THE REACTION SYSTEM: AUTOMATIC SOUND SEGMENTATION AND WORD SPOTTING FOR VERBAL REACTION TESTS

Gunnar Eisenberg, Thomas Sikora

Communication Systems Group
Technical University of Berlin, Germany
{eisenberg|sikora}@nue.tu-berlin.de

ABSTRACT

Reaction tests are typical tests from the field of psychological research and communication science in which a test person is presented some stimulus like a photo, a sound, or written words. The individual has to evaluate the stimulus as fast as possible in a predefined manner and has to react by presenting the result of the evaluation. This could be by pushing a button in simple reaction tests or by saying an answer in verbal reaction tests. The reaction time between the onset of the stimulus and the onset of the response can be used as a degree of difficulty for performing the given evaluation.

Compared to simple reaction tests verbal reaction tests are very powerful since the individual can simply say the answer which is the most natural way of answering. The drawback for verbal reaction tests is that today the reaction times still have to be determined manually. This means that a person has to listen through all audio recordings taken during test sessions and mark stimuli times and word beginnings one by one which is very time consuming and people-intensive.

To replace the manual evaluation of reaction tests this article presents the REACTION (*Reac*tion *Ti*me Determinat*ion*) system which can automatically determine the reaction times of a test session by analyzing the audio recording of the session. The system automatically detects the onsets of stimuli as well as the onsets of answers. The recording is furthermore segmented into parts each containing one stimulus and the following reaction which further facilitates the transcription of the spoken words for a semantic evaluation.

1. INTRODUCTION

There are three main classes of reaction tests, the plain *Reaction Time Test*, the *Stroop Test*, and the *Association Test*, which investigate different psychological phenomena. The typical setup for each of them is that a test person watches a screen on which a visual stimulus is presented. To gain the attention of the participant the stimulus is presented together with an alerting sound like a beep. After evaluating the stimulus the participant reacts by saying his answer. Simple java examples of non-verbal reaction tests can be found on the web [1, 2, 3].

In plain *Reaction Time Tests* the participant does not have to make any decisions about the presented stimulus [1]. He just has to acknowledge the perception of the stimulus as fast as possible. The test simply evaluates the reaction time's length. An example of a Plain Reaction Time Test could be that a red dot appears somewhere on the screen at random intervals in time. The participant has to say the word "dot" every time he discovers it.

Stroop Tests, named after their inventor, try to create some interference in the test person's consciousness between trained actions and cognitive abilities [2, 4]. Therefore the participant has to make a decision about the stimulus which is interfered by some opposing property of the stimulus itself. A well known example is reading color names (e.g. red, green, blue, etc.) which are printed in a different color or vice versa. Another example is naming the highest number out of a set of printed numbers with the smaller numbers being printed in a much bigger font than the higher numbers.

In *Association Tests* the test person is presented a picture or a word, often a noun (e.g. love, death, pleasure, etc.) on which he has to answer a certain emotional association (e.g. good, bad, embarrassing etc.) [3, 5].

Reaction test sessions are usually recorded on audio or videotape to be evaluated afterwards. On the audio track of the recordings the audible alert signals marking new stimuli and the answers of the participants are recorded. In this simple setup no additional information like time stamps or electronic markers for the onsets of new stimuli is recorded. This means that the recording is the only resulting material from the test session.

The advantages of using this simple setup is that it is very portable and investigators only have to take a minimum care of technical issues. The playback device usually is a laptop or sometimes a video cassette recorder with a TV-screen. The recording device is often an analogue dictating machine placed somewhere near the test person. Since until now the tests are manually evaluated afterwards, the poor recording quality is not impairing the evaluation as long as all answers can be understood.

To replace the manual evaluation an automatic evaluation system, like the one presented in this article, processes the sessions' recordings as input. It has to detect the recorded alert signals to determine the stimuli onsets and the onsets of the recorded answers. The system has to deal with the recording's poor quality like a high ground noise level, crackles, bad leveling and clipping.

2. PREVIOUS APPROACHES

The automatic measurement of reaction times in reaction tests breaks down into two tasks. One task is detecting the alert sounds' onsets marking the beginning of new stimuli. These onsets are the borders of segments, each containing a new stimulus and an answer. The other task is finding the onsets of the answered words. Both tasks have to be performed under noisy conditions.

Although there is actually no system which approaches the automatic evaluation of reaction tests directly there are approaches which perform tasks similar to the two subtasks mentioned above.

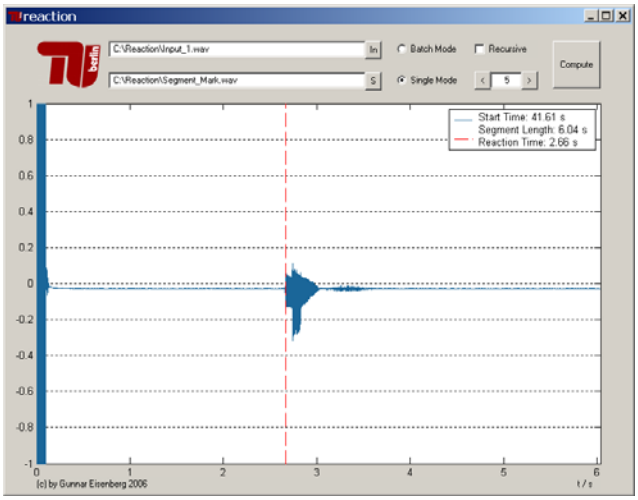

Figure 1: *REACTION's graphical user interface showing a computed segment together with the word's onset.*

Matsunaga et al. have presented a procedure for automatically segmenting broadcast news into speech, music and jingles (comparable to the given alert signal) and other classes [6]. In a noise free environment the detection rate for speech is 95.0 % and the detection rate for jingles is 87.7 %. The system has not been tested under noisy conditions.

Kim and Sikora have compared different algorithms for automatically segmenting sounds from different speakers in broadcast audio material [7]. The system does not need a priori information about the number of speakers and its recognition rate is 93.2 % for a scenario comparable to the scenario given in this work but with clean speech. Although the presented algorithms work well with clean speech the recognition rate drops with noisy environments.

Dufaux et al. have presented a system for automatic sound detection for noisy environments [8]. It detects impulsive sounds and is used for surveillance purposes. Their system has a recognition rate of up to 85.1 % for a SNR of 10 dB. The system could be useful for finding the alerts marking new segments but for an applicable system the recognition rate is still not high enough.

The work of Spina and Zue on automatic segmentation of general audio data [9] focuses on the training of segmentation systems which operate on noisy environments. Their work also shows the difficulty of trained recognition systems to deal with noise at all.

Various methods have been proposed for general onset detection which can also help solving the problem [10]. The recognition rates for onsets in a comparable scenario range from 70 % to 90 % and the problem of distinguishing between stimuli onsets and word onsets in an error prone environment would remain.

The cited approaches are developed to meet the requirements of a general case scenario. Therefore they turned out not to be robust enough to be directly applicable. As a result the REACTION system uses a different signal processing approach custom made for the given reaction test scenario.

3. THE REACTION SYSTEM

The user interface of the REACTION system can be seen in figure 1. The system needs two wave files in pcm-coded format,

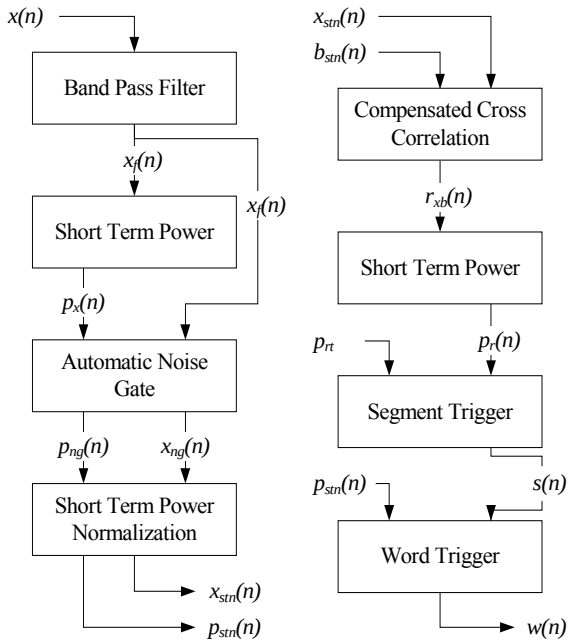

Figure 2: *Flowchart of the REACTION system showing the pre-processing chain (left) and the main processing chain (right). Input signals are shifted to the left, output signals are shifted to the right.*

mono or stereo with a minimum sample rate of 8000 Hz as input files for processing. One is the session's recording and the other is the short alert signal that marks the onsets of stimuli.

The process which is performed by REACTION is segmenting the session's recording by searching for the given alert signal so that each segment starts with the onset of a new stimulus. Further the onset of the test person's response is detected and the reaction time i.e. the time between the segment's start and the word's onset is determined. REACTION can operate on single sessions' recordings or in batch mode on several recorded sessions in one or more folders. The distinction between single or batch mode is done with the radio buttons in the upper right part of the interface. In batch mode the user can also set the system to crawl the selected folder recursively by checking the field "Recursive".

The program together with a manual and examples can be downloaded at our institute's website [11]. The usage is free for research purposes and in non commercial applications.

4. ALGORITHM

The two input signals of the system are $x(n)$ which is the session's recording and $b(n)$ which is the alert signal that marks the onsets of stimuli with n denoting the sample index. The algorithm's flowchart is shown in figure 2. It is divided in a pre-processing stage which operates on $x(n)$ and $b(n)$ and the main process. A part of a typical session's recording can be seen in figure 3.

4.1. Pre-Processing

The session's recording $x(n)$ is first band pass filtered with a second order Butterworth filter with cutoff frequencies at 50 Hz and 3900 Hz. This eliminates high frequency glitches and DC-

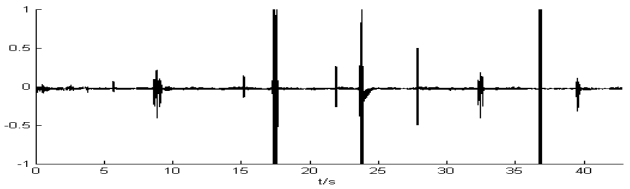

Figure 3: *Part of a typical input signal x(n). The ground noise level together with the bursts being alert sounds or spoken words can clearly be seen.*

offsets together with other low frequent rumble. After this initial filtering the signal is resampled to $f_s = 8000$ Hz for further processing.

For the resulting signal $x_f(n)$ the short term power $p_x(n)$ is computed with a window size of 25 ms, resulting in $N = 200$ for $f_s = 8000$ Hz:

$$p_x(n) = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} x_f{}^2(n+k) \,. \tag{1}$$

To eliminate ground noise an automatic noise gate is applied to the signal. Because of the nature of $x(n)$ most of its samples will neither contain speech nor parts of the alert but only the ground noise. Therefore a histogram is build to count the occurrences of the different values of $p_x(n)$. The value of of $p_x(n)$ which occurs most often will represent the ground noise level p_{gn}. All samples of $x_f(n)$ and $p_x(n)$ will be set to zero if their level is smaller than $2 \cdot p_{gn}$ resulting in the signals $x_{ng}(n)$ and $p_{ng}(n)$:

$$p_{ng}(n) = \begin{cases} 0 & \left| p_x(n) < 2\,p_{gn} \right. \\ p_x(n) & \left| p_x(n) \geq 2\,p_{gn} \right. \end{cases}, \tag{2}$$

$$x_{ng}(n) = \begin{cases} 0 & \left| p_x(n) < 2\,p_{gn} \right. \\ x_f(n) & \left| p_x(n) \geq 2\,p_{gn} \right. \end{cases}. \tag{3}$$

The signal is further normalized by a modified version of the short term power. Therefore the one sided decay envelope $v^*_{png}(n)$ of $p_{ng}(n)$ is computed:

$$v^*_{png}(n) = \max\left(p_{ng}(n), \quad v^*_{png}(n-1) \cdot \Delta \right). \tag{4}$$

The half value time for the exponential decay envelope is set to 220 ms, resulting in $\Delta = 99.961\%$ for $f_s = 8000$ Hz. The two sided decay envelope $v_{png}(n)$ is gained by applying equation (4) again to the reversed signal of $v^*_{png}(-n)$. The output signal $x_{stn}(n)$ and its power $p_{stn}(n)$ can be obtained by normalizing $x_{ng}(n)$ and $p_{ng}(n)$ to the power envelope as given by the following equations:

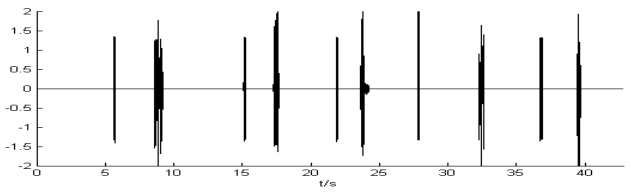

Figure 4: *The input signal after being pre-processed. The noise is gone and all bursts are normalized. The alert signals and spoken words can already be visually distinguished. The alert signals appear as cubic bursts whereas the words have a frayed shape.*

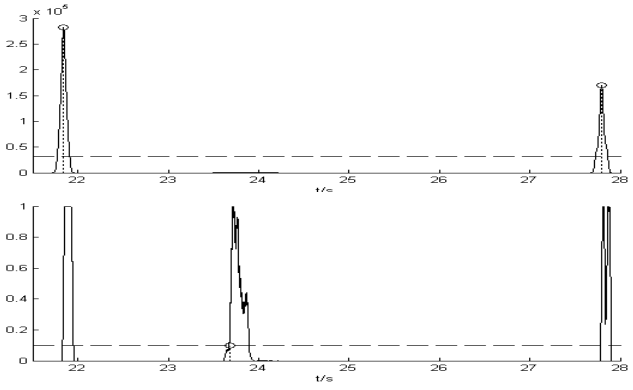

Figure 5: *The short term power of the cross correlation (top) and the short term power of the pre-processed input signal (bottom) together with horizontal dashed lines marking the static thresholds for new segments and word's onsets. The detected segment borders and word onsets are marked with vertical dotted pins.*

$$p_{stn}(n) = p_{ng}(n) / v_{png}(n), \tag{5}$$

$$x_{stn}(n) = x_{ng}(n) / \sqrt{v_{png}(n)}. \tag{6}$$

The pre-processing of the alert signal $b(n)$ to form the signal $b_{stn}(n)$ is formed accordingly to the steps described above. Only the automatic noise gate can be omitted because the nature of the signal is that it has no silent passages.

After having passed the pre-process stage the signals $x_{stn}(n)$ and $b_{stn}(n)$ are band limited, they are eventually noise gated and normalized in a way that their short term power is unity for the alert passages as well as for the spoken words. Figure 4 shows the signal from the example used in figure 4 after being pre-processed.

4.2. Main Processing

The second processing stage is the main process in which the alert signal's onsets and the words' onsets are determined. Since the signals have fixed properties after pre-processing this determination can be computed in a straight forward process.

First $x_{stn}(n)$ and $b_{stn}(n)$ are cross correlated to build the correlation signal $r_{xb}(n)$. To get rid of the typical phenomenon of oscillation of the correlation signal the short term power $p_r(n)$ of $r_{xb}(n)$ is computed according to equation (1), again using a window size of 25 ms.

Since $x_{stn}(n)$ and $b_{stn}(n)$ both are normalized in terms of their short term power no dynamic leveling needs to be applied for using the correlation's short term power $p_r(n)$ as a trigger to get the segment's onsets. It can directly be compared to a static threshold p_{rt}. This threshold is automatically determined to be 15 % of the maximum short term power value $p_{bb}(n)$ of the autocorrelation $r_{bb}(n)$ from $b_{stn}(n)$. Every local maximum of $p_r(n)$ marks a new segment as given by the following equations if it lies in a set of taps M_j whose according values of $p_r(n)$ lie above that threshold:

$$s_j(n) = \begin{cases} 1 & \left| n = \arg\max_{n \in M_j} p_r(n) \right. \\ 0 & \left| \textit{otherwise} \right. \end{cases}, \tag{7}$$

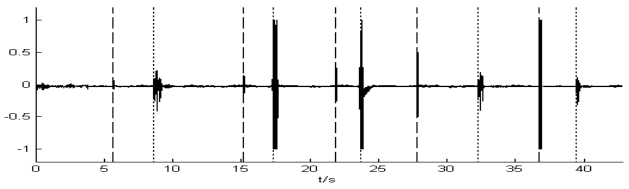

Figure 6: *The original input signal together with the segment borders (dashed lines) and the word's onsets (dotted lines).*

$$s(n) = \sum_j s_j(n). \tag{8}$$

The signal $s(n)$ which is also an output signal of the whole process, has the character of a trigger signal. It is 1 at the beginning of new segments and 0 elsewhere. To avoid multiple triggering the threshold has to be crossed for at least 10 ms (80 taps for f_s = 8000 Hz) which avoids triggering by glitches. Furthermore a new segment is only indicated if the last one is at least 50 ms gone (400 taps for f_s = 8000 Hz).

To find the word's onsets a slope technique is used. For every segment the word's onset is defined to be the first point in time where the normalized short term power $p_{stn}(n)$ of the signal $x_{stn}(n)$ reaches 25 % of its maximum, which is exactly 0.25 because of the normalization. To avoid triggering by glitches the threshold has to be crossed for at least 10 ms (80 taps for f_s = 8000 Hz). The signal $w(n)$ is derived from the words' onset times. It is 1 at the words' onsets and 0 elsewhere. Figure 5 shows parts of the signals $p_r(n)$ and $p_{stn}(n)$ for the example from figure 3 together with the generated triggers for segments and words' onsets. The resulting segmentation for the example signal can be seen in figure 6.

5. EVALUATION

The system was evaluated with real recordings of a reaction test. In this test 240 persons had to respond to 89 stimuli resulting in 21360 stimuli to be processed. The mean length of each test session's recording was 11.03 seconds and the total length of all evaluated recordings was 44:12 hours. The average reaction time determined in the tests was 3.30 seconds. The recordings were taken with an analogue dictating machine.

Although the quality of the recordings was quite poor, including the earlier mentioned flaws, the performance of the system was very good as it is depicted in figure 7. From the 21360 processed stimuli the REACTION system could segment 21162 segments (99.1 %) correctly. It has turned out that the system has never detected a new segment at a wrong point. Either the segment's border is detected correctly or it is missed completely. This behavior helps finding falsely segmented stimuli since they double the value of the determined (false) reaction time for the preceding segment. This marks these falsely segmented stimuli clearly as outliers in subsequent evaluations. Furthermore this behavior matches with outliers produced by semantic errors, i.e. when a person for some reason takes very long to respond to the presented stimulus. Therefore errors resulting from false segmentation can be ruled out quite easily afterwards.

From the 21162 correctly detected segments for 20583 words (97.3 %) the onset was detected correctly with an allowed tolerance of 15 ms. Compared to typical reaction times which are several seconds (in this case 3.30 s) the given tolerance is quite small. In total the number of correctly detected reaction times was 20583 (96.4 %).

Segmentation Rate	99.1 %
Onset Detection Rate	97.3 %
Reaction Time Detection Rate	96.4 %

Figure 7: *REACTION's Detection Rates*

6. CONCLUSIONS

The presented REACTION system can automatically detect reaction times from audio recordings of verbal reaction tests. It is indifferent against noise and other signal errors and because of its high recognition rate it is directly applicable and robust in everyday use.

7. REFERENCES

[1] G. Bradshaw, "Simple Choice Reaction Time," Available at http://epsych.msstate.edu/deliberate/SimpleRT/5.html, Accessed April 30, 2007

[2] E.Z. Yang, "Stroop Effect - Interactive Test," Available at http://www.thewritingpot.com/stroop/, Accessed April 30, 2007.

[3] T. Flynn, "Personality Test – Word Association Test," Available at http://www.similarminds.com/word/, Accessed April 30, 2007.

[4] C.M. MacLeod, P.A. MacDonald, "Interdimensional interference in the Stroop effect: uncovering the cognitive and neural anatomy of attention," *Trends in Cognitive Sciences*, vol. 4, no. 10, October 2000.

[5] G.R. Marshall, C.N. Cofer, "Associative Indices as Measures of Word Relatedness: A summary and Comparison of Ten Methods," *Journal of Verbal Learning and Verbal Behavior*, January 1964.

[6] S. Matsunaga, O. Mizuno, K. Ohtsuki, Y. Hayashi, "Audio Source Segmentation Using Spectral Correlation Features for Automatic Indexing of Broadcast News," in *Proc. 12th European Signal Processing Conference (EUSIPCO-2004),* Vienna, Austria, Sep. 2004.

[7] H.-G. Kim, T. Sikora, "Automatic segmentation of speakers in broadcast audio material," in *Proc. of SPIE*, Volume 5307, Storage and Retrieval Methods and Applications for Multimedia 2004, Dec. 2003.

[8] A. Dufaux, L. Besacier, M. Ansorge, F. Pellandini, "Automatic Sound Detection and Recognition for Noisy Environment" in *Proc. Xth European Signal Processing Conference (EUSIPCO-2000), Tampere,* Finland, Sep. 2000.

[9] M.S. Spina, V.W. Zue, "Automatic Transcription Of General Audio Data: Effect Of Environment Segmentation On Phonetic Recognition," in *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH '97),* Rhodes, Greece, Sep. 1997.

[10] S. Dixon, "Onset Detection Revisited," in *Proc. Workshop on Digital Audio Effects (DAFx'06),* Montreal, Canada, Sep. 2006.

[11] G.Eisenberg, "REACTION – Reaction Time Determination Software" Available at http://www.nue.tu-berlin.de/wer/eisenberg/reaction/, Accessed June 21, 2007.