A HIERARCHICAL POINT PROCESS MODEL FOR SPEECH RECOGNITION

Aren Jansen and Partha Niyogi

University of Chicago
Department of Computer Science
Chicago, IL USA

ABSTRACT

In this paper, we present a computational framework to engage distinctive feature-based theories of speech perception. Our approach involves: (i) transforming the signal into a collection of *marked point processes*, each consisting of distinctive feature landmarks determined by statistical learning methods, and (ii) using the *temporal statistics* of this sparse representation to probabilistically decode the underlying phonological sequence. In order to assess the viability of this approach, we benchmark our performance on broad class recognition against a range of HMM-based approaches using the CMU Sphinx 3 system. We find our system to be competitive with this baseline and conclude by outlining various avenues for future development of our methodology.

Index Terms— speech recognition, speech processing

1. INTRODUCTION

We are interested in the fundamental task of *pure speech recognition*—the ability of humans to interpret speech in terms of a sequence of phonological units without invoking any higher level (syntactic, semantic, or pragmatic) linguistic knowledge. Toward this end, we present a computational model for speech recognition that is inspired by several interrelated strands of research in phonology, acoustic phonetics, speech perception, and neuroscience. Our primary goal is to provide a computational platform to engage, quantify, and test various theories in these sciences, with the hope of gaining insights that may have eventual technological consequences.

In our system, distinctive features are the atomic units, as opposed to the phone-based representations (e.g. triphones) used in most modern speech recognition systems. The various acoustic correlates of distinctive features operate at different scales in time and frequency. Consequently, rather than having a "one size fits all" representation that is common in most systems, we select multiple representations tuned for different distinctions. These representations are processed by distinctive feature detectors that are designed to fire at important events or landmarks. These detectors each result in a sparse, point process representation of the speech signal.

The decoding of the signal proceeds by integrating the firing of the individual feature detectors in a hierarchical way. At the root of the hierarchy is the sonorant-obstruent feature that is the most basic and perceptually salient distinction among speech sounds. Vowels correspond to peaks of the sonority profile and provide anchor points that define syllable-sized analysis units. Probabilistic integration of detector outputs occurs at such syllabic time scales on the rationale that this is the smallest perceptually robust unit. Thus, the information content of the signal within each analysis unit is coded in the temporal statistics of the point process representation.

This foundation leads us to a system that performs reasonably compared to a vanilla HMM baseline and is distinct from any other built so far, though it shares many qualities with those inspired by acoustic phonetics, distinctive features, and event landmarks [1][2]. The practical benefits to our approach include: (i) The simplicity of our modular design may aid diagnostics and portability to new languages and environments; (ii) The hierarchical approach leads to fewer parameters than HMMs, allows reuse of training data for different distinctions, and allows efficient training with limited amounts of transcribed data; (iii) The system design with its specialized detectors and temporal coding provides a new way to characterize the statistics of speech signals and reason about issues of invariance and robustness.

2. SYSTEM ARCHITECTURE

2.1. Distinctive Feature Representation

The theory of distinctive features asserts that phonemes are not the primitive building blocks of language; rather, each phoneme is a complex of binary features that each distinguish natural classes of phonemes sharing some common characteristic. While feature systems are rooted in phonology, they have natural articulatory interpretations and corresponding acoustic and perceptual correlates, providing a useful computational starting point. Moreover, work by Goldsmith [3] and others suggest that features have a hierarchical internal organization. This structure implies nodes higher up in the tree correspond to features that are somehow more basic or fundamental and whose acoustic correlates are less context dependent. Furthermore, features contained in separate branches of the hierarchy are independent, allowing motivated context dependent processing. For our purposes, we consider the hierarchy shown in the plane of Fig. 1 involving the distinctive features *sonorant* [son], *consonantal* [cons], *continuant* [cont], and *nasal* [nasal].

Consequently, our entry point into the interpretation of the signal is a segmentation into sonorant and obstruent regions. Computationally, this segmentation may be accomplished using any available machine learning method, though we choose support vector machines (SVM) using the radial basis function (RBF) kernel. The hinge-loss weight and RBF width parameters are chosen using holdout validation. We employ mel-frequency cepstral coefficients (MFCCs) spanning the full frequency range (0-8 kHz), computed in 10 ms windows every 5 ms. The 39 MFCCs include one energy and 12 cepstral coefficients, along with their delta and acceleration (double-delta) coefficients. Once the SVM is trained, to determine the segmentation we simply threshold the classifier output.

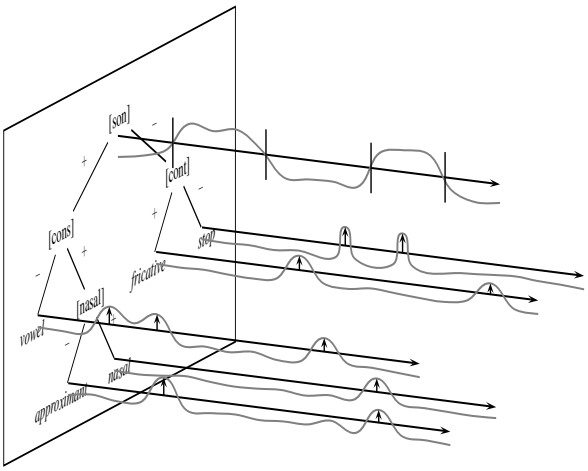

Fig. 1. Schematic diagram of the hierarchical timing tier representation. Landmarks are indicated by vertical arrows.

2.2. Feature Detectors for Subsequent Processing

The distinctive features corresponding to the lower tiers of the feature hierarchy make further distinctions into groups that coincide with the broad (manner) classes (vowels (V), approximants (A), nasals (N), fricatives (F), and stops (P)). For each, we build a detector that ideally should fire only in the presence of the feature. Such detectors may be analogized to neurons found in animals that fire selectively when certain complex acoustic attributes are present in the input stimulus [4][5].

There are three key ideas involved in the construction of a feature detector. First, for each feature of interest, a specialized acoustic representation is constructed in which that feature best expresses itself. Second, in this representational space, a classifier is built that is able to separate positive examples from negative ones on a frame-level. Third, the output of the classifier is further processed to provide a sparse representation in time as a point process composed of maxima points of the classifier outputs. At these points the features are most strongly expressed; following the general philosophy of [1] and others, we refer to these points in time as landmarks.

We end up with the hierarchical timing tier signal representation shown in Fig. 1. A few aspects of this representation are worth noting: (i) we have access to both the times and strengths at landmarks, resulting in a marked point process for each tier; (ii) the representation is very sparse, in stark contrast to typical frame-based representations of speech used in both traditional and landmark-based systems; (iii) the information in the representation is now coded in the temporal dynamics of these spikes and the statistics of inter-spike times will be correlated with the durations between articulatory events and ultimately the durations of various linguistic segments.

Given an imperfect sonority segmenter, we train each broad class SVM with examples across the entire phoneme space. To allow maximal separability, our broad class SVMs also employ the RBF kernel. Since each classifier processes the signal independently, their construction can be specialized according to the individual broad class content. While we choose MFCC features for the silence, fricative, nasal, approximant and vowel SVMs, the window length, frame rate, and frequency ranges used for each vary. Furthermore, our stop classifier employs energy and Wiener entropy parameters shown to be successful in this setting [6]. Since integration is not carried out on a common frame-level, we attain complete modularity of the system components.

The output of each SVM is a real number for each frame of the signal. In general, after thresholding this series, we define the landmarks as the position of any local maximum of the SVM output and the landmark strength as the corresponding maximal values. The one exception made to this landmark picking strategy is for the vowel detector; since the vowel landmarks will not be probabilistically determined, degenerate detections within a single vowel will result in insertions. To address this complication, we adapt the recursive "convex-hull" approach presented in [7] to compute a time-dependent baseline. While the amplitude of the local maxima may be large, under this scheme, neighboring candidates compete with respect to a baseline computed in the local region. Therefore, small variations of the detector output that would otherwise result in degenerate landmarks are rejected with an appropriate choice of threshold on the dynamic baseline-subtracted series.

2.3. Sonority Segment Decoding

The challenge now is to map the timing tier representation into a linear sequence of phonological units. If our feature detectors worked perfectly, this task would be trivial: simply read off the output of the feature detectors to obtain the corresponding broad class sequence. However, in the face of non-zero error rates, we need to model the statistical distribution of the pattern of firings associated with each underlying sequence and choose the most likely sequence given that pattern. The sonority segmentation defines a series of obstruent and sonorant regions. The vowel landmarks further subdivide the sonorant regions into series of sonorant intervocalic regions (from now on, we refer to them simply as intervocalic regions). Given the feature hierarchy, we may model obstruent and intervocalic regions independently.

Therefore, we require a probabilistic strategy for decoding the contents of both obstruent and intervocalic regions. To accomplish this, we have developed an approach for sonority segment decoding (SSD) based on a maximum *a posteriori* (MAP) estimate of the broad class sequence contained in each obstruent and intervocalic region. Consider an interval of a speech signal (T_1, T_2). The interval duration $T = T_2 - T_1$, combined with the activity of N broad class detectors, defines a set of observables $O = \{T, O_{X_1}, \ldots, O_{X_N}\}$, where each O_{X_i} denotes the observables for the class X_i detector. These consist of L_{X_i} time-strength pairs (one per detection) which we denote

$$O_{X_i} = \{(t_1^{X_i}, f_1^{X_i}), \ldots, (t_{L_{X_i}}^{X_i}, f_{L_{X_i}}^{X_i})\}, \tag{1}$$

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

At this point we can immediately write down a MAP estimate of the segment broad class sequence, $B^{\text{opt}} = \arg\max_B P(B|O)$. However, in the context of our hierarchical landmark-based system, we would like our model to also estimate which landmarks within the region were correct and which were misfires. With this information, we could later proceed with transcription refinement at true landmarks only. To address this, we can define a set of indicator variables $H = \{H_{X_1}, \ldots, H_{X_N}\}$, where $H_{X_i} = \{h_1^{X_i}, \ldots, h_{L_{X_i}}^{X_i}\}$ and $h_k^{X_i} = 1$ if the k^{th} detection of class X_i is a true positive, and 0 otherwise. Applying Bayes' rule, the MAP estimate taking H into account becomes

$$(B^{\text{opt}}, H^{\text{opt}}) = \arg\max_{B,H} P(O|B,H)P(H|B)P(B). \quad (2)$$

Since our approach is to first partition the utterance down to short syllabic analysis units consisting of a limited number of phonemes, we can accomplish optimization by simply calculating the likelihood for all possibilities. If we attempted the same exhaustive approach for word or sentence-long reconstruction units, this combinatoric problem would become prohibitively cumbersome.

We can further simplify the general MAP estimation problem of Eq. 2 with several conditional independence assumptions: (i) the behavior of the broad class detectors are independent of each other and the segment duration; (ii) the detection correctness pattern for the broad class detectors are independent; (iii) detection times for a particular class are independent of each other; (iv) strengths of the detections for a particular class are independent both of each other and the broad class sequence encountered; and, (v) detector strengths and times are independent. While the extent of the validity of these assumptions has not been rigorously established, they greatly reduce the number of training sentences required to estimate the component distributions. Thus, for a set of broad classes $\mathcal{C}$, the optimization problem of Eq. 2 reduces to

$$(B^{\text{opt}},H^{\text{opt}}) = \arg\max_{B,H} P(T|B)P(B)\times \prod_{X\in\mathcal{C}} P(H_X|B)\prod_{i=1}^{L_X} P(t_i^X|B,h_i^X)P(f_i^X|h_i^X). \quad (3)$$

Now, given an obstruent region, as determined by the sonority segmentation, and the set of contained obstruent detector observables, we need only compute Eq. 3 over possible $B \in \{\text{sil}, \text{P}, \text{F}\}^*$, where $\mathcal{C} = \{\text{sil}, \text{P}, \text{F}\}$. Within each sonorant region, L vowel landmarks determine a series of $L+1$ intervocalic regions. For each such intervocalic region, T is defined as the time elapsed either between adjacent vowel landmarks, between a landmark and an adjacent sonorant region boundary, or the entire sonorant region if there are no vowel landmarks. We perform the intervocalic optimization of Eq. 3 over possible $B \in \{\text{A}, \text{N}\}^*$, where $\mathcal{C} = \{\text{A}, \text{N}\}$.

This probabilistic framework requires the measurement of all prior distributions involved in Eq. 3 for both obstruent and sonorant intervocalic regions. These distributions can be obtained by simply running the various system components on transcribed training data and maintaining a record of the resulting observables involved in each distribution in Eq. 3. We use the computationally straightforward histogram method to estimate discrete variable distributions; for the scalar variables f, t, and T, we use uniform kernel density estimation, which introduces three kernel bandwidth parameters.

3. EXPERIMENTAL RESULTS

3.1. Sonority Segmentation and Detector Performance

The support vector machine for the sonority segmenter was trained on 100 randomly chosen TIMIT sx/i training sentences. Using 100 randomly chosen TIMIT sx/i test sentences, we recorded a frame-level test error of 6.44%. Since the entire phoneme need not be present in a given region for successful decoding, lack of segmentation precision does not necessarily preclude success in later stages. We find that for 95.0% and 89.3% of the sonorant and obstruent phonemes, respectively, a majority of their duration fall in an appropriate sonority segment. Note that phonemes that have an even smaller fractional overlap with a correct segment still may be correctly decoded if the corresponding landmark is located there.

Table 1. Representation and training parameters for the landmark detectors.

Detector	$T_{\text{win}}/T_{\text{step}}$	F_{range}	E_{train}	E_{phn}
Vowel	40/20 ms	0-4 kHz	10.3%	15.1%
Approx.	20/20 ms	0-8 kHz	19.2%	28.2%
Nasal	30/15 ms	0-8 kHz	6.0%	10.0%
Fricative	30/15 ms	0-8 kHz	6.9%	11.4%
Stop	35/5 ms	N/A	6.2%	17.4%
Silence	20/10 ms	0-8 kHz	6.0%	7.8%

Creating each of the six landmark detectors required the construction of a support vector machine trained to recognize frames of the target class. We work with a set of 100 randomly chosen TIMIT sx/i training sentences. For the vowel, approximant, nasal, fricative and silence detectors, we again use MFCCs, but the window size (T_{win}), step size (T_{step}), and frequency range (F_{range}) parameters vary according to Table 1. For the stop detector, we used the acoustic parameter prescription of [6] as an alternative to the MFCC representation, though we modify the frame rate to reduce computational costs. For training, all frames centered within the desired phoneme boundaries are considered positive examples except in the case of the stop detector, where only the closure-burst transition is considered a positive example. The frame-level training errors (E_{train}) and phoneme-level test error at the precision-recall break-even point (E_{phn}) are also listed in Table 1.

There are three additional points to note regarding detector performance: (i) a significant majority of errors are made between broad classes of the same sonority superclass, validating our choice of the sonorant feature as an appropriate initial point of speech segmentation; (ii) the vowel detector results in the lowest degenerate landmark rate (21 degenerate vowel landmarks in 1210 vowel phonemes), a direct result of the dynamic baseline algorithm; (iii) the weakest link by far is the approximant detector, which will be a vital point of future research.

3.2. Segment Decoding Performance

To separate the performance of the SSD model from that of the sonority segmenter and vowel detector, we conducted experiments using the actual sonority segmentations and vowel center points provided by the TIMIT transcription for both training and testing. Prior distribution data was collected from 1000 randomly chosen TIMIT sx/i training sentences. We evaluated the predictions relative to the actual sequences present using minimum string edit distance alignment. We determined optimal accuracy bin width parameters using an additional 100 training sentences.

We tested on all 1344 sx/i sentences contained in the TIMIT test set. There are 42 possible broad class sequences that may lie in any obstruent region and 12 possible sequences in any intervocalic region. Table 2 shows the transcription performance for several variations of the decode procedure. The first is a naive measure of baseline performance, where we simply chronologically sort the landmarks above the appropriate operating threshold in each obstruent region. The predicted sequence is simply the corresponding broad classes of these landmarks. The second method is the standard implementation of probabilistic decoding outlined in this paper. Finally, the two "Rank $\leq N$" methods assume we have access to an oracle that identifies the true obstruent or intervocalic region

sequence if it is one of the N most probable sequences. In all of the obstruent region decoding variations, we ignore silence landmarks in the performance analysis since their presence is not necessary in the final transcription.

The poor naive baseline performance illustrates the main problem with integrating multiple error prone detectors: correctness rates average, but insertion rates add together. However, for obstruent regions, SSD greatly cleans up false detections, admitting an insertion rate of only 6% while maintaining the correctness rate of the baseline. We find significantly lower intervocalic performance of both the naive baseline and the standard decode relative to obstruent region decoding. This is largely due to the exceptionally poor isolated performance of the approximant detector. Still, SSD more than doubles intervocalic accuracy over the baseline. In both region types, the ranking methods result in striking performance gains, portending great improvements in this module of the system when higher-level linguistic constraints are imposed.

3.3. Overall Performance

We now turn to the overall performance of our landmark-based broad class recognizer, implementing the sonority segmenter, landmark detectors, and probabilistic segment decoding. We tested our system and four continuous CMU Sphinx-3 HMM variations, using both context independent (CI) and dependent (CD) decoding with either broad class (BC) or individual phoneme (Ph) 3-state models. Each HMM was trained on all sx/i TIMIT training sentences, using 39-dimensional MFCCs, 8-mixture observation densities, no skip transitions, and no language model or transition probability rescaling. The phone-level HMM transcriptions are collapsed into broad classes. Note that in our system, probabilistic segment decoding is a context dependent approach, though the sonority segmentation and vowel landmark detection methods are context independent. Our model complexity is closest to the HMMs using broad class models, as we only implement one detector per broad class.

Minimum string edit distance alignment was performed for all five systems. Table 3 summarizes the broad class transcription performance on the TIMIT test set. Our system accuracy falls in the range of the various HMMs. The high insertion rates of the context dependent HMMs are primarily a result of not applying a language model to clean up spurious segments. Our system is a conservative guesser, resulting in a remarkably low insertion rate even without a language model; this is largely a result of landmark thresholding before decoding.

Table 2. Obstruent/intervocalic region decoding performance on 17525/12915 phonemes in 15766/36255 segments.

	Obstruent			*Intervocalic*		
Method	*Acc*	*Corr*	*Ins*	*Acc*	*Corr*	*Ins*
Baseline	42.0	79.2	37.2	25.5	54.0	28.5
Standard	77.0	83.0	6.0	53.0	69.9	16.9
Rank ≤ 2	89.2	92.1	2.9	85.1	90.4	5.3
Rank ≤ 3	93.8	94.7	0.9	95.1	96.8	1.7

Table 3. Broad class transcription performance for our system vs. various HMM approaches with no phone or word language model.

System	*Acc*	*Corr*	*Ins*	*Del*	*Repl*
Our System	70.3	76.0	5.7	11.3	12.7
HMM, CI/BC	64.5	67.4	2.9	17.9	14.7
HMM, CD/BC	67.4	90.3	22.8	1.7	8.0
HMM, CI/Ph	68.9	79.4	10.5	6.1	14.5
HMM, CD/Ph	73.3	91.7	18.3	1.7	6.6

4. CONCLUSION

We have presented a probabilistic framework for speech recognition incorporating the ideas of distinctive feature hierarchies, landmark detectors employing statistical learning, and point process temporal pattern modelling. We believe this framework provides a promising direction for research in speech recognition. Moreover, our system implementation involves several design choices that are not necessarily scientifically or computationally optimal, leaving significant room for improvement.

First, the hard decisions made in the sonority segmentation and vowel landmarks result in two significant bottlenecks. A more robust solution is to employ probabilistic approaches here as well. Second, improving individual landmark detectors would put less burden on the integration procedure. Possible approaches include: (i) implementing acoustic parameters as an alternative to MFCCs [7][8]; (ii) individually addressing specific phoneme-level detector inadequacies; (iii) implementing broad class transition detectors; and (iv) using alternative machine learning techniques. Third, the integration model may be improved; possible strategies include: (i) alternative prior distribution estimation techniques (e.g. parametric modelling or non-parametric kernel-smoothing), (ii) limiting the number of independence assumptions, (iii) more sophisticated landmark time normalization methods, and (iv) alternative statistical frameworks (e.g. [5]). Fourth, our SSD method provides exceedingly accurate N-best estimates, indicating a performance improvement similar to HMM methods when a language model is applied. Last but not least, we would ultimately like to extend the methods presented in this paper to a full phonetic transcription. This will involve expanding the distinctive feature hierarchy to distinguish between the individual phonemes within each broad class.

5. REFERENCES

[1] K. N. Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *J. Acoust. Soc. Am*, vol. 111, no. 4, pp. 1872–1891, 2002.

[2] A. Juneja and C. Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," in *Proc. of IJCNN*, 2003.

[3] J. Goldsmith, *Autosegmental Phonology*. Garland Press, 1979.

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

[5] Z. Chi, W. Wu, and Z. Haga, "Template-based spike pattern identification with linear convolution and dynamic time warping," *J. Neurophysiology*, Accepted 2005 (in press).

[6] P. Niyogi and M. M. Sondhi, "Detecting stop consonants in continuous speech," *J. Acoust. Soc. Am*, vol. 111, no. 2, pp. 1063–1076, 2002.

[7] Z. Xie and P. Niyogi, "Robust acoustic-based syllable detection," in *Proc. of ICSLP*, 2006.

[8] O. Deshmukh, C. Espy-Wilson, and A. Juneja, "Acoustic-phonetic speech parameters for speaker-independent speech recognition," in *Proc. of ICASSP*, 2002.