A Detection Framework for Locating Phonetic Events

P. Niyogi, P. Mitra, M. M. Sondhi

Bell Labs – Lucent Technologies
Murray Hill, NJ 07974, USA.

ABSTRACT

We consider the problem of detecting stop consonants in continuously spoken speech. We pose the problem as one of finding the optimal filter (linear or non-linear) that operates on a particular appropriately chosen representation. We discuss the performance of several variants of a canonical stop detector and consider its implications for human and machine speech recognition.

1. INTRODUCTION

We are exploring a framework for speech recognition that utilizes the notion of distinctive features. An important problem that has to be solved for the success of such an approach is the accurate and robust detection of phonetic events. The acoustic cues for the different phonetic events are distributed non-homogeneously in the time-frequency plane, so separate detectors will be constructed for each of them. This is in contrast to approaches that use the same representation for all sound classes without special attention to their particular attributes. In this paper, we focus on the problem of detecting stop consonants in continuous speech. Stops present a challenging case because of their highly transient acoustic characteristics. Progress on this problem will be useful for speech recognition as well as automatic speech segmentation. Some aspects of this work that we would like to highlight are:

a) the signature of a stop consonant is a closure followed by a sharp release of broadband energy, especially at high frequencies. We therefore represent a stop by its spectrum and its Wiener entropy that provides a measure of spectral flatness to characterize the broadband nature of the burst. All spectra have been computed using multi-tapered spectral methods [5].

b) We propose to solve the problem of stop detection by constructing an optimal filter that operates on the above representation such that the output is high when there is a stop and low otherwise. Previous attempts at stop detection have typically attempted to take derivatives in appropriately chosen energy bands [2]. While this is intuitively a reasonable thing to do, differential operators are not necessarily optimal for stop detection. The filter derived by our method depends on the optimality criterion chosen. We compare the performance of several optimal filters with that of the differential operator, and show significant improvement.

c) We obtain ROC curves for the stop detection problem on the multi-speaker TIMIT database. When the correct detection rates range from 70 to 90 percent, the insertion rates range from 5 to 20 percent. (equal error rate — 16 percent). This is shown to be reasonably competitive with traditional HMM based methods. Crucially, however, our filtering framework requires 33 parameters trained from 4 speakers — a vast reduction in the number of parameters and consequently the amount of training data.

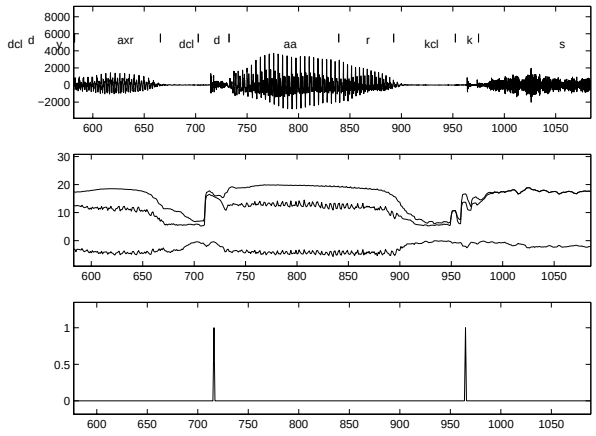

Figure 1: Portion of the speech waveform $s(n)$,(top panel), the associated three-dimensional feature vector, $\mathbf{x}(n)$ (middle panel), and the desired output $y(n)$ bottom panel marking the times of the closure-burst transition.

d) We show that in the above approach most of the false negatives are due to poorly released stops; many of the false positives can be explained as glottal stops or closures followed by strident fricatives (these are perceptually often like stops).

e) While detailed results are presented for the case of stops, our approach can be utilized to detect other kinds of phonetic events as well, and we will comment on these in the paper.

2. THE STOP DETECTION PROBLEM

Stop consonants are produced by causing a complete closure of the vocal tract followed by a sudden release. Hence they are signalled in continuous speech by a period of extremely low energy (corresponding to the period of closure) followed by a sharp, broad band signal (corresponding to the release). As a result, stops consonants are highly transient (dynamic) sounds that have a varying duration lasting anywhere from 5 to 100 ms. In American English, the class of stops consists of the sounds $\{$ p,t,k,b,d,g$\}$.

In order to build a detector for stop consonants in running speech, the speech signal, $s(t)$, is characterized by a vector time series with three dimensions — (i) log(total Energy) (ii) log(Energy above 3kHz) (iii) spectral flatness measure based on Wiener Entropy defined as $\int log(S(f,t))df - log(\int S(f,t)df)$. All quantities are computed using 5 ms windows moved every 1 ms. Multitapered estimates [5] are computed for the spectra from which energies and Wiener entropy are then calculated. Thus, we have $\mathbf{x}(n) = [x_1(n)\ x_2(n)\ x_3(n)]'$ where n represents time (discretized in units of milliseconds) and x_1 through x_3 are the three

acoustic quantities that are measured every 1 ms. Energies at 1 ms intervals potentially allow us to track rapid transitions that would otherwise be smoothed out by a coarser temporal resolution. This is particularly important since previous studies (e.g. [4]) indicate that burst durations for voiced stops could be as short as a few milliseconds. The Wiener entropy based flatness measure can be interpreted as a Kullback-Liebler divergence between $S(f,t)$ and a flat spectrum. It is also related to the predictability of the process $s(t)$.

We need to find an operator on the feature vector time series that will return a single dimensional time series that takes on large values around the times that stops occur and small values otherwise. The most natural points in time that mark the presence of stops are the transition from closure to burst release. Shown in fig. 1 is an example of a speech waveform $s(n)$, the associated feature vector time series $\mathbf{x}(n)$ and a desired output $y(n)$. The technical goal is to find an operator h on the time series $\mathbf{x}(n)$ that produces an output $y_h(n) = h \circ \mathbf{x}(n)$ such that $\parallel y - y_h \parallel$ is small in some sense (norm). Specifically, we choose the optimal operator (from some class $\mathcal{H}$ of operators) according to the criterion

$$h_{opt} = \arg\min_{h \in \mathcal{H}} R(h) = \arg\min_{h \in \mathcal{H}} E[(y - y_h)^2] \qquad (1)$$

it is easy to show that this is equivalent to approximating (by y_h) the conditional density time series $p(n) = E[\{y(n)\}|\{\mathbf{x}(n)\}] = P(y(n) = 1|\{\mathbf{x}(n)\})$ for the case when $y(n)$ takes values in $\{0,1\}$. Thus $p(n)$ is the conditional probability of a stop at time n given the time series $\{\mathbf{x}(n)\}$.

In this paper, we consider only linear convolution operators $h \circ \mathbf{x} = h * \mathbf{x}$. In actual practice, we deviate from the formulation of eqn. 1 since we don't have access to the true distribution that generates the time series $\{\mathbf{x}(n), y(n)\}$ and so cannot compute $R(h)$. We actually approximate $R(h)$ by an empirical risk $R_{emp}(h)$ computed from labelled examples (training data) given by:

$$R_{emp}(h) = \sum_{k=1}^{N} \sum_{n=1}^{N_k} w^{(k)}(n)(y^{(k)}(n) - y_h^{(k)}(n))^2$$

Here, N is the number of sentences in the training set. Each sentence corresponds to a particular realization of the process $(\mathbf{x}, y)$ and N_k is the length of the kth sentence. Let the kth sentence in the training set have m_k stops with corresponding closure-burst transitions occuring at times n_{kl} ($l \in \{1, \ldots, m_k\}$) respectively. Then $y^{(k)}(n)$ is 0 everywhere except for values of $n = n_{kl}$ where it takes the value 1. The weighting function, $w^{(k)}(n)$, is also 0−1-valued with $w^{(k)}(n) = 1$ everywhere except for $0 < |n - n_{kl}| < W$ where it takes the value 0. Finally, the kth filtered output, $y_h^{(k)}(n)$, is given by $y_h^{(k)}(n) = \sum_{i=1}^{3} \sum_j x_i^{(k)}(n-j)h_i(j)$.

Some remarks are in order:

1. This is an optimal filter design problem whose solution can be solved by adaptive means using Recursive Least Squares techniques. The filter can be trained from data to optimally match the desired output y. Taking derivatives of energy (correspondingly differences of energy at successive times) corresponds to a particular choice of the linear filter h.

2. The function $w^{(k)}(n)$ serves to weight the data so that parts of the signal near a stop transition (but not exactly at it) are not taken into consideration — it acts as a "don't care" region because it is not completely clear what a desirable output is near a transition. Further, from a numerical point of view, this allows the output $y_h^{(k)}$ some time to move smoothly from 0 to 1 and back again to 0 at the stops. In our experiments, the value of W was set at 6, i.e., a don't care region was effective from 5 ms before to 5 ms after a closure-burst transition. An optimal choice of W was not attempted.

3. In our experiments, we set $h_i(m)$ to be zero if $|m| \geq 6$. Thus there were ($33 = 3 \times 11$) free parameters for the filter that were then optimally learned from the training data in the manner described. On a test sentence, stops were detected by thresholding the output y_h obtained by filtering the feature vector $\mathbf{x}$ with h.

3. EXPERIMENTAL RESULTS

We present results of several stop detection algorithms on the TIMIT database. All results are presented on dialect region 4 of the test set containing 32 speakers, 16 male and 16 female saying ten sentences each, resulting in a total of 320 sentences. At every point in time (ms), the detection algorithm could potentially postulate the existence of a stop — clearly, as in any detection problem, one will need to balance the false acceptance rate (percentage of false detects, i.e., insertions) against the false rejection rate (percentage of stops not detected). As one varies the threshold for acceptance, corresponding ROC curves are generated and shown in fig. 2.

The overall conclusion from these experiments is that it is possible to attain an equal error rate of about 16 % on TIMIT speakers with a 33 parameter linear filter. It is possible to improve this to 12 % by moving to a non-linear filter with capacity control but we do not describe those results here.

A. This takes energy differences. The specific operation is given by $y_h = \sum_{i=1}^{2}(x_i(n) - x_i(n-1))$.

B. Taking optimized differences but only on energy components of $\mathbf{x}$. Thus $y_h = \sum_{i=1}^{2} \sum_j x_i(j)h_i(n-j)$ but the values of $h_i(m)$'s are now chosen optimally by minimizing $R_{emp}(h)$ constructed appropriately.

C. Taking optimized differences with all three components of the vector $\mathbf{x}$, i.e., the two energy components and the Wiener entropy. This is the complete formulation described in the previous section.

Some further points need to be made here:

1. In order to go from the output y_h to a set of candidate times n_i where stops are postulated, we need a decision rule. An appropriate one to use is to threshold y_h and pick peaks after thresholding. Each candidate peak n_i, was considered to be a correct detection if it was within 20 ms of the TIMIT labelling of the closure-burst transition, else it was considered to be a false insert.

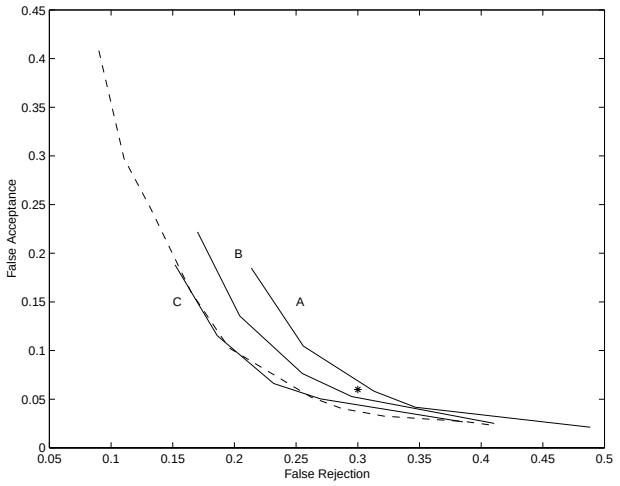

Figure 2: ROC curves for detection of stop consonants using three different algorithms.

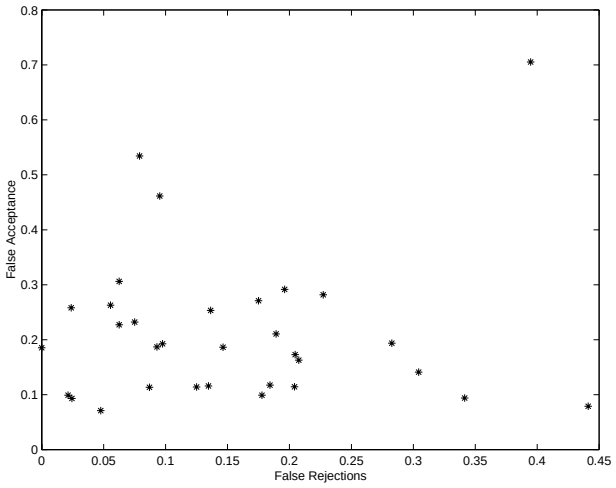

Figure 3: False acceptance and rejection rates by speaker.

2. Since there are only 33 parameters in the full-scale linear filter, the optimal parameters can be derived from very few training data. Specifically, in this case, we selected 4 speakers at random (2 male and 2 female) from the TIMIT training data base with 10 sentences from each, making a total of 40 sentences on which the detectors were trained.

3. Researchers have considered the problem of detecting phonetic events in running speech [1, 2, 3]. Unfortunately, they have not published ROC curves, nor compared their performance to other methods.

4. The '*' shown in the plot corresponds to the performance of a full blown HMM (32 mixtures; 3 state left-to-right models; 47 phonemes; free grammar; 450,000 parameters). The HMM was trained on an extremely data base with similar acoustic characteristics and run on the TIMIT sentences. The HMM output was decoded to segment the signal into stops and non-stops. Each closure-burst transition was considered to be correctly detected if it fell *anywhere* within a segment postulated as a stop by the HMM. This is a concession to the fact that the HMM is not designed to specifically locate the closure-burst transition.

5. The procedure outlined above can be extended to detection of other phonetic events, as well as to improve stop detection. From an algorithmic point of view what is needed for each such extension is a choice of representation, a choice of the operator h, and a decision rule. We are currently investigating other broad class transitions, e.g., fricative-vowel and vowel-nasal. As our models become more complicated, more parameters will be required and some method of capacity control will be required.

Let us examine more closely the nature of the errors made by the detector in C.

3.1. Errors by Speakers

Recall that there are 32 speakers in directory 4 of the TIMIT database. Here we examine how the stop detector performs on each of these speakers. We look at a few cases of bad performance to get some insight into the nature of these cases. Fig. 3 shows the acceptance and rejection rates for each of the 32 speakers (using algorithm C) for a point (18 % false rejection) on the ROC curve of fig. 2.

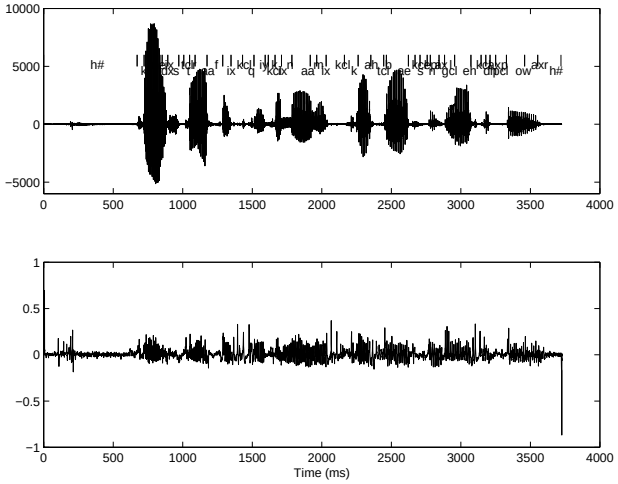

Figure 4: Sentence (top) and detector output (bottom) for speaker on whom performance is poor

Notice that there are some speakers for which the performance of the current detection algorithm is quite poor. It turns out that each of the speakers with high false acceptance rates was male with low pitch and occasionaly creaky voice with considerable glottalization. False firings of the stop detector often occured at the pitch pulses. Fig. 4 shows the output of the stop detector (before thresholding) on a sentence on which performance is particularly bad. Speakers with high false rejection rates typically had many poorly articulated stops. Fig. 5 shows the portion of a sentence corresponding to the stop “p”. Notice how the stop is very poorly articulated leading to poor detector performance. These two figures demonstrate some of the typical problems that are encountered with the current stop detection algorithms.

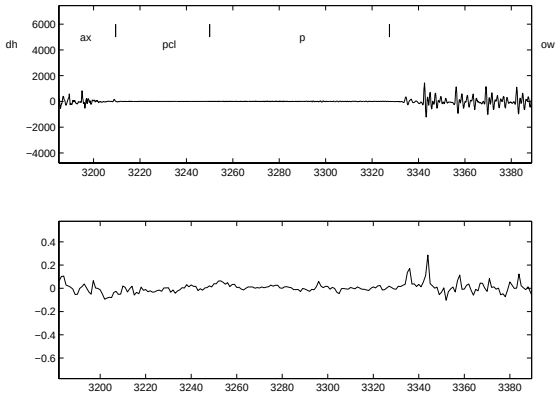

Figure 5: The stop "p" (top) and detector output (bottom) for speaker on whom performance is poor.

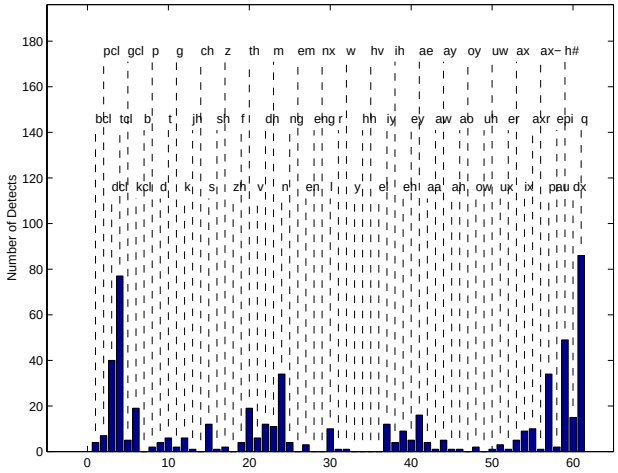

Figure 6: False insertions, left phonemic context.

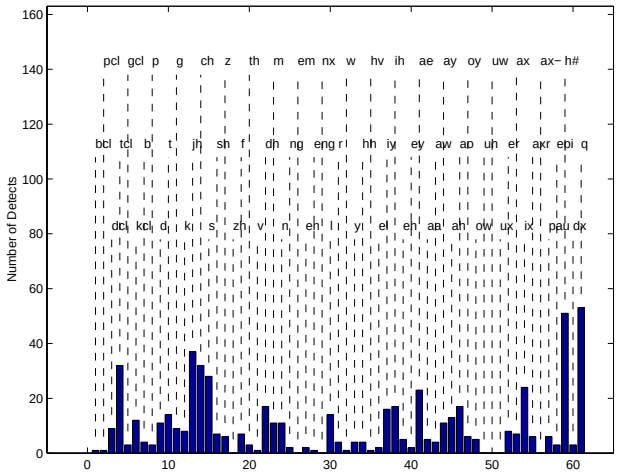

Figure 7: False insertions, right phonemic context.

3.2. Errors by Phonetic Class

Here we examine the detections/insertions to get a sense of how often they occur during different phonetic events in the speech signal. For convenience, we pick a particular point on the ROC curve of detection algorithm C with false rejection of 23 percent and false acceptance of 5 percent. For each point in time that was marked as a closure-burst transition by the stop detector, we located the closest true phonetic boundary (provided by the manual TIMIT segmentation) and noted the phonetic identity to the left and right of that boundary. Figs. 6-7 show the number of false insertions for each phonetic class (left and right). The TIMIT notation has been used for the phonemes. One notices some false inserts occuring during closures and releases of stops. These correspond to firings of the detector that are more than 20 ms away from the closure-transition boundary. Excluding these, the most common left contexts are "q" (glottal stop); "h#" (silence; presumably preceding the sentence); "pau" (pause); "n" (presumably due to the nasal closure); "th" (dental fricative; this has a partial closure and broadband nature). The most common right contexts are "q"; "ch"; "jh"(affricates having some stop like properties); "s" (strident fricative); "ae"; ix" (all vowels; presumably preceded by glottalization or silence or having noticeable pitch pulses). Most of these errors are not unreasonable and ways to eliminate them have to be considered.

4. CONCLUSIONS

We have considered the problem of detecting stop consonants in continuous speech. We have utilized a simple representation using log-energies and Wiener entropy to characterize the speech signal. Stops correspond to certain characteristic transitions in this feature space and we show how to use a filtering framework to extract the stops with reasonable accuracy and very little training data.

Many phonetic events, particularly those characterized by transitions, e.g. broad class boundaries, nasals etc. can be handled by a similar approach. While we have utilized a simple linear filter to extract the stop, one can, in principle, use more complex nonlinear filters to extract such phonetic events. Finally, the output of the filters can be interpreted as a *a-posteriori* density for the event and might be useful as an intermediate representation for other speech recognition and segmentation tasks.

5. REFERENCES

1. Glass, J. R. and Zue, V. W. "Detection and Recognition of Nasal Consonants in American English", Proceedings of ICASSP, pp. 2767-2770, 1986.
2. Liu, S. "Landmark Detection for Distinctive Feature-Based Speech Recognition", Ph.D.. Thesis. MIT, Cambridge, MA. 1995.
3. Mermelstein, P. "On Detecting Nasals in Continuous Speech," *Journal of the Acoustical Society of America*, pp. 581-587, 1977.
4. Niyogi, P. and Ramesh, P., "Incorporating Voice Onset Time to Improve Letter Recognition Accuracies," Proceedings of ICASSP, 1998.
5. D. J. Thomson, "Spectrum Estimation and Harmonic Analysis," Proc. IEEE, vol. 70, 1055-1096, 1982.