Robust Keyword Spotting with Rapidly Adapting Point Process Models

Aren Jansen[1], *Partha Niyogi*[1,2]

[1]Department of Computer Science, [2]Department of Statistics
University of Chicago, Chicago, IL USA
aren@cs.uchicago.edu, niyogi@cs.uchicago.edu

Abstract

In this paper, we investigate the noise robustness properties of frame-based and sparse point process-based models for spotting keywords in continuous speech. We introduce a new strategy to improve point process model (PPM) robustness by adapting low-level feature detector thresholds to preserve background firing rates in the presence of noise. We find that this unsupervised approach can significantly outperform fully supervised maximum likelihood linear regression (MLLR) adaptation of an equivalent keyword-filler HMM system in the presence of additive white and pink noise. Moreover, we find that the sparsity of PPMs introduces an inherent resilience to non-stationary babble noise not exhibited by the frame-based HMM system. Finally, we demonstrate that our approach requires less adaptation data than MLLR, permitting rapid online adaptation.

Index Terms: keyword spotting, point process model, noise adaptation

1. Introduction

The goal of any speech recognition task is to map a continuous pressure wave to a symbolic linguistic sequence. Each approach to the problem must choose the intermediate representational level of detail upon which to build its statistical models. Frame-based approaches discard a significant amount of the original waveform and model variability in terms of a more parsimonious (and invariant) vector time series representation. While this remains the dominant approach, the optimal representation has yet to be conclusively determined. Insights from acoustic phonetics [1] and auditory neuroscience [2, 3] suggest that linguistic information may be more naturally and efficiently coded in sparse temporal point patterns corresponding to acoustic events in the speech signal and neural firing patterns in the auditory cortex.

In our previous work [4], we demonstrated that the durational statistics of suitably chosen time points can be used to extract keyword occurrences from continuous speech with the same accuracy as an equivalent frame-based keyword-filler HMM system. In addition to the inherent theoretical potential for reduced computational and model complexity, there is reason to believe that this sort of sparse temporal modeling may also improve recognition robustness. Indeed, while the framed speech signal does include some redundancy that might be useful, it also includes linguistically irrelevant portions that can be difficult to model robustly.

In an attempt to evaluate our intuition, we examine the robustness of our sparse point process-based keyword spotting system relative to the equivalent frame-based keyword-filler HMM system. We also evaluate a new unsupervised strategy for noise adaptation, which involves adjusting only the threshold of the low-level feature detectors that produce the point process representation. For completeness, we begin with a brief overview of the core keyword spotting system developed in [4]. This is followed by a specification of the corresponding adaptation strategy, the baseline keyword-filler HMM system, and the performance obtained by both in the presence of additive white, pink, and babble noise.

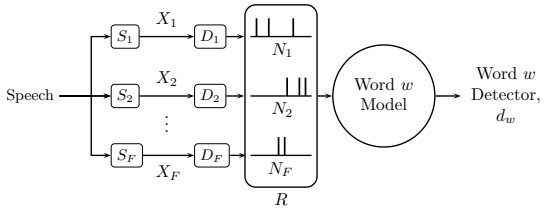

Figure 1: High-level architecture for the core system.

2. Core System Overview

Our core keyword spotting system consists of three main components (see Figure 1): an acoustic front end, a set of feature detectors, and word detector. Given a set $\mathcal{P}$ of phonological features and/or acoustic properties, the acoustic front end consists of one signal processor S_p for each $p \in \mathcal{P}$ that produces a vector time series representation, $X_p = \{x_1 x_2 \dots x_{\tau_p}\}$, where each $x_i \in \mathbb{R}^{k_p}$. A feature detector D_p for each $p \in \mathcal{P}$ transforms X_p into a point pattern $N_p = \{t_1, t_2, \dots, t_{n_p}\}$ comprised of the points in time that feature p is most strongly expressed or most perceptually salient.

The composite set of point patterns $R = \{N_p\}_{p \in \mathcal{P}}$ defines the sparse point process representation on which all subsequent modeling is based. A word model for each keyword w of interest is used to map R into a detector function $d_w(t)$ designed to take high values at times that the word is uttered and low values otherwise. Appropriately thresholded local maxima of d_w for each w define the points in time that the keyword is spotted.

2.1. Acoustic Front End and Feature Detectors

For the experiments conducted in this paper, we define our feature set $\mathcal{P}$ to be the standard 48 element set of phones defined in [4]. We used a standard 39-dimensional mel frequency cepstral coefficient (MFCC) front end, which includes velocity and acceleration coefficients. This front end is shared across all feature detectors (i.e., X_p is taken to be the same MFCC vector time series X of length τ for all p).

We define the detector D_p for each phone $p \in \mathcal{P}$ as a composition of two operations. First, we apply a phone-dependent detector function $g_p : \mathbb{R}^{k_p} \rightarrow [0, 1]$ to the vector time series X to produce a scalar time series $\{g_p(x_1), \dots, g_p(x_\tau)\}$ that should approach one when phone p is produced and zero oth-

erwise. These functions are derived from a basic monophone acoustic model consisting of a C-component per phone Gaussian mixture model (GMM) of the form

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

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 $\mathbf{\Sigma}$. Then, we may define $g_p(x)$ as

$$g_p(x) = P(p|x) = \frac{P(x|p)P(p)}{\sum_{p \in \mathcal{P}} P(x|p)P(p)}, \quad (2)$$

where $P(p)$ may either be assumed to be uniform or estimated from typical speech data. Second, we apply a thresholded peak finding function that computes the point pattern N_p as

$$N_p = \{i\Delta | g_p(x_i) > \delta_p \text{ and } g_p(x_i) > g_p(x_{i\pm1})\}, \quad (3)$$

where δ_p is the phone p detector threshold and Δ is the sampling interval of X.

2.2. Point Process Keyword Models

Our word model assumes there are two underlying stochastic processes that generate the observed point process representation $R = \{N_p\}_{p \in \mathcal{P}}$. The first is a homogeneous Poisson process that generates the observations in regions outside instances of the target keyword. The second is an inhomogeneous Poisson process that generates the point pattern corresponding to instances of the target keyword.

Our keyword w detector function, $d_w(t)$, may be defined in terms of the (log) likelihood ratio

$$d_w(t) = \log \left[\frac{P(R|\theta_w(t)=1)}{P(R|\theta_w(t)=0)} \right], \quad (4)$$

where R is the point process representation and $\theta_w : \mathbb{R} \rightarrow \{0, 1\}$ is an indicator function of time that takes the value 1 when the word utterance begins and 0 otherwise.

Given a time t and a candidate word duration $T \in \mathbb{R}^+$, we partition the point process representation R observed for an utterance of total duration L into three subsets: $R_l = R|_{(0,t]}$, $R_{t,T} = R|_{(t,t+T]}$, and $R_r = R|_{(t+T,L]}$. We assume conditional independence between subsets and assume that R_l and R_r are generated by the same homogeneous background process. Thus, the likelihoods of R_l and R_r cancel out in the ratio. Introducing the duration nuisance parameter T and noting that $P(R|\theta_w(t)=0)$ does not depend on T, Equation 4 reduces to

$$d_w(t) = \log \int \frac{P(R_{t,T}|T, \theta_w(t)=1)}{P(R_{t,T}|T, \theta_w(t)=0)} P(T|\theta_w(t)=1) dT. \quad (5)$$

The word duration distribution, $P(T|\theta_w(t) = 1)$, may be estimated directly from a set of word examples. The remaining two terms are modeled as follows:

1. For $P(R_{t,T}|T, \theta_w(t) = 1)$, we begin by normalizing $R_{t,T}$ to the interval $(0, 1]$; that is, we map $R_{t,T}$ to $R'_{t,T}$ such that for each $t_i \in R_{t,T}$ there is a corresponding $t'_i \in R'_{t,T}$ where $t'_i = [t_i - (t - T)]/T$. Given this mapping, we make the simplifying assumption that

$$P(R_{t,T}|T, \theta(t)=1) = \frac{1}{T^{|R_{t,T}|}} P(R'_{t,T}|\theta(t)=1).$$

This equivalence assumes that the observations for each instance of the word are generated by a common, T-independent inhomogeneous Poisson process operating on the interval $(0, 1]$ that is subsequently scaled by T to a point pattern on the interval $(t, t + T]$. Assuming an inhomogeneous Poisson process has generated $R'_{t,T}$, its likelihood takes the form

$$P(R'_{t,T}|\theta_w(t)=1) = \prod_{p \in \mathcal{P}} e^{-\int_0^1 \lambda_p(s) ds} \prod_{s \in N'_p} \lambda_p(s), \quad (6)$$

where $\lambda_p(s)$ is the process rate parameter at normalized time $s \in (0, 1]$ for feature p and N'_p are the elements of $R'_{t,T}$ for phone p. Learning a new word thus amounts to estimating the rate parameter functions, $\{\lambda_p\}_{p \in \mathcal{P}}$, from the point patterns observed when presented with examples of the word. This may be accomplished with kernel smoothing or maximum likelihood estimation of a parametric model.

2. For the $P(R_{t,T}|T, \theta_w(t) = 0)$ distribution, we need only consider a homogeneous Poisson process model that depends solely on the total number n_p of event arrivals observed for each feature p and the total duration of the segment (in this case T). The likelihood of $R_{t,T}$ given this homogeneous Poisson process model takes the form

$$P(R_{t,T}|T, \theta_w(t)=0) = \prod_{p \in \mathcal{P}} [\mu_p]^{n_p} e^{-\mu_p T}, \quad (7)$$

where μ_p is the background process rate parameter for phone p. Training this model amounts to estimating the rate parameters $\{\mu_p\}$ as the average detector firing rates over a large collection of arbitrary speech.

Given a novel utterance, we may evaluate the detector function by sliding a set of windows with durations distributed across the support of $P(T|\theta_w(t) = 1)$ and approximating the integral expression of Equation 5 with a sum. The interested reader may find more details of the core system in [4].

3. Adaptation Strategy

In general, there are two approaches one could take to adapt the above system to a novel acoustic environment. The first option is to adapt or retrain the keyword and background model parameters (i.e., λ_p and μ_p for each p) in the new environment. This approach requires a significant amount of additional keyword examples, and would thus be untenable in most practical situations as an online adaption strategy.

The second option is to adapt the feature detectors that produce the point process representation. Under this approach, one path forward would be to adapt the GMM parameters involved in Equation 1; indeed, this is a typical strategy used for continuous density HMM speaker and noise adaptation (see Section 4). However, our point process framework allows a much simpler alternative: adapt the detector thresholds $\{\delta_p\}_{p \in \mathcal{P}}$ of Equation 3.

Our strategy is to adjust phone detector thresholds to maintain the background firing rates measured in clean speech. Formally, let δ_p be the phone p detector threshold used in training the keyword and background models on clean speech. The background model contains a firing rate estimate μ_p for each phone detector measured in this clean setting. Now, given some amount of speech in the noisy environment, we can measure the new background firing rate μ'_p as a function of a new phone detector threshold value δ'_p. The goal, then, is to find the value of

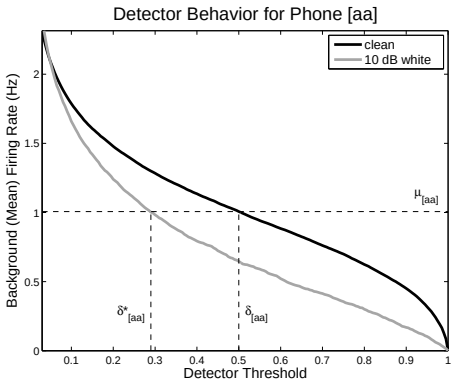

Figure 2: Empirical background firing rates for [aa] detector as a function of threshold.

δ'_p that produces a rate μ'_p in noisy speech that is closest to the original value μ_p. Thus, we define the adapted threshold δ^*_p as

$$\delta^*_p = \arg\min_{\delta'_p} |\mu_p - \mu'_p(\delta'_p)|. \quad (8)$$

Since this procedure adapts the phone detector thresholds to maintain clean speech firing rates, we simply reapply the original keyword and background models in the noisy environment.

Figure 2 shows the background firing rate of the detector for phone [aa] as a function of threshold for both clean speech and white noise corrupted speech at 10 dB SNR. As one might expect, the addition of noise reduces the detection certainty of this vocalic speech sound. Thus, to maintain the original firing rate measured when $\delta_{[\mathrm{aa}]} = 0.5$, the adapted threshold must take the reduced value of $\delta^*_{[\mathrm{aa}]} = 0.28$.

Our adaptation strategy is founded on two fairly strong assumptions: (i) the times and relative strengths of the local maxima of the detector functions evaluated across the utterance are preserved; and (ii) the addition of noise corrupts each phone detector the same way inside the keyword as it does elsewhere. The degrees of success observed in Section 5 serve as a measure of the validity of these assumptions in various noise conditions.

4. Baseline Keyword-Filler HMM System

We compare our point process-based keyword spotter to the standard keyword-filler hidden Markov model (HMM) specified in [5]. We allow for multiple pronunciations of each keyword and employ the above-described monophone acoustic model for state observation densities. The complete details of our implementation is provided in [6] for the interested reader.

To adapt the keyword-filler HMM system, we apply maximum likelihood linear regression (MLLR) to adapt means of the Gaussian components in the monophone model [7].[1] The strategy of this approach is to find the linear transformation that, when applied to the Gaussian means, will maximize the likelihood of the adaptation data. In our implementation, we compute separate transformations for each phone GMM. Since we are using full covariance matrices, we limit the search to diagonal linear transformations to allow a closed form solution.

In theory, MLLR can be applied unsupervised using an initial Viterbi or forward-backward alignment, which has been shown to be viable for speaker adaptation. However, in the context of noise adaptation, we found unsupervised MLLR to degrade performance relative to that of the original model. Thus, in the experiments below, the phonetic identity of each frame of the noisy adaptation data was provided for MLLR.

[1] MLLR can also be used to adapt covariances, but the relative gains over adapting means alone have been shown to be marginal [7].

5. Experimental Results

We consider the performance of the point process model (PPM) and keyword-filler HMM systems both in clean speech and in the presence of white, pink (equal energy per octave), and babble noise at various signal-to-noise ratios (SNR).[2] We trained the monophone acoustic model with $C = 8$ mixtures per state and one state per phone, using the entire TIMIT sx/i training set. This monophone model was used both to construct the PPM phone detectors and HMM observation densities. Keyword HMMs and PPMs for *greasy*, *wash*, *dark*, and *year* were trained using all 462 instances of each contained in the TIMIT sa1 training sentences. The Poisson intensity functions, $\{\lambda_p(t)\}_{p\in\mathcal{P}}$, were estimated using Epanechnikov kernel smoothing with bandwidth 0.1. Finally, clean speech background models were trained using all sx/i training sentences.

Each model was tested on all sa1/sx/si TIMIT test data, totaling 1.3 hours of speech containing at least 168 instances of each keyword. For each noise condition, we provided both systems the same 20 minutes of noisy training speech for adaptation and measured performance on noisy test data with the community standard figure-of-merit (FOM) score, which is defined as the average keyword detection rate when we allow 1, 2, ..., 10 false alarms per hour.

Table 1 shows the FOM performance (%) for the adapted (A) and non-adapted (N) point process models (PPM) and keyword-filler hidden Markov model (HMM) keyword spotting systems for the four keywords. The system with the best absolute FOM performance for each keyword and noise level/type is shown in bold. These results exhibit several trends:

1. Unsupervised PPM adaptation consistently provides sizeable gains under pink and white noise. Furthermore, PPM adaptation was only observed to significantly harm performance in the case of keyword *year* in pink noise at 20 dB SNR. Note that in this singular case, the non-adapted FOM score was inflated above the clean speech value. This peculiar behavior results from low levels of noise functioning to suppress false alarms to a greater degree than true detections.
2. The MLLR adaptation improvements are not as consistent, but in some cases lead to significant improvements under low levels of white and pink noise. It is important to emphasize, however, that MLLR was applied fully supervised. (As indicated above, we were unable to get improvements from unsupervised MLLR.) Note also that MLLR adaptation can significantly harm performance, particularly in low noise levels where the unadapted system is still doing well.
3. Neither adaptation strategy provides consistent improvements for babble noise, although the potential harm of PPM adaptation is much more limited. (The average adaptation losses under babble noise are 1.4 and 10 points for the PPM and HMM systems, respectively.) Interestingly, the non-adapted PPM keyword spotter is significantly more robust to babble noise at 20-30 dB SNR than the non-adapted HMM system. Since babble noise is non-stationary, it corrupts the speech signal non-uniformly in time. The frame-based HMM system is a model of the entire speech signal, and

[2] We used noise samples from the NOISEX-92 database [8].

Table 1: *Performance (in % FOM) for each keyword for various types and levels of noise. (A = Adaptation used, N = No adaptation)*

Noise		*greasy*				*wash*				*dark*				*year*			
		PPM		HMM		PPM		HMM		PPM		HMM		PPM		HMM	
Type	SNR	A	N	A	N	A	N	A	N	A	N	A	N	A	N	A	N
–	clean	–	**97.1**	–	91.0	–	**92.1**	–	72.9	–	76.8	–	**88.2**	–	**49.5**	–	36.1
white	30 dB	**89.2**	88.5	80.6	80.8	**83.8**	81.5	66.1	75.3	56.4	47.1	**72.5**	58.5	**56.1**	48.9	10.6	28.9
	20 dB	**79.3**	46.0	59.9	56.0	**69.5**	68.8	58.0	61.9	**42.1**	15.2	19.3	7.0	**25.3**	23.8	12.5	8.4
	10 dB	**29.5**	0.6	22.9	6.4	**23.9**	16.0	19.0	9.9	**10.4**	0.5	0.0	0.0	**7.6**	4.7	5.3	0.0
	0 dB	**2.3**	0.0	0.0	0.0	**3.8**	0.0	0.6	0.0	**1.1**	0.0	0.0	0.0	**0.4**	0.0	0.0	0.0
pink	30 dB	**94.3**	91.5	86.0	82.4	**84.3**	83.5	63.3	61.8	61.1	53.6	**80.4**	74.6	**56.4**	53.5	20.1	25.0
	20 dB	**73.0**	65.1	60.4	61.5	**73.0**	66.8	70.2	61.0	41.3	24.8	**51.9**	25.1	38.8	**56.6**	14.6	19.8
	10 dB	**44.5**	9.9	21.8	15.7	24.0	15.5	**27.3**	16.7	**9.8**	1.6	0.0	0.0	**10.9**	5.8	8.3	0.6
	0 dB	**5.1**	0.0	0.0	0.0	**1.1**	1.1	0.0	0.0	**1.6**	0.0	0.0	0.0	**0.7**	0.0	0.0	0.0
babble	30 dB	**97.4**	97.0	88.9	88.3	87.0	**89.2**	77.9	73.7	68.5	69.1	**70.4**	67.0	61.9	**63.3**	20.5	38.3
	20 dB	87.4	**90.8**	82.3	78.9	**82.4**	80.5	64.2	58.1	**56.6**	46.6	52.3	41.1	**34.8**	32.5	20.6	14.7
	10 dB	51.1	**55.2**	31.1	44.0	**34.8**	30.9	31.1	39.2	**25.0**	10.9	23.9	21.2	3.4	4.1	**8.5**	2.0
	0 dB	**7.2**	3.6	3.1	4.3	2.8	1.4	**5.8**	1.5	**3.4**	0.8	2.3	0.0	0.0	0.0	0.0	**0.6**

Table 2: *Average robustness levels of the systems.*

Noise		Adapted		Non-adapted	
Type	SNR	PPM	HMM	PPM	HMM
white	30 dB	92.4	72.7	84.9	84.6
	20 dB	65.8	50.5	47.5	44.4
	10 dB	21.3	16.5	7.0	5.2
	0 dB	2.2	0.2	0.0	0.0
pink	30 dB	95.5	82.0	90.7	82.3
	20 dB	71.7	65.5	71.6	58.6
	10 dB	26.7	21.1	10.2	10.5
	0 dB	2.5	0.0	0.3	0.0
babble	30 dB	102.3	85.3	103.7	95.0
	20 dB	80.9	73.7	76.8	63.4
	10 dB	32.5	31.9	28.2	32.9
	0 dB	3.7	3.5	1.6	2.1

thus must hold up against the entirety of the babble corruption. Our point process models, on the other hand, can be spared if the surges of babble intensity do not coincide with the times of detector firings.

Table 2 lists for each system the average robustness level, which we define as the FOM performance in noise measured relative to the corresponding clean speech value, averaged across keywords. The average adapted PPM robustness levels range from marginally to significantly higher than the adapted HMM system for all noise types and levels.

Finally, we consider the adapted system performance as a function of the amount of adaptation data used. We examine one case where phone detector threshold adaptation was particularly beneficial (keyword *greasy* in white noise at 20 dB SNR) and one case where MLLR adaptation was beneficial (*dark* in pink noise at 20 dB SNR). Figure 3 shows the FOM performance for both adapted systems as we vary the amount of adaptation data from 10 seconds to 20 minutes.[3] PPM adaptation improves performance down to 10 seconds, while MLLR sends HMM performance below the non-adapted value when given less than one minute of adaptation data.

6. Conclusions

We have presented a new unsupervised approach for adapting a point process-based keyword spotter system to noisy environments. We demonstrated that this method outperforms supervised MLLR adaptation of a equivalent keyword-filler HMM system in white and pink noise. Finally, we have found the non-adapted point process models to display an inherent robustness to moderate levels of non-stationary babble noise.

[3]Since the performance depends on the particular data chosen, the values reported are averages over multiple draws of the adaptation data.

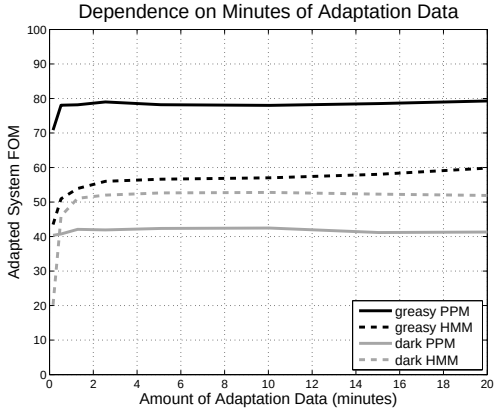

Figure 3: Adaptation data dependence.

7. 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] N. Suga, "Basic acoustic patterns and neural mechanisms shared by humans and animals for auditory perception," in *Listening to Speech: An Auditory Perspective (S. Greenberg and W. A. Ainsworth, Eds.)*, pp. 159–182. Lawrence Erlbaum Associcates, Mahwah, NJ, 2006.

[3] B. A. Olshausen and D. J. Field, "Sparse coding of sensory inputs," *Curr. Opin. Neurobiol.*, vol. 14, pp. 481–487, 2004.

[4] A. Jansen and P. Niyogi, "Point process models for spotting keywords in continuous speech," *IEEE Trans. Audio Speech Lang. Process.*, 2009, (To appear; also U. Chicago Tech. Rep. TR-2008-09).

[5] Igor Szöke, Petr Schwarz, Pavel Matějka, Lukáš Burget, Martin Karafiát, and Jan Černocký, "Phoneme based acoustics keyword spotting in informal continuous speech," in *Lecture Notes in Computer Science - TSD 2005 (V. Matousek et al.)*, pp. 302–309. Springer-Verlag, Berlin, 2005.

[6] A. Jansen and P. Niyogi, "An experimental evaluation of keyword-filler hidden Markov models," Tech. Rep. TR-2009-02, U. Chicago, Apr. 2009.

[7] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Comp. Speech Lang.*, vol. 10, pp. 249–264, 1996.

[8] Signal Processing Information Base, "Noise data," [Online] Available: http://spib.rice.edu/spib/select_noise.html, 1995.