Point Process Models for Event-Based Speech Recognition

A. Jansen[*] and P. Niyogi[†]

February 27, 2008

Abstract

Several strands of research in the fields of linguistics, speech perception, and neuroethology suggest that durational modelling of a 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.

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).

[*] Department of Computer Science

[†] Departments of Computer Science and Statistics, University of Chicago.

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 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 $\mathcal{F}$ 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

[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; Pruthi and Espy-Wilson, 2004; 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.

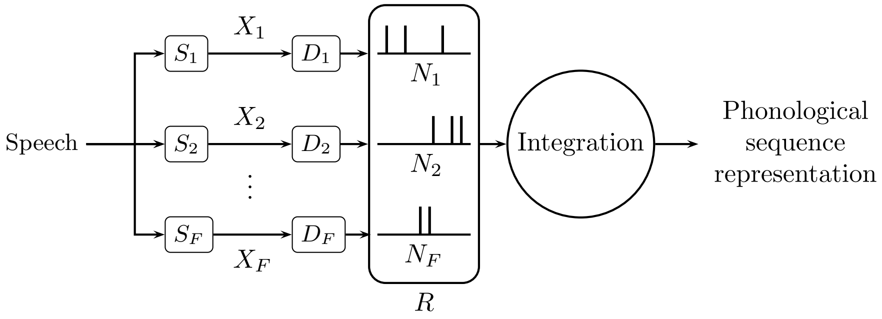


Figure 1: Architecture of our event-based framework. In general, we construct one signal processor S_i for each acoustic property of interest ($F = |\mathcal{F}|$), which produces a specialized representation X_i. Each representation is input to a detector D_i for the property, producing a point process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

over some alphabet $\mathcal{P}$, which may, for example, be the set of phones, broad classes, distinctive features, articulatory variables, or even syllables or words. Figure 1 shows a schematic of our architecture.

In this paper, we assume one has a detector for each phonological unit $p \in \mathcal{P}$ (i.e. $\mathcal{F} = \mathcal{P}$), each producing a point process $N_p = \{t_1, \ldots, t_{n_p}\}$, where each $t_i \in \mathcal{R}^+$. 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 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.

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 multi-scale 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, 2007). While decoding these constrained-length obstruent sequences may be viewed as a large multi-class classification 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 segment of speech given the point process representation defined above. The naive and hidden Markov model-based approaches can be applied globally to an entire utterance. However, for the explicit timemark and Poisson process models, we must first process the utterance into relatively short segments whose space of possible underlying sequences is limited by phonological constraints.

2.1 Naive Approach

The simplest method of converting a set of point processes $\{N_p\}$ 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 $p_i \in \mathcal{P}$ where $t_i^{p_i} < t_j^{p_j}$ for $i < j$, the sequence is determined by

$$S = p_1 p_2 \dots p_N.$$

The 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 Point Process Hidden Markov Model (PPHMM)

A hidden Markov model can not be directly applied to an asynchronous set of point processes $R = \{N_p, M_p\}_{p \in \mathcal{P}}$, where each $t \in N_p$ may be any real number for all $p \in \mathcal{P}$. However, if our point processes are synchronous (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 $V = \vec{v}_1\vec{v}_2 \ldots \vec{v}_T$ defined by

$$\vec{v}_l[j] = \begin{cases} f_k \in M_{p_j} & \text{if } \exists k \text{ s.t. } t_k \in N_{p_j} \text{ and } t_k = l\Delta t \\ 0 & \text{o/w} \end{cases} \tag{1}$$

We can then proceed to applying a continuous density HMM model to recover the hidden state sequence $S = s_1 s_2 \ldots s_T$ for $s_t \in \mathcal{P}$ by maximizing the joint probability over S and V. Under the Markov assumption, this term takes the form

$$\log P(V, S) = \sum_{t=1}^{T} [\log P(\vec{v}_t | s_t) + \log P(s_t | s_{t-1})]. \tag{2}$$

The transition probabilities $P(s_t|s_{t-1})$ can be determined by counting the frame-level transitions according to the transcription. For modelling the distributions $P(\vec{v}_t|s_t)$ over new input vector space, which tends to have sparse support, applying a 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 V is binary-valued, we model the emission densities using B-component multivariate binomial mixture models of the form

$$P(v_t | p_t = p) = \sum_{b=1}^{B} \omega_{pb} \mathcal{B}(\vec{q}_{pb})(\vec{v}_t). \tag{3}$$

where $\sum_{b=1}^{C} \omega_{pb} = 1 (\omega_{pb} > 0)$ for each $p \in \mathcal{P}$. Here, the function $\mathcal{B}(\vec{q}_{pb})$ is the b^{th} binomial mixture component in the context of phonological unit p, given by

$$\mathcal{B}(\vec{q}_{pb})(\vec{v}) = \prod_i \left(q_{pb}[i]\right)^{v[i]} \left(1 - q_{pb}[i]\right)^{(1-v[i])}, \tag{4}$$

where $q_{pb}i \in [0, 1]$ is the b^{th} component probability of a detection in the i^{th} component in the context of p. 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 Δv for all coordinates. Assuming the coordinates are conditionally independent, we may write

$$P(v_t | p_t = p) = \prod_{j=1}^{|\mathcal{P}|} H_{pj}(\vec{v}_t[j]), \tag{5}$$

where H_{pj} is the histogram estimate of the distribution of the j^{th} coordinate in the context of phonological unit p.

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., $\vec{v}_t = \vec{0}$) at times when no landmarks occurred. The emission probability distributions estimated for each state will each yield a constant value $K_p = P(\vec{0}|p)$ when the zero vector is encountered. If we set aside the transition probabilities for a moment, it follows that for all t such that $\vec{v}_t = \vec{0}$, the optimal state is always $p_t = \arg\max_p K_p$, 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 $\mathcal{P}' \equiv \{\mathcal{P}, \epsilon\}$, 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

Consider a maximum *a posteriori* (MAP) estimate of the phonetic sequence S given the observed point process representation $R = \{N_p, M_p\}_{p \in \mathcal{P}}$ and the duration of the segment $T = T_2 - T_1$, given by

$$S^* = \arg\max_{S \in \mathcal{P}^*} \log P(S|R,T) = \arg\max_{S \in \mathcal{P}^*} P(R|S)P(T|S)P(S), \tag{6}$$

where we have assumed conditional independence between the point process and the segment duration. We would like to deal with the term $P(R|S)$ by explicitly modelling the times and strengths of landmarks observed. Since all landmarks within a given segment lie in the interval $[T_1, T_2]$, we begin by normalizing the segment 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

$$P(R|S) = \prod_{p \in \mathcal{P}} \prod_{i=1}^{n_p} P(t_i^p, f_i^p | S). \tag{7}$$

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

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

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

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

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

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

2.4 Poisson Process Model

The marked point process representation makes a Poisson process model a natural choice for the $P(R|S)$ term of Equation 6. 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 segment 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 processes $N_p = \{t_1, \ldots, t_{n_p}\}$, one for each $p \in \mathcal{P}$, contained in the interval $(0, T]$. If $N_p(t) \equiv |\{t_i | t_i \leq t\}|$ is the number of landmarks in the interval $(0, t]$, then for a homogeneous Poisson process, we may write

$$P_{a,b}(k) \equiv P[N_p(b) - N_p(a) = k] = \frac{(\lambda\tau)^k e^{-\lambda\tau}}{k!},$$

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

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

Since the process is memoryless, the probability of the whole point process becomes

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

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

$$P(R|S) \propto \prod_p [\lambda(p, S)]^{n_p} e^{-\lambda(p,S)T}, \qquad (8)$$

where $\lambda(p, S)$ depends both on the generating phonological sequence and the phonological unit of the point process 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 training segments containing the sequence S, and the total number K of landmarks of type p observed in those segments, the maximum likelihood estimate of $\lambda(p, S)$ is

$$\lambda^*(p, S) = \arg\max_\lambda K \log \lambda - \lambda NT = K/NT. \qquad (9)$$

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

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

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

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

where

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

It follows that the maximum likelihood estimation of the rate parameter of the d^{th} segment piece for phonological unit p and generating sequence S is given by

$$\lambda_d^*(p, S) = K_d D/NT, \qquad (10)$$

assuming we have been provided with N training segments containing a total of K_d landmarks in the d^{th} segment piece. Finally, the conditional probability of the whole representation given a generating sequence S can be computed as

$$P(R|S) \propto \prod_{p\in\mathcal{P}} \prod_{d=1}^{D} [\lambda_d(p,S)]^{n_{p,d}} e^{-\lambda_d(p,S)\Delta T} \tag{11}$$

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 parameter. In this case, the sole spatial dimension corresponds to the mark space $[0, 1]$, 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

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

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

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

where

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

The maximum likelihood estimation of the rate parameter of the m^{th} mark space division for phonological unit p and generating sequence S is given by

$$\lambda_m^*(p,S) = K_m/NT, \tag{12}$$

assuming we have been provided with N training segments of sequence S containing a total of K_m landmarks in the m^{th} mark space division. The conditional probability of the whole representation given a generating sequence S can be computed as

$$P(R|S) \propto \prod_{p\in\mathcal{P}} \prod_{m=1}^{M} [\lambda_m(p,S)]^{n_{p,m}} e^{-\lambda_m(p,S)T} \tag{13}$$

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

3 Experiments in Obstruent Segment Decoding

In this section, we consider the speech recognition subtask of decoding consonants in obstruent segments of the speech signal. 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, 2007) 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 a central role of the sonorant-obstruent distinction.

Given an obstruent segment (T_1, T_2) of duration $T = T_2 - T_1$ and the point process representation restricted to the segment, $R' = R|_{(T_1,T_2)}$, we would like to find the most likely sequence $S = p_1 \ldots p_n$, where $p_i \in \mathcal{O} =$ the set of obstruent phones and $\mathcal{P}$ is the set of all phones. This amounts to finding the S that maximizes $\log P(S, V)$ for the PPHMM method presented in Section 2.2 or that maximizes $P(S|R)$ for the explicit time-mark and Poisson process models of Sections 2.3 and 2.4, respectively. Given the linguistic constraints on the length of obstruent sequences, there are only 385 possible obstruent sequences in the TIMIT corpus[2] 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 using all test sentences. We defined our phonological unit set $\mathcal{P}$ to be the standard 48 phone set defined by Lee and Hon (1989) and used in later work by Sha and Saul (2007). The definition of this set in terms of TIMIT labels is shown in Table 3.

3.1 Constructing the Point Process Representation

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

1. Given W windows of the signal collected every Δ_ϕ seconds, construct for each $\phi \in \mathcal{F}$ 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] 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.

Table 1: The list of 48 phones used in our experiments and the corresponding TIMIT labels included for each (reproduced from Lee and Hon (1989)).

Phone	Example	Incl	Phone	Example	Incl
iy	b*ea*t		en	butt*on*	
ih	b*i*t		ng	si*ng*	eng
eh	b*e*t		ch	*ch*urch	
ae	b*a*t		jh	*j*udge	
ix	ros*e*s		dh	*th*ey	
ax	th*e*		b	*b*ob	
ah	b*u*tt		d	*d*ad	
uw	b*oo*t	ux	dx	bu*tt*er	
uh	b*oo*k		g	*g*ag	
ao	ab*ou*t		p	*p*op	
aa	c*o*t		t	*t*ot	
ey	b*ai*t		k	*k*ick	
ay	b*i*te		z	*z*oo	
oy	b*oy*		zh	mea*s*ure	
aw	b*ough*		v	*v*ery	
ow	b*oa*t		f	*f*ief	
l	*l*ed		th	*th*ief	
el	bott*le*		s	*s*is	
r	*r*ed		sh	*sh*oe	
y	*y*et		hh	*h*ay	hv
w	*w*et		cl (sil)	(unvoiced closure)	{p,t,k}cl
er	b*ir*d	axr	vcl (sil	(voiced closure)	{b,d,g}cl
m	*m*om	em	epi (sil)	(epenthetic closure)	epi
n	*n*on	nx	sil	(silence)	h#, pau

2. Construct a detector function $g_\phi : \mathbb{R}^{k_\phi} \rightarrow \mathbb{R}$ for each $\phi \in \mathcal{F}$ 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, we can compute the point process (N_ϕ, M_ϕ) for feature ϕ according to

$$
\begin{aligned}
N_\phi &= \{i\Delta_\phi | g_\phi(x_i) > \delta \text{ and } g_\phi(x_i) > g_\phi(x_{i\pm1})\} \\
M_\phi &= \{g_\phi(x_i) | i\Delta_\phi \in N_\phi\}.
\end{aligned}
$$

   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 $\mathcal{F}$ to be the set of phones $\mathcal{P}$ (i.e., there is a one-to-one correspondence between features $\phi \in \mathcal{F}$ and phones $p \in \mathcal{P}$). 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 regressor 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 an segment of phone $p_l \in \mathcal{P}$, we would like to compute a set of detector functions $g_p : \mathbb{R}^{39} \rightarrow [0, 1]$ 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 $p \in \mathcal{P}$, given by

$$
P(x|p) = \sum_{c=1}^{C} \omega_{pc} \mathcal{N}(\vec{\mu}_{pc}, \mathbf{\Sigma}_{pc})(x), \tag{14}
$$

where $\omega_{pc} > 0$ and $\sum_{c=1}^{C} \omega_{pc} = 1$ for each $p \in \mathcal{P}$; and $\mathcal{N}(\vec{\mu}, \mathbf{\Sigma})$ is a normal distribution with mean $\vec{\mu}$ and full covariance matrix $\mathbf{\Sigma}$. 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

$$
g_p(x) = P(p|x) = \frac{P(x|p)P(p)}{\sum_{p \in \mathcal{P}} P(x|p)P(p)}, \tag{15}
$$

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 performance for various detector reliabilities.

Figure 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 segments for independent decoding. With the transcription-provided truth and model prediction 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.

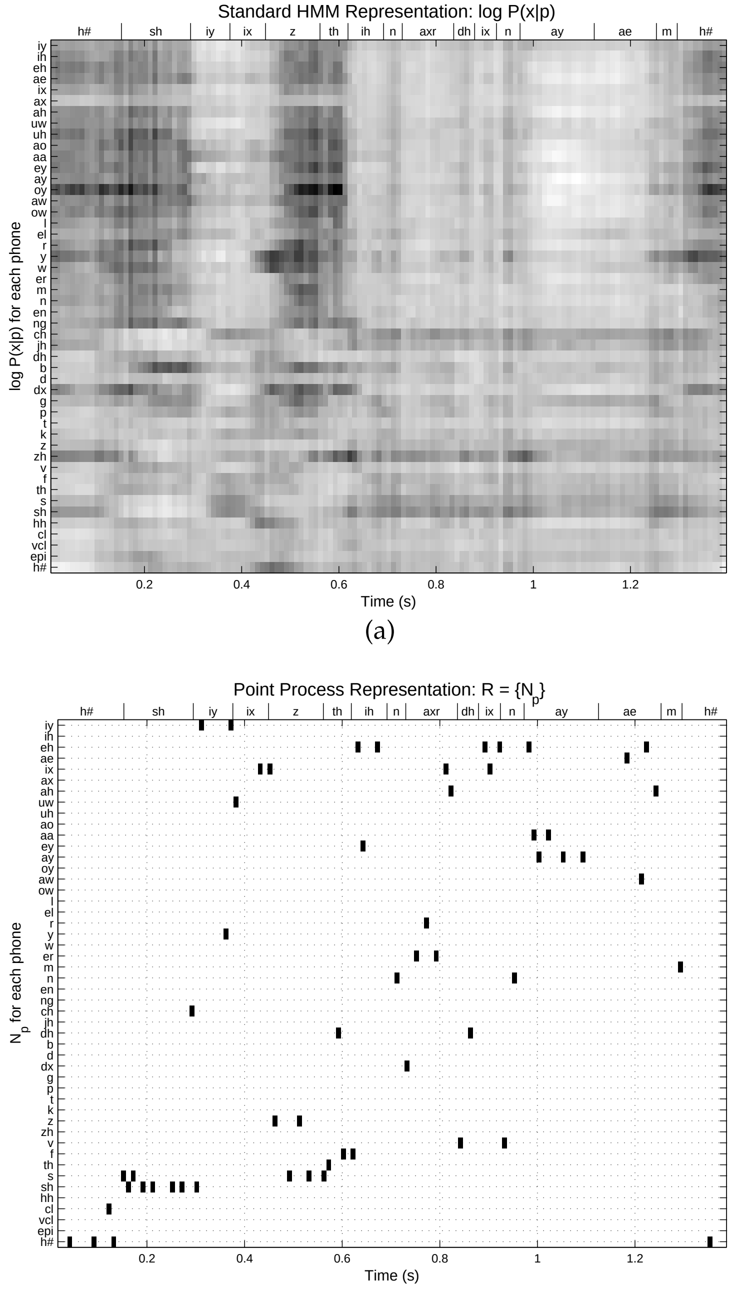
We proceeded by scoring the predicted sequences using minimum string edit distance alignment with the truth sequence in each obstruent segment. This results not only in a measurement of the recognition accuracy/error rates, but also a breakdown of the errors into insertion, deletion, and substitution 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 the point process representation produced by obstruent phone detectors (i.e., $\{N_p, M_p\}_{p \in \mathcal{O}}$). 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 2 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



(a)

(b)

Figure 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, $R = \{N_p\}_{p\in\mathcal{P}}$ for $\delta = 0.5$.

Table 2: 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

Table 3: 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

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.

3.3.2 Point Process HMM Results

To apply HMM methods, we constructed vector time series from the point process representations as described in Section 2.2. We attempted to model the sparse marked point process vector time series data with a Gaussian mixture model (the standard density model for HMM systems), which resulted in performance below the naive baseline. We 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 3 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 4 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 $\Delta v = 0.05$ 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 4: 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

Table 5: 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

3.3.3 Explicit Time-Mark Model Results

For the explicit time-mark model, we solved the optimization problem of Equations 6 and 7 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 5 shows the obstruent recognition accuracy resulting from the explicit time-mark model. We observe the 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 segment duration T in the probabilistic model. That is, the dependence on segment 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

Table 6: 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

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 Equation 6 and 11 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 6 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[3] (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 segment duration T.

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] 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).

Table 7: 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

3.3.5 Baseline HMM Results

Finally, to provide a reference point 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 distributions $P(x|p)$ used as their emit probabilities are equivalent to those used to construct our point process representation. Therefore, comparison of their system and ours functions isolate the adequacy of our point process representation and models relative to a basic HMM approach.

Our implementation of this HMM baseline matched the full phonetic recognition performance published by Sha and Saul. The corresponding obstruent segment recognition performance is listed in Table 7. We observe the usual improvement in recognition accuracy as we increase the number of mixture components, but with stable insertion and deletion rates.

3.4 Discussion

Table 8 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, 2007), 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 is the best overall approach studied in this paper, statistically equivalent or outperforming the other methods at all detector reliabilities. More surprisingly, this Poisson process model, operating only on the sparse point process representation, matches or outperforms the standard HMM 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 point process HMM method accuracy is statistically equivalent to an HMM 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 PPHMM is statistically to the Poisson process model at $C = 8$, any HMM-based 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.

   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 point process HMM method, at least without interpolation.

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

Table 8: Best obstruent phone recognition accuracies for each method.

C	Naive	PPHMM	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

on a 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. One possibility is keyword spotting or 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, 2007) 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; 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 auditory 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

Asohan Amarasingham, Ting-Li Chen, Stuart Geman, Matthew T. Harrison, and David L. Sheinberg. Spike count reliability and the Poisson hypothesis. *Journal of Neuroscience*, 26(3):801–809, 2006.

Yali Amit, Alexey Koloydenko, and Partha Niyogi. Robust acoustic object detection. *J. Acoust. Soc. Am*, 118(4), 2005.

Emery N. Brown. Theory of point processes for neural systems. In *Methods and Models in Neurophysics (Chow CC, Gutkin B, Hansel D, Meunier C, Dalibard J)*, chapter 14, pages 691–726. Elsevier, Paris, 2005.

Zhiyi Chi, Wei Wu, and Zach Haga. Template-based spike pattern identification with linear convolution and dynamic time warping. *J. Neurophysiology*, 97(2):1221–1235, 2007.

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

Karl-Heinz Esser, Curtis J. Condon, Nobuo Suga, and Jagmeet S. Kanwal. 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, 1997.

Zoltan M. Fuzessery and Albert S. Feng. Mating call selectivity in the thalamus and midbrain of the leopard frog (Rana p. pipiens): single and multiunit responses. *Journal of Comparitive Psychology*, 150:333–334, 1983.

Davi Geiger, Tyng-Luh Liu, and Michael J. Donahue. Sparse representations for image decompositions. *Int. J. Comput. Vision*, 33(2):139–156, 1999.

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

Robert Gütig and Haim Sompolinksy. The tempotron: A neuron that learns spike timing-based decisions. *Nature Neuroscience*, 9(3):420–428, 2006.

Aren Jansen and Partha Niyogi. A probabilistic speech recognition framework based on the temporal dynamics of distinctive feature landmark detectors. Technical Report TR-2007-07, U. of Chicago, Computer Science Dept., 2007.

Kai-Fu Lee and Hsiao-Wuen Hon. Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11):1641–1648, 1989.

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

Daniel Margoliash and Eric S. Fortune. Temporal and harmonic combination-sensitive neurons in the zebra finch's HVc. *Journal of Neuroscience*, 12:4309–4326, 1992.

Partha Niyogi and M. M. Sondhi. Detecting stop consonants in continuous speech. *J. Acoust. Soc. Am*, 111(2):1063–1076, 2002.

Bruno A. Olhausen. Learning sparse, overcomplete representations of time-varying natural images. In *Proc. of ICIP*, 2003.

Stephen G. Parker. *Quantifying the Sonority Hierarchy*. PhD thesis, University of Massachusetts-Amherst, 2002.

David Poeppel, William J. Idsardi, and Virginie van Wassenhove. Speech perception at the interface of neurobiology and linguistics. *Philosophical Transactions of the Royal Society of London B*, 2007.

Tarun Pruthi and Carol Espy-Wilson. Acoustic parameters for automatic detection of nasal manner. *Speech Communication*, 43:225–239, 2004.

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, 2007.

Fei Sha and Lawrence K. Saul. Comparison of large margin training to other discriminative methods for phonetic recognition by hidden Markov models. In *Proc. of ICASSP*, 2007.

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

Kenneth N. Stevens and Sheila E. Blumstein. The search for invariant acoustic correlates of phonetic features. In *Perspectives on the Study of Speech (P. Eimas and J. L. Miller)*, chapter 1, pages 1–38. Erlbaum, Hillsdale, NJ, 1981.

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

Wilson Truccolo, Uri T. Eden, Matthew R. Fellows, John P. Donoghue, and Emery N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiology*, 93:1074–1089, 2005.

Rebecca Willett. Multiscale intensity estimation for marked Poisson processes. In *Proc. of ICASSP*, 2007.

Zhimin Xie and Partha Niyogi. Robust acoustic-based syllable detection. In *Proc. of ICSLP*, 2006.