BURST DETECTION BASED ON MEASUREMENTS OF INTENSITY DISCRIMINATION

John-Paul Hosom* and Ronald A. Cole**

* Center for Spoken Language Understanding (CSLU)
  Oregon Graduate Institute of Science and Technology (OGI)
  20000 N.W. Walker Rd., Portland Oregon, 97291 USA
  e-mail: hosom@cse.ogi.edu    www: http://cslu.cse.ogi.edu/

** Center for Speech and Language Research (CSLR)
  University of Colorado, Boulder
  3215 Marine Avenue, Boulder Colorado, 80303 USA
  e-mail: cole@cslr.colorado.edu    www: http://cslr.colorado.edu/

ABSTRACT

Detection of burst-related impulses, such as those accompanying plosive stop consonants, is an important problem for accurate measurement of acoustic features for recognition (e.g., voice-onset-time) and for accurate automatic phonetic alignment. The proposed method of burst detection utilizes techniques for identifying and combining information about specific acoustic characteristics of bursts. One key element of the proposed method is the use of a measurement of intensity discrimination based on models from perceptual studies. Our experiments compared the proposed method of burst detection to the support vector machine (SVM) method, described below. The total error rate for the proposed method is 13.2% on the test-set partition of the TIMIT corpus, compared to a total error rate of 24% for the SVM method.

1. INTRODUCTION

Burst-related impulses (bursts) occur at the instant of release of plosive phonemes, and can be described as a sudden impulse-like increase in energy due to release of air from the mouth. These impulses are often followed by aspiration or frication noise, which in turn may be followed by a voiced phoneme. Identification of bursts can be combined with voicing detection results to estimate voice-onset-time (VOT), which may then be used in plosive-phoneme classification. In addition, burst detection can assist in accurate alignment of phonemes with their corresponding speech signal.

Accurate detection of bursts is a challenging problem because of the transient nature of bursts, the variety of phonetic contexts in which they occur, the similarity of bursts to certain types of noise, and differences in speaker and channel conditions. Several methods have been developed for locating the instant of burst release in plosive phonemes; including neural-network classification, HMM classification, thresholding of energy derivatives in various frequency bands, and the use of support vector machines (SVM). We will briefly describe the SVM method developed by Niyogi et al. because (a) its reported accuracy is the best found in the literature; (b) testing has been done on the commonly-available TIMIT corpus; and (c) a comparison between the SVM method and other methods has been given. We will then introduce a measure of intensity discrimination and show how this measure can be used to achieve accurate burst detection.

2. THE SVM METHOD

The SVM method, as described by Niyogi et al. [1], works by classifying a set of binary features (“burst” and “non-burst”) using a support vector machine; SVMs are binary classifiers that are considered to provide good generalization to unseen data. Niyogi et al. implemented two SVMs that are capable of linear and non-linear classification, respectively. The input to the SVM at each 1-msec frame consists of log energy of the entire spectrum, log energy of the frequency region from 3 to 8 kHz, and a spectral flatness measure, all computed with a 5-msec window. A detected burst is considered to be classified correctly if it occurs within 20 msec of the closure-burst label boundary obtained from phonetically hand-labeled TIMIT utterances. Training was done on randomly selected dialect regions of 40 sentences from the training partition of the TIMIT corpus, with 133 positive examples and 10760 negative examples. Testing was done on the test partition of one dialect region of the TIMIT corpus, using 320 sentences from 32 speakers. Niyogi et al. constructed ROC curves for their various classification methods, varying a parameter called \( U \) that “controls the trade-off between empirical fit to the data and capacity of the learning machine.” In addition, they evaluated a phoneme-based HMM approach and a derivative-of-energy approach on the same data.

Due to the large number of examples of non-bursts compared to the number of examples of bursts, an ROC curve in which performance is computed based on the total number of frames yields an extremely low number of false acceptances compared to the number of false rejections. To address this issue, Niyogi et al. constructed their ROC curves by evaluating the number of detected bursts with respect to the number of burst and non-burst phonemes instead of the number of burst and non-burst frames. The best performance is obtained by the non-linear SVM, with a total error rate of about 24%. The linear SVM has a total error rate of about 32%, the derivative-of-energy
approach has a total error rate of about 40%, and the HMM approach has a total error rate of 36%.

3. INTENSITY DISCRIMINATION

The use of intensity discrimination as a feature for automatic speech processing is motivated by perceptual studies on the smallest detectable change in intensity, conducted by several psychologists and summarized by Moore [2]. In the psychological studies, two-alternative forced-choice experiments were conducted, and subjects were asked to indicate which of the two stimuli contained the signal with an increased intensity. As Moore reports, a general pattern of intensity discrimination is clear despite variations in the methods and stimuli. Intensity discrimination can be modeled as a relative change in energy on the log scale:

$$\Delta L = 10 \log_{10} \left( \frac{I + \Delta I}{I} \right)$$

where $\Delta L$ is a measure related to the perceived change in intensity, $I$ is the intensity of the signal, and $\Delta I$ is the absolute change in intensity. Intensity, as defined by Moore, is the sound power transmitted through a given area, although it can also be used to describe “any quantity relating to the amount of sound, such as power or energy.” A fixed threshold for perceived intensity can be determined; if the value of $\Delta L$ is below this threshold, then the intensity change is not detected. Typical thresholds for detection are between 0.5 and 1 dB.

This model has been found to provide a good description of human detection of intensity changes. Depending on how the parameters of this model are chosen for automatic speech processing, a simple measure of intensity discrimination at different resolutions can be obtained. We apply the formula given by Moore directly, with intensity measured by the energy of the signal, and the window sizes for computing $I$ and $\Delta I$ dependent on whether we are interested in long-term or brief changes in the signal.

For a general measure of phonetically-relevant changes in the signal, we use a window size of 250 msec for $I$ and a window size of 40 msec for $\Delta I$. The window length for $I$ has been chosen to correspond to roughly the duration of one syllable, and the window length for $\Delta I$ corresponds to the minimum duration of a speech segment required for assigning phonetic quality and “the interval in which acoustic stimulation begins to assume an independent identity”, as reported by Greenberg [3]. As can be seen in Figure 1, this measure of intensity discrimination provides a reasonable indication of the onsets and offsets of major phonetic events, as local maxima and minima in the intensity-discrimination measurement correspond to major phonetic changes.

Measurements of intensity discrimination may be useful not only as a feature for detecting phoneme-level changes in the signal, but also for detecting other events in which changes in intensity are a factor.

4. PROPOSED BURST DETECTION

In our proposed method of burst detection, we use knowledge of the physical processes involved in production of bursts in order to classify burst-related impulses. Each burst is produced by a closure of the oral cavity in order to produce an increase in internal air pressure, followed by a sudden release of the constriction, which causes an abrupt increase in energy of the signal. Because of this process, bursts are characterized by at least 15 msec of low energy (during the closure), which is followed by a sudden increase in energy (at the instant of release), which is followed by a gradual decline in energy (during the release). Furthermore, the radiation characteristic of sound emanating from the mouth causes the burst at the instant of release to take on the qualities of an impulse, with a relatively flat spectrum and short duration. The spectral envelope of the burst is shaped to some degree by the surrounding phonemes. The proposed method detects bursts by applying the following criteria:

(a) There must be a relative increase in energy at the instant of release,

(b) The increase in energy must occur over most frequency bands, and

(c) The burst must have certain spectral properties that distinguish it from environmental noise (such as clicks).

These criteria can be satisfied by using Moore’s measure of intensity discrimination to estimate relative changes in energy, combining separate frequency-band information into a single...
measurement using Bayes’ rule, and using a neural-network classifier to incorporate spectral properties into the classification process.

The proposed method then works as follows:

1. Intensity discrimination is applied to bark-scale frequency bands. The window sizes for $I$ and $I'$ are small, in order to maximize the discrimination of bursts.

2. Normalization and equal-loudness weighting of the frequency bands is applied to the results of step (1), in order to give the frequency bands that are perceptually more important greater weight in the final result.

3. Assuming independence of the frequency bands, the weighted results of intensity discrimination are combined using Bayes’ rule. Each band is assumed to provide evidence for one of two conditions: a burst, or lack of a burst.

4. A threshold is used to select a number of “candidate bursts” for further processing.

5. A neural network is used to evaluate all candidate bursts, with input from the Bayes’ rule result and several frames of spectral information (PLP coefficients), and a binary burst/non-burst output.

Results from this process are illustrated in Figure 2, which shows an example waveform (containing letters of the alphabet) and its corresponding spectrogram, the higher six bands of intensity discrimination, the combined result of intensity discrimination, and the results of neural-network classification.

This method uses four parameters for locating candidate bursts: the window sizes for $I$ and $I'$, the window size for computing the delta in $I$, and the threshold value. Initial values for these parameters were determined from speech-specific knowledge and visual inspection of their effects. These initial values were then modified in small increments, and the resulting candidate bursts were evaluated on a development partition of the TIMIT corpus. Then, several sets of parameters with the best performance were selected for locating the candidate peaks on which to train neural networks. As the neural network is trained only on the detected candidate bursts, insertion errors can be reduced by the network, but deletion errors can not be recovered from. The initial networks were trained with the value of the candidate peak, 13 PLP coefficients, no delta values, and a “context” window of frames at $-5, 0, 5$ msec relative to the frame of interest. Once the final parameter set was chosen from the development-set results, then network training was changed to include delta values and a larger context window of frames from $+30$ to $+30$ msec relative to the frame of interest (at 5-msec intervals). Finally, the set of parameters and the neural network with the best performance on the development-set data were selected as the final system.

Evaluation was done in the same way as Niyogi et al., with a detected impulse considered correctly identified if it lies within 20 msec of the manually-labeled closure-burst boundary, and
the percentage of insertions and deletions measured relative to the number of plosive and non-plosive phonemes.

5. CORPORA

Three corpora were used in the development and evaluation of the proposed method: the TIMIT, OGI Stories [4], and OGI Portland Cellular [4] corpora. The TIMIT corpus contains read speech from 630 speakers from eight dialect regions of the United States. The sentences were designed to be phonetically rich, and were recorded with a Sennheiser noise-canceling, head-mounted microphone in a quiet environment. The speech was digitized at 16 KHz with 16-bit resolution. The OGI Stories corpus contains utterances of extemporaneous speech, where each utterance is approximately 40 seconds in length. These data were recorded over telephone channels. Speakers were recruited from throughout the United States, and were asked to speak on the topic of their choice for one minute. A total of 692 utterances are available, of which more than 200 have been transcribed with time-aligned phonetic labels. The data were recorded from an analog line, and digitized in 8 KHz 16-bit linear format. The Portland Cellular corpus consists of utterances obtained from speakers who were using cellular telephones. Like the Stories corpus, the Portland Cellular corpus contains extemporaneous speech on a topic of the speaker’s choice, and 200 calls have been transcribed at the phonetic level. The data were captured digitally from a T1 connection and saved in 8 KHz, 8-bit *-law format.

The evaluation of the four parameters was done on the TIMIT corpus, training of the networks was done on the TIMIT, Stories, and Portland Cellular corpora, development-set evaluation of the networks was done on the TIMIT corpus, and test-set evaluation was done on the TIMIT, Stories, and Portland Cellular corpora. These different corpora were used in training the networks and in final evaluation, in order to build and evaluate a system on different channel conditions. Development-set evaluation of parameters and networks was done on the TIMIT corpus, in order to have a single evaluation result on which to base the parameter and network selection. Training was done on 2000 examples from the TIMIT corpus, 4000 examples from the Stories corpus, and 2000 examples from the Portland Cellular corpus; these quantities were selected to provide nearly equal amounts of positive and negative examples.

6. RESULTS AND CONCLUSION

For the candidate peaks in the development set, the number of insertions ranged from about 30% to about 75%, and the number of deletions ranged from 1.25% to 6%, depending on the window sizes and threshold values. From the development set results, the parameter set of \( \{ I \text{ window size} = 22.0, \delta \text{ window size} = 24.0, \text{delta window size} = 14.0, \text{ and threshold} = 0.075 \} \) was selected for the final system. Test-set evaluation was done on 1344 sentences (6261 bursts and 45420 non-burst phonemes) from the TIMIT corpus, 42 sentences (2629 bursts and 19967 non-burst phonemes) from the Stories corpus, and 33 sentences (946 bursts and 8431 non-burst phonemes) from the Portland Cellular corpus. Results of these evaluations are given in Table 1.

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Insertion (%)</th>
<th>Deletion (%)</th>
<th>Total Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIMIT</td>
<td>5.14</td>
<td>8.06</td>
<td>13.20</td>
</tr>
<tr>
<td>Stories</td>
<td>11.56</td>
<td>8.34</td>
<td>19.91</td>
</tr>
<tr>
<td>Portland Cellular</td>
<td>25.26</td>
<td>8.08</td>
<td>33.34</td>
</tr>
</tbody>
</table>

Table 1. Insertion, deletion, and total error rate for each of three corpora (test-set results).

It can be seen from this table that the total error rate on the TIMIT corpus is 13.20%, which is a 45% reduction in error compared to the best total error rate reported by Niyogi et al. on this corpus. Also, it is interesting to note that the increased noise, different channel conditions, and lower sampling rate did not greatly affect the deletion rate, but had a dramatic impact on the insertion rate. One possible explanation for this is that people change their speaking style to compensate for degraded channel conditions, thereby enunciating bursts clearly enough to be detected at roughly the same rate for any channel. The increased noise for the Stories and Portland Cellular corpora may explain the increased insertion rate, as these channels are more likely to have non-speech phenomena that resemble burst-related impulses. Finally, it should be noted that performance of this method (without optimizing the code in any way) is about real-time on a Pentium Pro 200 MHz CPU.

7. ACKNOWLEDGEMENTS

The authors would like to thank the CSLU member companies for their support of this work. This work was sponsored in part by the National Science Foundation (grant numbers GER-9354959, IRI-9614217, and IIS-9970061); the views expressed in this paper do not necessarily represent the views of the NSF.

8. REFERENCES