counting the frame-level transitions occurring in the training data. (Note that all TIMIT utterances, which were used exclusively in the present study, begin with leading silence and thus predicate q_0 to be a silence state.) For modelling the distributions $P(\vec{x}_t|q_t)$ over the new input vector space, which tends to have sparse support, applying the traditional Gaussian mixture model (GMM) is not a natural choice, nor does it work in practice. We instead consider two more appropriate models: (i) binomial mixture models (BMM) for the unmarked point process representation, and (ii) histogram method estimation for the marked representation.

For the case of an unmarked point process representation, where the vector time series X is binary-valued, we model the emission densities using B-component multivariate binomial mixture models of the form

$$P(x_t|q_t = q) = \sum_{b=1}^{B} \omega_{qb}\mathscr{B}(\vec{\pi}_{qb})(\vec{x}_t), \tag{4}$$

where $\sum_{b=1}^{B}\omega_{qb} = 1$ $(\omega_{qb} > 0)$ for each $q \in \mathscr{Q}$. Here, the function $\mathscr{B}(\vec{\pi}_{qb})$ is the bth binomial mixture component in the context of state q, given by

$$\mathscr{B}(\vec{\pi}_{qb})(\vec{x}) = \prod_i \left(\pi_{qb}[i]\right)^{x[i]}\left(1 - \pi_{qb}[i]\right)^{(1-x[i])}, \tag{5}$$

where $\pi_{qb}[i] \in [0,1]$ is the bth mixture component probability of a detection in the ith dimension of X in the context of state q. A maximum likelihood estimate of the BMM parameters may be found using the expectation-maximization (EM) algorithm.

If we consider a marked point process representation, the vector time series is no longer binary-valued and the BMM is no longer applicable. Instead, we consider a histogram estimate of the vector space with a common bin width Δx for all coordinates. Assuming the coordinates are conditionally independent, we may write

$$P(x_t|q_t = q) = \prod_{j=1}^{|\mathscr{P}|} H_{qj}(\vec{x}_t[j]), \tag{6}$$

where H_{qj} is the histogram estimate of the distribution of the jth coordinate in the context of state q. This formulation is equivalent to a classical, quantized (component-wise) observation density HMM-based approach (Frangoulis, 1989), but where most of the observation vectors components are zero.

Finally, it is important to note that the sparse nature of the point process representation can produce a significant amount of zero vectors (i.e., $\vec{x}_t = \vec{0}$), which occur at times when no landmarks are present. The emission probability distributions estimated for each state will each yield a constant value $K_q = P(\vec{0}|q)$ when the zero vector is encountered. If we set aside the transition probabilities for a moment, it follows that for all t such that $\vec{x}_t = \vec{0}$, the optimal state is always $q_t = \arg\max_{q\in\mathscr{Q}} K_q$, which could conceivably lead to serious insertion problems. Therefore, it is vital that the transition probabilities be able to prevent falling into this default state every time the zero vectors occur. If not, a possible solution is to define an augmented state space $\mathscr{Q}' \equiv \{\mathscr{Q}, \epsilon\}$, where ϵ is a null state to model the zero vectors. Then, occurrences of this null state in the decoding can simply be omitted.

2.3. Explicit time-mark model (ETMM)

Consider a maximum *a posteriori* (MAP) estimate of the phonological sequence S that generates the supersegment in the interval $[T_1, T_2]$, given the observed point process representation $R = \{N_p, M_p\}_{p\in\mathscr{P}}$ and the interval duration $T = T_2 - T_1$. This MAP estimate takes the form

$$S^* = \arg\max_{S\in\mathscr{P}^*} P(S|R,T) = \arg\max_{S\in\mathscr{P}^*} P(R|S)P(T|S)P(S), \tag{7}$$

where we have assumed conditional independence between the point process representation and the interval duration.[5] At this point, we have an optimization problem with a high-level form that is similar to that used for stochastic (phonetic) segment models (Ostendorf, 1996), as we factor the objective function into terms for both the segment duration and the observations in the segment. However, we deviate from this established paradigm in two key ways:

[4] It should be noted that in the context of supersegment decoding, it may be beneficial to introduce explicit supersegment durational modelling into the HMM framework using one of the standard approaches proposed in the past for phonetic duration modelling. These include hidden semi-Markov models (Levinson, 1986), expanded state HMMs (Russell and Cook, 1987), post-processor duration penalties (Juang et al., 1985), and inhomogeneous HMMs (Ramesh and Wilpon, 1992). While exhibiting varying levels of success when used in the context of phone duration modelling, these approaches introduce significant technical complications without community-standard solutions. We leave an investigation of these methods in the context of supersegment duration modelling for future work.

[5] It is worthwhile to reflect on the inclusion of the term $P(S)$. The reader may notice that such a source of phonological information, which amounts to a unigram language model over sequences that occur in the supersegments, was not included in the HMM-PP formulation of Section 2.2. It has been suggested that this may give the ETMM and Poisson process model approaches an unfair advantage. However, it is important to recall that HMM-based methods, including the HMM-PP approach, make use of frame-level state transition probabilities in their global Viterbi decode. In the context of decoding obstruent regions, which we focus on in this paper, these state transition probabilities can give the HMM-based approaches an advantage of its own. That is, obstruent region predictions can be beneficially influenced, through transition constraints, by the surrounding sonorant phone contexts; this is not possible with obstruent sequence-level unigram probabilities alone.

(i) we consider models of longer, phonologically-motivated supersegments (e.g. regions of constant sonority), and (ii) we construct models of the temporal dynamics of the point process representation in these supersegments, as opposed to models of the observation space across the frames of each supersegment.

In particular, we would like to deal with the term $P(R|S)$ by explicitly modelling the times and strengths of landmarks observed. Since all landmarks lie in a fixed interval $[T_1, T_2]$, we begin by normalizing the supersegment length and landmark times to the interval [0,1]. We make the simplifying assumption that all landmarks are independent, allowing us to factor $P(R|S)$ into

$$P(R|S) = \prod_{p\in\mathscr{P}} \prod_{i=1}^{n_p} P(t_i^p, f_i^p|S). \tag{8}$$

Training requires the estimation of the distribution over $(t,f) \in [0,1]^2$ for each S. Given sets of supersegment training examples containing each possible S, these distributions can be found using standard techniques such as histogram or kernel smoothing methods once given the observed landmarks.

In our experiments, we employ a uniform kernel density estimator for $P(T|S)$ and $P(t^p, f^p|S)$. For the univariate $P(T|S)$ distributions, this takes the form

$$P(T|S) = \frac{1}{N\Delta T}\sum_{i=1}^{N} K\left(\frac{T - T_i}{\Delta T}\right), \tag{9}$$

where $K(x) = 1[|x| < 1], \Delta T$ is the smoothing bandwidth, and $\{T_i\}_{i=1}^N$ are the durations of N supersegment training examples containing S. For the bivariate kernel density estimates of $P(t^p, f^p|S)$, we write

$$P(t,f|S) = \frac{1}{L\Delta t\Delta f}\sum_{i=1}^{L} K\left(\frac{t - t_i^p}{\Delta t}\right) K\left(\frac{f - f_i^p}{\Delta f}\right), \tag{10}$$

for time and strength bandwidths Δt and Δf, respectively, where the data $\{(t_i^p, f_i^p)\}_{i=1}^L$ are the time-strength pairs for all L landmarks of class p observed in supersegments containing S.

2.4. Poisson process model

In the previous section, we considered explicit modelling of the point process arrival times in each supersegment of interest. However, the marked point process representation suggests a Poisson process model as an alternative, natural choice for the $P(R|S)$ term of Eq. (7). This model comes in two varieties: homogeneous and inhomogeneous. A homogeneous Poisson process assumes that in any differential time interval dt the probability of an arrival is λdt, where $\lambda \in \mathbb{R}^+$ is the process rate parameter. This probability is independent of spiking history, resulting in a memoryless point process. For the inhomogeneous case, the constant rate parameter is generalized to a time-dependent function $\lambda(t)$, but the memoryless property still holds. Finally to handle a marked point process, we can consider a rate parameter $\lambda(t, f)$, which depends on both the time t and the strength f of the landmark. As done for the explicit time-mark model, we must normalize the landmark times in each obstruent region to the interval [0,1] for each Poisson process model variant discussed below.

2.4.1. Homogeneous Poisson process

Consider a collection of independent point patterns $N_p = \{t_1, \ldots, t_{n_p}\}$, one for each $p \in \mathscr{P}$, contained in the interval $(0, T]$. If $\eta_p(t) \equiv |\{t_i \in N_p | t_i \leqslant t\}|$ is the number of landmarks in the interval $(0, t]$, then for a homogeneous Poisson process, we may write

$$\mathbb{P}_{a,b}(k) \equiv \mathbb{P}[\eta_p(b) - \eta_p(a) = k] = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!}, \tag{11}$$

where $\tau = b - a$. It follows the probability that the first arrival occurs *after* time t is $\mathbb{P}[t_1 > t] = \mathbb{P}_{0,t}(0) = e^{-\lambda t}$. Therefore, the probability that the first landmark lies in the interval $(t, t + dt]$ is $\mathbb{P}[t_1 \in (t, t + dt]] = \lambda e^{-\lambda t}dt$, which leads to a corresponding density function

$$f(t) = \lambda e^{-\lambda t}. \tag{12}$$

Since the process is memoryless, the likelihood[6] of the whole point pattern becomes

$$P(N_p) = \mathbb{P}_{t_{n_p},T}(0) \times f(t_1) \prod_{i=2}^{n_p} f(t_i - t_{i-1}) = \lambda^{n_p} e^{-\lambda T}. \tag{13}$$

It follows that the conditional likelihood of the whole representation $R = \{N_p\}$, given the phonological sequence S, takes the form

$$P(R|S) = \prod_p [\lambda(p,S)]^{n_p} e^{-\lambda(p,S)T}, \tag{14}$$

where $\lambda(p, S)$ depends both on the generating phonological sequence S and the phonological unit p of the point pattern being evaluated.

Training this model, then, amounts to estimating $\lambda(p, S)$ for each (p, S) pair. In particular, if we are given N normalized-length supersegment training examples containing the sequence S, and the total number K of landmarks of type p observed in those examples, the maximum likelihood estimate of $\lambda(p, S)$ is

$$\lambda^*(p,S) = \arg\max_{\lambda} K\log\lambda - \lambda NT = K/NT. \tag{15}$$

2.4.2. Inhomogeneous Poisson process

For the inhomogeneous case, we consider a piecewise continuous rate parameter over D divisions of the interval $(0, T]$ given by $\lambda(t) = \lambda_d$ for $d = \text{ceiling}(t/\Delta T)$, where $\Delta T = T/D$. In this case, the Poisson process can be

[6] Usually, we will use the notation P to denote likelihood of the data, i.e., the density evaluated at the data points. We use $\mathbb{P}(E)$ to denote the probability of the event E.

factored into D independent processes operating in each piece of the interval. That is, if

$$N_{p,d} \equiv N_p|_{I(d)}, \tag{16}$$

where $I(d) = ((d-1)\Delta T, d\Delta T]$, and $|N_{p,d}| = n_{p,d}$, then the likelihood of an individual pattern is determined by

$$P(N_p) = \prod_{d=1}^{D} P(N_{p,d}) \tag{17}$$

where

$$P(N_{p,d}) = (\lambda_d)^{n_{p,d}} e^{-\lambda_d \Delta T}. \tag{18}$$

It follows that the maximum likelihood estimation of the rate parameter of the dth subinterval for phonological unit p and generating sequence S is given by

$$\lambda_d^*(p,S) = K_d D/NT, \tag{19}$$

assuming we have been provided with N supersegment training examples containing a total of K_d landmarks in the dth subinterval. Finally, the conditional likelihood of the whole representation given a generating sequence S can be computed as

$$P(R|S) = \prod_{p\in\mathscr{P}} \prod_{d=1}^{D} [\lambda_d(p,S)]^{n_{p,d}} e^{-\lambda_d(p,S)\Delta T}. \tag{20}$$

2.4.3. Marked Poisson process

The generalization of either the homogeneous or inhomogeneous Poisson process model to handle marked point processes is straightforward if we consider spatially dependent rate parameters. In this case, the sole spatial dimension corresponds to the mark space, which we assume is normalized to $[0,1]$, resulting in a mark-dependent rate parameters $\lambda(t,f)$ ($\lambda(f)$ for the homogeneous case). We again implement a piecewise continuous approximation by splitting the mark space into M divisions with $\lambda(f) = \lambda_m$ for $m = \text{ceiling}(fM)$. As before, the Poisson process factors into M independent processes operating in each division of the mark space. For a homogeneous marked Poisson process, we can define

$$N_{p,m} \equiv \{t_i \in N_p | f_i \in M_p|_{I(m)}\}, \tag{21}$$

where $I(m) = ((m-1)/M, m/M]$ and $|N_{p,m}| = n_{p,m}$. It follows that the likelihood of an individual point pattern for a particular phonological unit p is given by

$$P(N_p) = \prod_{m=1}^{M} P(N_{p,m}). \tag{22}$$

where

$$P(N_{p,m}) = (\lambda_m)^{n_{p,m}} e^{-\lambda_m T}. \tag{23}$$

The maximum likelihood estimation of the rate parameter of the mth mark space division for phonological unit p and generating sequence S is given by

$$\lambda_m^*(p,S) = K_m/NT, \tag{24}$$

assuming we have been provided with N supersegment training examples containing sequence S containing a total of K_m landmarks in the mth mark space division. The conditional likelihood of the whole representation given a generating sequence S can be computed as

$$P(R|S) = \prod_{p\in\mathscr{P}} \prod_{m=1}^{M} [\lambda_m(p,S)]^{n_{p,m}} e^{-\lambda_m(p,S)T} \tag{25}$$

The marked Poisson process generalizes to the inhomogeneous case in exactly the same way described for the unmarked case.

3. Experiments in obstruent supersegment decoding

In this section, we consider the speech recognition subtask of decoding consonants in obstruent supersegments, i.e., all obstruent regions of the speech signal that lie between two sonorant phones. This speech recognition subtask, while not typically performed in isolation, arises naturally if one first segments the speech signal into sonorant and obstruent regions and decodes each independently. Our previous work (see Jansen and Niyogi, 2008) has demonstrated the computational viability of this approach. Furthermore, perceptual studies (see Parker, 2002) and computation models of speech perception (see Poeppel et al., 2007) provide scientific motivation for such a central role of the sonorant–obstruent distinction.

Given an obstruent region $[T_1, T_2]$ of duration $T = T_2 - T_1$, we would like to find the most likely sequence $S = p_1 \dots p_n$, where $p_i \in \mathscr{O}$ and $\mathscr{O}$ is the set of obstruent phones. For the HMM-PP method, this amounts to performing a global Viterbi decode of the entire utterance and retrieving the predicted sequence in the interval $[T_1, T_2]$ (discarding any sonorant states that might have spilled into the region). That is, if S_{global} is the global phonetic decode, then S is taken to be the subsequence of S_{global} that is restricted to both $\mathscr{O}$ and $[T_1, T_2]$.

For the explicit time-mark and Poisson process models of Sections 2.3 and 2.4, respectively, obstruent region decoding amounts to finding the S that maximizes $P(S|R')$, where $R' = R|_{[T_1,T_2]}$. Given the linguistic constraints on the length of obstruent sequences, there are only 385 possible obstruent sequences in the TIMIT corpus.[7] This limit facilitates the feasibility of direct $P(S|R')$ computation for each possible sequence.

All experiments were conducted using the TIMIT speech corpus, consisting of a total 3696 training and 1344 test sentences, read by both males and females spanning the continental United States. We held out 100 randomly chosen training sentences for any required nuisance parameter tuning, and trained all models using the remaining 3596 sentences. All performance evaluations were conducted

[7] While TIMIT only contains a subset of the possible sequences present in the English language, we believe longer sequences remain sufficiently rare in natural settings to ignore for our purposes.

using all test sentences. We defined our phonological unit set $\mathscr{P}$ to be the standard 48 phone set defined by Lee and Hon (1989) and used in later work by Sha and Saul (2007).

3.1. Constructing the point process representation

We require a map from the speech signal $s(t)$ to a collection of point patterns $R = \{N_\phi, M_\phi\}_{\phi\in\mathscr{F}}$, where $\mathscr{F}$ is some set acoustic or linguistic properties that is adequate to differentiate the phonological units in $\mathscr{P}$. This mapping is accomplished using the following three components:

(1) Given W windows of the signal collected every Δ_ϕ seconds, construct for each $\phi \in \mathscr{F}$ an acoustic front end that produces a k_ϕ-dimensional vector representation $X_\phi = x_1, \ldots, x_W$, where $x_i \in \mathbb{R}^{k_\phi}$. Each representation X_ϕ should be capable of isolating frames in which feature ϕ is expressed and, to that end, the window and step sizes may be varied accordingly.
(2) Construct a detector function $g_\phi : \mathbb{R}^{k_\phi} \to \mathbb{R}$ for each $\phi \in \mathscr{F}$ that takes high values when feature ϕ is expressed and low values otherwise. Each detector may be used to map X_ϕ to a detector time series $\{g_\phi(x_1), \ldots, g_\phi(x_W)\}$.
(3) Given a threshold δ, compute the point pattern (N_ϕ, M_ϕ) for feature ϕ according to

$$N_\phi = \{i\Delta_\phi | g_\phi(x_i) > \delta \quad \text{and} \quad g_\phi(x_i) > g_\phi(x_{i\pm1})\} \tag{26}$$
$$M_\phi = \{g_\phi(x_i) | i\Delta_\phi \in N_\phi\}.$$

Here, we assume $N_\phi = \{t_1, \ldots, t_{n_\phi}\}$ and $M_\phi = \{f_1, \ldots, f_{n_\phi}\}$ are ordered such that $t_{i+1} > t_i$ and $f_i = g_\phi(x_j)$, where $j = t_i/\Delta_\phi$.

In our experiments presented in this paper, we take our feature set $\mathscr{F}$ to be the set of phones $\mathscr{P}$ (i.e., there is a one-to-one correspondence between features $\phi \in \mathscr{F}$ and phones $p \in \mathscr{P}$). While the point process representation can theoretically (and perhaps, ideally) be constructed from multiple acoustic representations tuned for each phonetic detector, we implemented a single shared front end for all of the phone detectors. In particular, we employed the rastamat package (Ellis, 2005) to compute a traditional 39-dimensional Mel-frequency cepstral coefficient (MFCC) feature set for 25 ms windows sampled every 10 ms. This included 13 cepstral coefficients computed over the full frequency range (0–8 kHz), as well as 13 delta and 13 delta–delta (acceleration) coefficients. Cepstral mean subtraction was applied on the 13 original coefficients, and principal component diagonalization was subsequently performed for the resulting 39 dimensional vectors.

In general, the simplest approach to constructing the detector functions is to independently train a one-vs.-all regression function for each phonological unit using any suitable machine learning method. That is, given L labelled MFCC training examples $\{(x_l, p_l)\}_{l=1}^{L}$, where each $x_l \in \mathbb{R}^{39}$ is contained in a segment of phone $p_l \in \mathscr{P}$, we would like to compute a set of detector functions $g_p : \mathbb{R}^{39} \to [0, 1]$ such that $g_p(x) = P(p|x)$. In our implementation, we used the normalized MFCC vectors for each phone to estimate the $P(x|p)$ distributions assuming a C-component GMM for each $p \in \mathscr{P}$, given by

$$P(x|p) = \sum_{c=1}^{C} \omega_{pc} \mathscr{N}(\vec{\mu}_{pc}, \boldsymbol{\Sigma}_{pc})(x), \tag{27}$$

where $\omega_{pc} > 0$ and $\sum_{c=1}^{C}\omega_{pc} = 1$ for each $p \in \mathscr{P}$; and $\mathscr{N}(\vec{\mu}, \boldsymbol{\Sigma})$ is a normal distribution with mean $\vec{\mu}$ and full covariance matrix $\boldsymbol{\Sigma}$. The maximum likelihood estimate of these GMM parameters are found using the expectation-maximization (EM) algorithm on the training data $\{(x_l, p_l)\}_{l=1}^{L}$. These distributions determine the family of detector functions, $\{g_p\}$, as

$$g_p(x) = P(p|x) = \frac{P(x|p)P(p)}{\sum_{p\in\mathscr{P}} P(x|p)P(p)}, \tag{28}$$

where $P(p)$ is the frame-level probability of phone p as computed from the training data. Note that for each model presented below, we measured performance for $C \in \{1, 2, 4, 8\}$ to study the dependence on detector reliabilities.

Fig. 2 shows for an example sentence the evaluation of $\log P(x|p)$ and the corresponding point process representation after applying a threshold of $\delta = 0.5$ (the threshold that results in optimal Poisson process model performance). The drastic reduction of information resulting from the conversion produces an exceedingly sparse point process representation.

3.2. Evaluation procedure

From each test sentence, we used the accompanying transcription to produce a set of obstruent regions to be decoded. Recall that for the HMM and naive method, the obstruent region predictions are extracted from the global decode, discarding any sonorant phones that spill over. With the transcription-provided truth and model prediction for each obstruent region in hand, the set of 48 phones were collapsed into the standard 39 units according to the equivalence sets {cl,vcl,epi,sil}, {l,el}, {n,en}, {sh,sh}, {ao,aa}, {ix,ih}, and {ax,ah}. To facilitate comparison with HMM methods, which cannot predict repeated phones, we also collapsed such occurrences.

We proceeded by scoring the predicted sequences using minimum string edit distance alignment with the truth sequence in each obstruent supersegment. This results not only in a measurement of the recognition accuracy/error rates, but also a breakdown of the errors into insertion, deletion, and substitution types, which we provide in the discussion of each model.

3.3. Results

3.3.1. Naive baseline results

Since we are interested in decoding obstruent regions, the naive baseline approach requires only the subset of

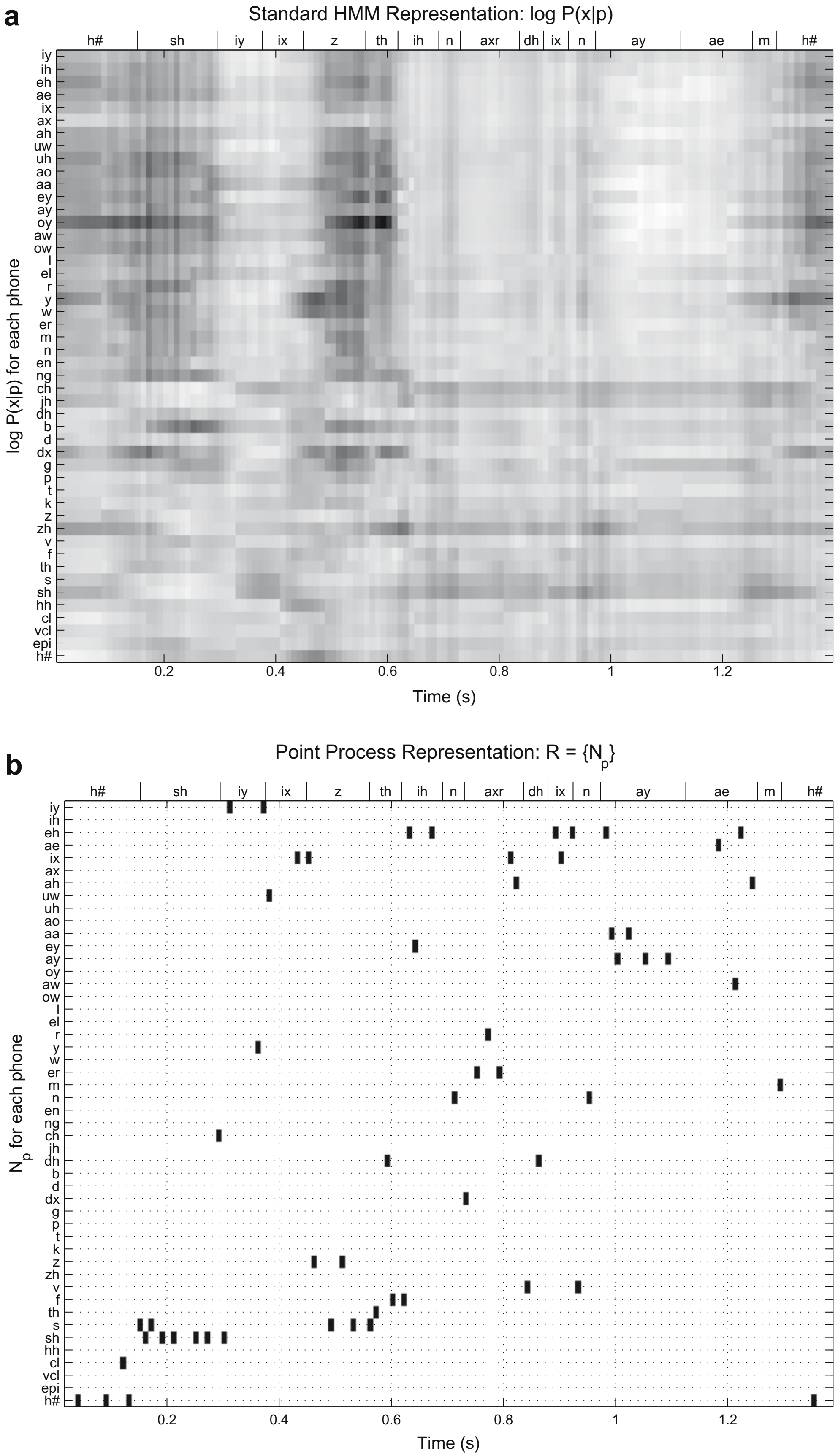

Fig. 2. (a) The lattice of $\log P(x|p)$ values for the utterance "she is thinner than I am", where higher probability is lighter. (b) The corresponding (unmarked) point process representation, $R = \{N_p\}_{p\in\mathscr{P}}$ for $\delta = 0.5$.

the point process representation produced by obstruent phone detectors (i.e., $\{N_p, M_p\}_{p\in\mathscr{C}}$). To determine an operating threshold, we varied the value from 0 to 1 in increments of 0.05 and chose the setting that maximizes the recognition accuracy on the holdout set. It is important to note that the optimal value for this naive approach is not necessarily the optimal value when implementing other methods. In particular, since this naive approach is primarily susceptible to insertion errors, achieving maximal accuracy necessitates a comparatively high threshold setting. The probabilistic models we consider allow us to consider lower probability landmarks without such high insertion rates.

Table 1 shows the obstruent recognition accuracy using this naive approach for several values of C, the number of GMM components used to construct the feature detectors. The increasing detector reliability with higher values of C results in accuracy gains, as expected. However, we also find that for the lower two values of C, a lower threshold value is required to achieve optimal accuracy. Note that if we set the threshold to achieve correctness rates in line with the other methods, the resulting accuracies become negative (i.e., the insertion rate exceeds the correctness rate). This fact illustrates the necessity of a suitable probabilistic model to clean spurious firings of the noisy detectors.

Table 1
Obstruent phone recognition performance for the naive (baseline) method.

C	δ	Accuracy	% Corr	% Ins	% Del	% Sub
1	0.90	34.0	43.9	9.9	37.3	18.8
2	0.90	38.4	54.4	16.0	25.7	19.9
4	0.95	41.4	53.5	12.1	30.8	15.7
8	0.95	44.4	56.9	12.5	27.7	15.4

3.3.2. HMM-PP results

To apply HMM methods to the point process representation, we constructed the sparse vector time series as described in Section 2.2. As mentioned above, we performed a global Viterbi decode on the entire TIMIT utterances and used the TIMIT phonetic transcription to retrieve the predictions in the obstruent regions (any non-obstruent phones that spilled over into these regions were thrown away before scoring).

We initially attempted to model the sparse marked point process vector time series data with a Gaussian mixture model (the standard density model for typical HMM-based systems), which resulted in performance below the naive baseline. We then performed experiments using both the binomial mixture model for the unmarked point process representation and the explicit model using the histogram method for the marked representation. Table 2 shows the obstruent recognition accuracy for various detector reliabilities, where we have employed a 2-component BMM to model the vector time series constructed from the unmarked point process representation. For each value of C, a detector threshold of $\delta = 0.5$ and no null state produced optimal results. We find a steep increase in the deletion and substitution rates as the detector set becomes less reliable, while low insertion rates are achieved across the board.

Table 2
Obstruent phone recognition performance for an HMM with binomial mixture models applied to the unmarked point process representation.

C	δ	Accuracy	% Corr	% Ins	% Del	% Sub
1	0.5	47.6	49.9	2.3	22.3	27.3
2	0.5	54.8	57.2	2.4	18.1	24.7
4	0.5	58.9	61.4	2.5	15.5	23.1
8	0.5	60.7	63.7	3.1	14.5	21.8

Table 3 lists the obstruent recognition accuracy for a applying the histogram estimate observation densities given a marked point process representation. A detector threshold of $\delta = 0.5$, no null state, and a coordinate bin width of $\Delta x = 0.05$ produced optimal results for all detector reliabilities. Low insertion rates coupled with a significant reduction in substitution errors result in accuracy improvements over the unmarked representation using BMMs.

Table 3
Obstruent phone recognition performance for an HMM with histogram estimates of the observation densities, as applied to a marked point process representation.

C	δ	Accuracy	% Corr	% Ins	% Del	% Sub
1	0.5	51.1	53.1	2.0	22.2	24.7
2	0.5	58.1	60.4	2.3	17.1	22.4
4	0.5	61.6	64.0	2.4	14.9	21.0
8	0.5	63.6	66.2	2.6	14.0	19.8

3.3.3. Explicit time-mark model results

For the explicit time-mark model, we solved the optimization problem of Eqs. (7) and (8) over the 385 possible obstruent phone sequences. In our implementation, we performed uniform kernel density estimation of the distributions $P(T|S)$ and $P(t,f|S)$. As described in Section 2.3, this introduces three kernel bandwidth parameters with optimal values ($\Delta t = 0.3, \Delta T = 0.05, \Delta f = 0.2$) determined using holdout validation (maximizing accuracy on the holdout set). Finally, the distribution $P(S)$ was measured using normalized counts.

Table 4 shows the obstruent recognition accuracy resulting from the explicit time-mark model. We observe the

Table 4
Obstruent phone recognition performance for the explicit time-mark model.

C	δ	Accuracy	% Corr	% Ins	% Del	% Sub
1	0.0	51.7	63.0	11.3	5.2	31.8
2	0.0	57.8	66.5	8.6	5.0	28.5
4	0.0	60.4	68.4	8.0	5.0	26.6
8	0.0	61.4	69.3	7.9	5.3	25.4

expected increase in system accuracy as the detector set improves with increasing numbers of GMM components. This improvement results from a simultaneous decrease in both insertion and substitution errors. However, we observe a fairly stable deletion rate, indicating the importance of the region duration T in the probabilistic model. That is, the dependence on supersegment duration can give precedence to longer sequences in the face of missed detections, reducing deletions errors in favor of a mixture of additional correct phones and substitution errors.

One major drawback to this approach is the substantial training data required to accurately estimate the 385×48 distributions of the form $P(t^p, f^p|S)$, which is especially troublesome for the rare sequences. Interestingly, we found that using no threshold ($\delta = 0$) led to optimal performance in all cases, a setting produces a point process representation that contains a large abundance of low probability landmarks. We believe such low probability landmarks in the distribution estimation procedure bulks up the statistics for rare sequences, alleviating training data shortfalls and resulting in overall performance gains. For this reason, our intuition suggests that the optimal threshold would increase as we provide more training data or use distribution estimation techniques better suited to small sample sizes. Such investigation lies outside the scope of this paper.

3.3.4. Poisson process model results

The Poisson process model requires the evaluation of Eqs. (7) and (20) over the 385 possible obstruent phone sequences. We again used uniform kernel density estimation of the distributions $P(T|S)$ (optimal bandwidth $\Delta T = 0.05$) and determined $P(S)$ using normalized counts. To estimate $P(R|S)$, we must compute the family of rate parameters required by the model assumption. In the most general case (inhomogeneous, marked), we can completely define the model architecture by selecting the number of time and mark interval divisions (D and K, respectively), as well as the optimal detector threshold.

Table 5 shows the obstruent recognition accuracy for an inhomogeneous unmarked Poisson process model. We divide the time interval into three homogeneous regions to roughly correspond with the typical maximum obstruent sequence length of three phones[8] (in the model presentation above, this corresponds to $D = 3$). With this model architecture, we found the optimal threshold to be $\delta = 0.5$. This is also an intuitive choice, as it corresponds to an optimal Bayes binary classification for each landmark (i.e., is the phone more likely present than not). We find that the performance gain from increasing detector reliability arises from a decrease in substitution errors, while the insertion and deletion rates remains roughly constant. We believe the low insertion rate across the board is primarily a result of the threshold imposed. As in the explicit time-mark model results, the stable deletion rate is maintained by the explicit modelling of supersegment duration T.

Table 5
Obstruent phone recognition performance for the inhomogeneous unmarked Poisson process model.

C	δ	Accuracy	% Corr	% Ins	% Del	% Sub
1	0.5	56.6	61.6	5.0	5.6	32.8
2	0.5	60.3	65.6	5.4	5.2	29.2
4	0.5	62.5	67.6	5.1	5.0	27.4
8	0.5	63.2	68.7	5.5	5.2	26.2

As might be expected, a homogeneous architecture (i.e. $D = 1$) led to poor performance, both for marked and unmarked representations. More surprisingly, we found that including marks in the inhomogeneous model architecture led to a consistent decrease in accuracy as we increased the number of mark divisions (i.e. $K > 1$). This may point to the validity of the optimal Bayes classification threshold or may simply be a consequence of limited training data. Due to the inferior performance, we omit the listing for these model configurations.

3.3.5. Baseline HMM results

Finally, to provide a reference point[9] from the mainstream speech recognition community, we implemented the vanilla HMM baseline defined by Sha and Saul (2007) (i.e., the maximum likelihood variant in their study). Not coincidentally, our front end prescription (see Section 3.1) is identical to Sha and Saul's. This means the Gaussian mixture distributions $P(x|p)$ used as the baseline HMM's emit probabilities are the same used to construct our point process representation. Therefore, comparison of their system and ours functions to isolate the adequacy of our point process representation and models relative to a basic HMM approach. Note that the state space definition of this baseline system matches that used in the HMM-PP method of Section 2.2.

Our implementation of this HMM baseline matched the full phonetic recognition performance published by Sha and Saul. As done for the HMM-PP approach, we performed a global Viterbi decode and used the TIMIT phonetic transcription to retrieve the predictions in the obstruent regions (again, any non-obstruent phones predicted were thrown away before scoring). The corresponding obstruent region recognition performance is listed in Table 6. We observe the usual improvement in recognition accuracy as we increase the number of mixture components, but with stable insertion and deletion rates.

[8] In the TIMIT database, the 378 of the 385 possible obstruent phone sequences have length less than or equal to 3 (not including closure silences).

[9] The chosen baseline system is by no means the state-of-the-art in TIMIT phonetic recognition. The best results achieved for this task are provided in Glass (2003).

Table 6
Obstruent phone recognition performance for a baseline HMM.

C	Accuracy	% Corr	% Ins	% Del	% Sub
1	51.1	63.6	12.6	7.8	28.6
2	57.5	68.9	11.4	6.5	24.6
4	61.3	72.1	10.8	6.0	21.9
8	63.3	74.1	10.8	5.9	20.0

3.4. Discussion

Table 7 summarizes the best obstruent recognition accuracy obtained from each of the methods presented in this paper. Several trends emerge from this comparison table:

(1) All probabilistic point process models perform significantly better than the naive method. While this may not be a surprising fact, the nearly 20 point margins demonstrate how noisy the detector set is and how effective each probabilistic model is at cleaning up false positives. To illustrate this fact further, we can consider the naive performance when setting the threshold to result in similar correctness levels as the probabilistic models. If, for example, we threshold the $C = 8$ detector set to produce a comparable 70% correctness rate, the naive method produces a dismal 23% accuracy. Furthermore, if we apply the Poisson process threshold of 0.5, we observe an insertion rate of 149%.

(2) The inhomogeneous unmarked Poisson process model outperforms the explicit time-mark model for all detector set reliabilities. This represents significant progress relative to our previous work (Jansen and Niyogi, 2008), which employed a variant of ETMM. The Poisson process model has lower complexity (in terms of the number of parameters) and is thus better estimated with limited training data. Also, we believe the Poisson process model is better suited to a unreliable detector set, as it factors in inactivity of detectors that had fired in the training data for a candidate generating sequence. The explicit model, on the other hand, directly evaluates the active detectors only, so a missed detection is not penalized in computing the overall probability of the candidate generating sequence. This provides an explanation for the optimal zero threshold for ETMM: low probability landmarks allow otherwise inactive detectors to have a say.

(3) The inhomogeneous unmarked Poisson process model either outperforms or is comparable to all other methods for $C = 8$ (most reliable phone detectors) and outperforms all other methods for lower values of C (less reliable phone detectors). More surprisingly, this Poisson process model, operating only on the sparse point process representation, is comparable to or outperforms the standard HMM baseline using the complete vector time series representation. As detector reliabilities decrease, the Poisson process model exhibits significantly improved robustness. We again believe this to be a consequence of appropriate built-in penalties for detector inactivity.

(4) The HMM-PP method accuracy matched the HMM baseline for all detector reliabilities. This somewhat surprising fact illustrates the sufficiency of the sparse point process representation for phonetic decoding of obstruent regions. It is important to note that while HMM-PP performance is marginally better than that of the Poisson process model at $C = 8$, the standard HMM used in the HMM-PP method requires a vector time series representation. In the context of this paper, this does not pose a problem, as we construct the point process representation from a vector time series, and thus a synchronous clock rate is automatically provided. However, the ultimate utility of a point process representation for speech will arise when we construct a linguistically or neurobiologically motivated *asynchronous* front end. Note that while multiple stream frame-based methods can accommodate varying levels of asynchrony with varying levels of success (see Nock and Ostendorf, 2003), the viability of ETMM and Poisson process models is invariant to the level of asynchrony.

To illustrate this point, we performed an experiment where the stop consonant detectors were constructed with a MFCC front-end, but sampled every 7.5 ms as opposed to the 10 ms step size used for the other detectors. In this case, the Poisson process model resulted in the same performance. However, this small degree of asynchrony precluded application of the HMM-PP method as presented, i.e., without the introduction some means of synchronization (e.g. interpolation or synchronization state, as in (Bourlard et al., 1996)). We leave for future work an experimental comparison

Table 7
Best obstruent phone recognition accuracies for each method.

C	Naive	HMM-PP	ETMM	Poisson	HMM
1	34.0	51.1	51.7	56.6	51.1
2	38.4	58.1	57.9	60.3	57.5
4	41.4	61.6	60.4	62.5	61.3
8	44.4	63.6	61.4	63.2	63.3

Table 8
Sonorant consonant phone recognition accuracy for both the inhomogeneous Poisson process model and HMM baseline.

C	Poisson	HMM
1	72.4	60.5
2	75.2	64.3
4	76.2	67.8
8	78.0	70.7

between point process methods and multiple stream frame-based methods operating on linguistically-motivated asynchronous representations.

Finally, to provide an idea of how this approach might fare on full phonetic recognition, we extended our approach to the relatively easier task of decoding sonorant intervowel regions. This task amounts to determining for each region the most likely of the 61 possible sonorant consonant phone sequences that occur in the TIMIT database, given the observed point process representation. Table 8 shows the sonorant consonant recognition accuracy for the HMM baseline (again, extracted from the global Viterbi decode) and an inhomogeneous unmarked Poisson process model, where we have divided the time interval into four homogeneous regions. As we degrade detector reliability, the Poisson insertion and deletion rates remained roughly constant. This stability is maintained by the explicit modelling of supersegment duration T. The drops in accuracy of the Poisson method with decreasing values of C are thus primarily due to increased substitution rates. The vast majority of such substitutions are between phones of the same broad class.

It is important to note that in the case of sonorant intervowel regions only, explicit modelling of the supersegment duration provided a significant advantage to our supersegment-level Poisson method over the HMM baseline. However, such durational modelling rests heavily on accurate vowel–sonorant consonant boundaries, which are not trivial to automatically determine. Thus, the sonorant intervowel performance comparison should be taken with a grain of salt. To investigate this matter further, we used the baseline Sha and Saul HMM-based recognizer to automatically determine (i) a segmentation into obstruent, sonorant intervowel, and vowel regions; and (ii) the phonetic identities of the vowels. We then used our Poisson model to decode the obstruent and sonorant intervowel regions as described above. Note that this particular strategy discards the phone transition probabilities that straddle the sonorant–obstruent–vowel segmentation boundaries (the HMM baseline uses these probabilities to constrain recognition). Still, we achieved a full phonetic recognition accuracy that fell only one point short of the original HMM baseline performance. While using an HMM-based recognizer in this role is not the intent of our larger speech recognition framework, this result exhibits the merit of a sonority-based segmentation strategy.

4. Conclusions and future work

We have presented several statistical speech recognition models applicable to a landmark-based point process representation of speech. From our experiments in obstruent phone recognition, we have found that these methods are capable of recovering the underlying linguistic content from an exceedingly sparse set of landmarks with accuracy comparable to a basic HMM operating on the complete frame-based representation. We find the most promising and robust approach to be a standard inhomogeneous Poisson process model.

There are several directions for further research that follow naturally from the findings presented in this paper:

(1) Ultimately, we would like to extend this detector-based approach to standard recognition tasks. Possibilities include keyword spotting and small vocabulary recognition, achievable by building a point process model for each word of interest (in much the same way we build a model for each obstruent phone sequence). To build a large vocabulary recognition engine, we may extend our previously developed framework (see Jansen and Niyogi, 2008) to full phonetic recognition by integrating the findings presented here. Preliminary experiments in these directions have been promising.

(2) In this paper, we constructed our point process representation by piggybacking off a standard MFCC and GMM frame-based front end. While this choice facilitated performance comparison with the HMM baseline, it is not necessarily the most scientifically plausible. A complete exploration of point process representation construction strategies remains, an endeavor for which significant progress has already been made (see Stevens and Blumstein, 1981; Stevens, 2002; Niyogi and Sondhi, 2002; Pruthi and Espy-Wilson, 2004; Amit et al., 2005; Xie and Niyogi, 2006). The ideal point process representation will require a linguistically and/or neurobiologically motivated design to maximize the benefits of applying coding models proposed by the cognitive neuroscience community.

(3) We have only scratched the surface of the set of possible statistical models applicable to a point process representation of speech. In particular, implementing and testing models designed to work on limited training examples will prove vital to creating robust landmark-based recognition systems with human-comparable performance. For example, the Poisson process model may be improved with more sophisticated rate parameter (intensity) estimation techniques, such as kernel smoothing or parametric modelling (see Willett (2007) for an example in a different context). Additional models arising from the computational neuroscience community may also be considered (see Legenstein et al. (2005) and Gütig and Sompolinksy (2006) for examples).

(4) Further interface of the automatic speech recognition (ASR) community with cognitive neuroscience researchers may prove fruitful. The results presented in this paper demonstrate that looking to research in those fields can lead to insights in the design and development of ASR systems. Moreover, evaluation of the efficacy of scientifically-motivated ASR strategies can also quantify the plausibility of current models of speech perception. For example, recent

statistical analysis of neuronal activity in the visual cortex of monkeys has suggested that a slowly varying inhomogeneous Poisson process model is not ideal (Amarasingham et al., 2006). Similar hypotheses for speech perception could be tested in the context of ASR by implementing them in the framework presented in this paper.

References

Amarasingham, A., Chen, T.-L., Geman, S., Harrison, M.T., Sheinberg, D.L., 2006. Spike count reliability and the Poisson hypothesis. J. Neurosci. 26 (3), 801–809.

Amit, Y., Koloydenko, A., Niyogi, P., 2005. Robust acoustic object detection. J. Acoust. Soc. Amer. 118 (4).

Bourlard, H., Dupont, S., Ris, C., 1996. Multi-stream Speech Recognition. Tech. Rep. IDIAP-RR 96-07, IDIAP.

Brown, E.N., 2005. Theory of point processes for neural systems. In: Chow, C.C., Gutkin, B., Hansel, D., Meunier, C., Dalibard, J. (Eds.), Methods and Models in Neurophysics. Elsevier, Paris, pp. 691–726 (Chapter 14).

Chi, Z., Wu, W., Haga, Z., 2007. Template-based spike pattern identification with linear convolution and dynamic time warping. J. Neurophysiol. 97 (2), 1221–1235.

Deng, L., Sun, D.X., 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. J. Acoust. Soc. Amer. 95 (5), 2702–2719.

Ellis, D.P.W., 2005. PLP and RASTA (and MFCC, and inversion) in Matlab. (online web resource). URL <http://www.ee.columbia.edu/dpwe/resources/matlab/rastamat/>.

Esser, K.-H., Condon, C.J., Suga, N., Kanwal, J.S., 1997. Syntax processing by auditory cortical neurons in the FM–FM area of the mustached bat *Pteronotus parnellii*. Proc. Natl. Acad. Sci. USA 94, 14019–14024.

Frangoulis, E., 1989. Vector quantisation of the continuous distributions of an HMM speech recogniser based on mixtures of continuous distributions. In: Proc. of ICASSP, pp. 9–12.

Fuzessery, Z.M., Feng, A.S., 1983. Mating call selectivity in the thalamus and midbrain of the leopard frog (Rana p. pipiens): single and multiunit responses. J. Comparitive Psychol. 150, 333–334.

Geiger, D., Liu, T.-L., Donahue, M.J., 1999. Sparse representations for image decompositions. Internat. J. Comput. Vision 33 (2), 139–156.

Glass, J.R., 2003. A probabilistic framework for segment-based speech recognition. Computer Speech Lang. 17, 137–152.

Greenberg, S., Carvey, H., Hitchcock, L., Chang, S., 2003. Temporal properties of spontaneous speech – a syllable-centric perspective. J. Phonetics 31 (3), 465–485.

Gütig, R., Sompolinksy, H., 2006. The tempotron: a neuron that learns spike timing-based decisions. Nature Neurosci. 9 (3), 420–428.

Hasegawa-Johnson, M., 2002. Finding the best acoustic measurements for landmark-based speech recognition. Accumu: J. Arts Technol. Kyoto Computer Gakuin.

Jansen, A., Niyogi, P., 2008. Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition. J. Acoust. Soc. Amer. 124 (3), 1739–1758.

Juang, B.H., Rabiner, L.R., Levinson, S.E., Sondhi, M.M., 1985. Recent developments in the application of hidden Markov models to speaker independent isolated word recognition. In: Proc. ICASSP.

Lee, C.-H., 2007. An overview on automatic speech attribute transcription (asat). In: Proc. Interspeech, pp. 1825–1828.

Lee, K.-F., Hon, H.-W., 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust. Speech Signal Process. 37 (11), 1641–1648.

Legenstein, R., Nger, C., Maass, W., 2005. What can a neuron learn with spike-timing-dependent plasticity? Neural Comput. 17 (1), 2337–2382.

Levinson, S.E., 1986. Continuously variable duration hidden Markov models for automatic speech recognition. Computer Speech Lang. 1 (2), 29–45.

Li, J., Lee, C.-H., 2005. On designing and evaluating speech event detectors. In: Proc. Interspeech, pp. 3365–3368.

Livescu, K., Glass, J., 2004. Feature-based pronunciation modeling for speech recognition. In: Proc. HLT/NAACL.

Ma, C., Taso, Y., Lee, C.-H., 2006. A study on detection based automatic speech recognition. In: Proc. Interspeech.

Mak, B., Tam, Y.-C., 2000. Asynchrony with trained transition probabilities improves performance in multi-band speech recognition. In: Proc. ICSLP, pp. 149–152.

Margoliash, D., Fortune, E.S., 1992. Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. J. Neurosci. 12, 4309–4326.

Niyogi, P., Sondhi, M.M., 2002. Detecting stop consonants in continuous speech. J. Acoust. Soc. Amer. 111 (2), 1063–1076.

Niyogi, P., Mitra, P., Sondhi, M.M., 1998. A detection framework for locating phonetic events. In: Proc. ICSLP.

Nock, H.J., Ostendorf, M., 2003. Parameter reduction schemes for loosely coupled HMMs. Computer Speech Lang. 17 (2–3), 233–262.

Olhausen, B.A., 2003. Learning sparse, overcomplete representations of time-varying natural images. In: Proc. ICIP, pp. 41–44.

Ostendorf, M., 1996. From HMMs to segment models: stochastic modelling for CSR. In: Lee, Chin-Hui, Soong, Frank K., Paliwal, Kuldip K. (Eds.), Automatic Speech and Speaker Recognition: Advanced Topics. Springer, pp. 185–209 (Chapter 8).

Ostendorf, M., Kannan, A., Kimball, O., Rohlicek, J., 1992. Continuous word recognition based on the stochastic segment model. In: Proc. DARPA Workshop on Continuous Speech Recognition.

Parker, S.G., 2002. Quantifying the Sonority Hierarchy. Ph.D. Thesis, University of Massachusetts-Amherst.

Poeppel, D., Idsardi, W.J., van Wassenhove, V., 2007. Speech perception at the interface of neurobiology and linguistics. Philosophical Transactions of the Royal Society of London B.

Pruthi, T., Espy-Wilson, C., 2004. Acoustic parameters for automatic detection of nasal manner. Speech Comm. 43, 225–239.

Ramesh, P., Wilpon, J.G., 1992. Modeling state durations in hidden Markov models for automatic speech recognition. In: Proc. ICASSP.

Russell, M.J., Cook, A.E., 1987. Experimental evaluation of duration modelling techniques for automatic speech recognition. In: Proc. ICASSP.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T., 2007. Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Machine Intell. 29 (3), 411–426.

Sha, F., Saul, L.K., 2007. Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models. In: Proc. ICASSP, pp. 313–316.

Stevens, K.N., 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Amer. 111 (4), 1872–1891.

Stevens, K.N., Blumstein, S.E., 1981. The search for invariant acoustic correlates of phonetic features. In: Eimas, P., Miller, J.L. (Eds.), Perspectives on the Study of Speech. Erlbaum, Hillsdale, NJ, pp. 1–38 (Chapter 1).

Stevens, K.N., Manuel, S.Y., Shattuck-Hufnagel, S., Liu, S., 1992. Implementation of a model for lexical access based on features. In: Proc. ICSLP.

Suga, N., 2006. Basic acoustic patterns and neural mechanisms shared by humans and animals for auditory perception. In: Greenberg, Steven, Ainsworth, William A. (Eds.), Listening to Speech: An Auditory Perspective. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 159–182.

Sun, J., Deng, L., 2002. An overlapping-feature-based phonological model incorporating linguistic constraints: applications to speech recognition. J. Acoust. Soc. Amer. 111 (2), 1086–1101.

Truccolo, W., Eden, U.T., Fellows, M.R., Donoghue, J.P., Brown, E.N., 2005. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. J. Neurophysiol. 93, 1074–1089.

Willett, R., 2007. Multiscale intensity estimation for marked Poisson processes. In: Proc. ICASSP, pp. 1249–1252.

Xie, Z., Niyogi, P., 2006. Robust acoustic-based syllable detection. In: Proc. ICSLP.

Zhang, Y., Diao, Q., Huang, S., Hu, W., Bartels, C., Bilmes, J., 2003. DBN based multi-stream models for speech. In: Proc. ICASSP, pp. 836–839.