TECHNOLOGY DIGEST
TD-08-050
©DECEMBER 2008
TIMELY TECHNOLOGY TRANSFER

The work described in this document was performed by Transportation Technology Center, Inc.,
a wholly owned subsidiary of the Association of American Railroads.

Estimation of Temporal Word Boundaries: A Building Block for Determining Human Alertness

Parham Shahidi, Steve C. Southward, and Mehdi Ahmadian, Virginia Tech and
David D. Davis, TTCI

Summary

Under Association of American Railroads (AAR) sponsorship through the Technology Scanning Strategic Research Initiative, Virginia Tech and Transportation Technology Center, Inc. (TTCI) have developed the tools needed to determine train crew and roadway worker alertness from voice communications. The system does not require comprehension of the speech, instead it uses measures like rate of speech (words per minute) to assess alertness. Alertness is determined by comparing speech when the subject is known to be alert (i.e., from a baseline measurement) to speech at the time of analysis. The first application for this system is likely to be analysis of train crew-dispatcher radio communications. This *Technology Digest* describes development of the tool used to determine the start and end of a word in the communications record.

A novel real-time algorithm has been developed for estimating temporal word boundaries (i.e., beginning and ending of each word) in measured speech without the need for interpreting individual words. This algorithm serves as a foundational building block for a method of estimating a variety of key metrics, such as word production rate, phrase production rate, and words per phrase, that are indicative of human mental states. The method will be used to develop a system for monitoring locomotive crew alertness. The majority of existing speech processing algorithms rely on prerecorded speech corpora (collection of language examples).[1,2] The real-time algorithm presented here is unique in that it employs a simple and efficient pattern matching method to identify temporal word boundaries by monitoring threshold crossings in the speech power signal. This algorithm eliminates the need to interpret the speech, and still produces reasonable estimates of word boundaries. The proposed algorithm has been tested with a batch of experimentally recorded speech data and with real-time speech data. During this testing phase, the following metrics were successfully extracted from the speech signal:

- Duration of gaps between words
- Duration of gaps between phrases
- Number of words
- Number of words over time (word production rate)
- Duration of words

The extraction of these metrics brings along the capability to estimate time-varying, real-time statistical measures such as minima, maxima, and mean values, which can in turn be used to correlate the metrics through an alertness inference engine.[3] The output of the inference engine is an alertness quotient that estimates the alertness of the train crew.

Please contact **David Davis (719) 584-0754** with questions or concerns regarding this *Technology Digest.* E-mail: David_Davis@ttci.aar.com.

BACKGROUND AND INTRODUCTION

Automated detection of temporal word boundary locations is a required function in many speech processing applications. Speech characteristics, such as word production rate, phrase boundaries and phrase production rate, rely on the knowledge of the correct number of words, as well as approximate temporal word boundaries. The primary objective of this research is to estimate the word boundaries without using linguistic knowledge of the speech.

The necessity of such a functional aspect in speech processing systems is gaining importance as speech processing is introduced into a constantly increasing number of everyday applications. Reasons for the deployment of automated services vary from convenience over security to economic measures, which simply favor automated systems instead of human personnel. Commercial examples include services such as ordering goods over the telephone, encrypting data by spoken passwords, mapping routes, and monitoring human states. Each one of these services deals with sensitive information and therefore requires prevention of information theft and preservation of the end user's privacy. Even though security has not been an issue in previous projects, similar approaches have been undertaken.[3,4]

The primary motivation for this research project is the need for a speech processing system for monitoring train driver alertness. Ultimately, this research will lead to the development of a nonintrusive method for quantifying an estimate of driver alertness based on analysis of noninterpreted speech. This is of high interest in the transportation industry for improving vehicle safety. However, for some vehicle applications, privacy rights prevent the use of common speech recognition engines for this task and therefore require the development of an alternative approach.

Common speech processing engines use word corpora, which provide a database with which segments of speech can be recognized and further processed.[1,2] The nature of the approach employed in this research project eliminates the necessity for a database and by doing so also adds the advantage of requiring less computational power, since fewer comparisons have to be performed. This key fact allows for the incorporation of the system in real-time applications. Alternative solutions for temporal word boundary detection have been developed, but these are based on computationally expensive methods that incorporate estimates of the signal direction and coherency, which make real-time deployment prohibitive.[5]

The method proposed by this research has been intentionally designed for simplicity in order to ensure its viability for real-time speech processing applications. The driver alertness monitoring application does not require precise knowledge of the temporal word boundaries. The proposed algorithm has been developed and tested on sample speech datasets in both a batch mode and a real-time mode.

Figure 1 shows a block diagram that outlines the individual functions of the algorithm. Sampled data from a microphone is passed into a speech processor and is subsequently processed through the word onset and offset detectors. After these steps, any estimated temporal boundaries are transferred into the word assembler where the speech signal is reconstructed. At this point in the algorithm, the actual computation of the metrics begins. From the reassembled speech signal, the word rate as well as the gap and the speech durations is estimated.

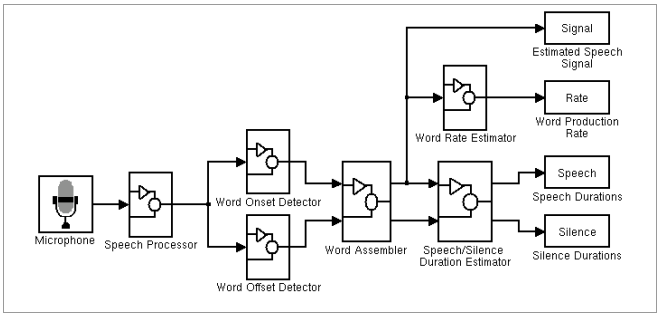

Figure 1. Block Diagram of Temporal Word Boundary Estimation Algorithm

The speech processor in Figure 1 is a subsystem that contains several additional sub-processes as indicated in Figure 2. Since it is known that the intensity of the speech signal fluctuates with respect to the presence of speech,[6] and in particular, words, the analysis method postulates that incoming words can be segmented in accordance with the intensity of the speech power signal. To measure the intensity of the speech signal, the processor first generates an estimate of the mean-square power in the speech signal.

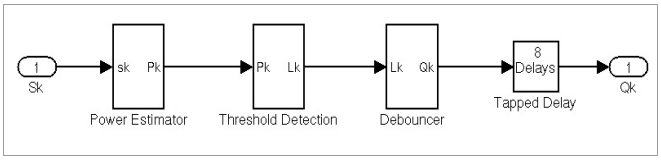

Figure 2. Block Diagram of Speech Processor Subsystem

To remove unwanted noise and improve the signal-to-noise ratio, the power estimator also filters the squared signal with a Butterworth 5th order low-pass filter with a break frequency set at 50 Hz.

In the threshold detector, the power signal is then compared to fixed programmable thresholds to generate a logical output signal. When the power increases past the "cut-on" threshold, a logical TRUE is generated. When the power decreases past the "cut-off" threshold, a logical FALSE is generated. The threshold values are design parameters that are set for a particular signal-noise ratio (SNR) condition.

The next step is to process the signal with a debounce algorithm. This function eliminates spurious rising and falling edges that can create false word boundaries. The mechanism used here does not allow a new edge to occur for duration of 40 milliseconds whenever a transition is detected. This time interval is also a design parameter that can be selected based on SNR.

The "tapped delay" at the end of the processor finally segments the speech into vectors which contain eight samples each. This format is required for the correct functionality of the onset/offset detectors.

EXPERIMENTAL VALIDATION

Batch Evaluation

The algorithm was first tested in batch mode with several different pre-recorded clean speech sample sets, one of which is presented here. The results below were obtained using a 7.73-second duration sample set. The speech data is a passage from written text that was spoken by the author and digitally sampled.[7]

Table 1 shows a comparison between the actual temporal word boundaries and the estimates from the proposed algorithm. The actual temporal word boundaries were manually extracted using a conventional audio signal processing software on the computer.

Table 1. Experimental Validation of Temporal Word Onset Detection

Spoken Word	Actual Onset (sec)	Onset Estimate (sec)	Actual Offset (sec)	Offset Estimate (sec)
One	0.792		0.967	
Of	0.967	0.8429	1.084	1.1065
Its	1.106	1.1627	1.311	1.3368
Main	1.335	1.4124	1.652	1.6857
Features	1.677	1.8106	2.209	2.1748
Is	2.214	2.2651	2.588	2.5705
That	2.811	2.8590	2.898	3.0596
It	2.896	3.1247	3.034	3.3413
Uses	3.072	3.3967	3.562	3.5740
A	3.611	3.6870	3.752	3.7719
Very	3.817	3.8501	4.160	4.4830
Large	4.160	4.6526	4.608	5.1813
Number	4.628	5.2387	5.09	5.4837
Of	5.186	5.7804	5.526	6.1950
Numerical	5.729	6.3075	6.416	6.4267
Problems	6.523	6.5206	7.183	6.7802

For this speech sample set, the proposed temporal boundary detection algorithm recognized a total of 15 words; although the actual number of words was 16. Table 1 also shows that the proposed algorithm occasionally detected a word onset and/or offset that has a higher offset from the actual value than others do.

Figure 3 shows a plot of the raw speech waveform used for this validation test, as well as the estimated power signal used for threshold comparison, and the actual and experimental word onset and offset times plotted as square waves, where a rising edge indicates a word onset. As indicated in Figure 3, the experimentally detected boundaries envelop the speech signal and approach the actual boundaries as stated above.

One metric for analyzing the accuracy of the proposed algorithm is the root mean square (RMS) value of the error between the actual and estimated word boundaries. For this purpose, the errors associated with the results that correspond to actual values were computed. The result obtained from this calculation was an RMS error value of 0.3179 seconds for the word onsets and 0.297 seconds for the offsets.

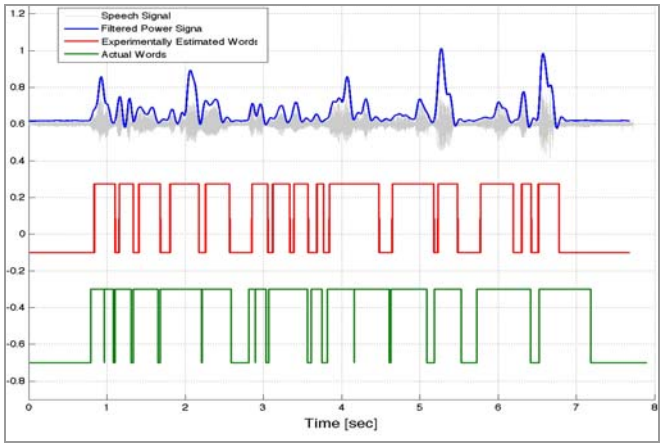

Figure 3. Raw Speech Data with Actual and Experimental Temporal Word Detection Results

These results demonstrate the accuracy of the proposed algorithm in detecting word boundaries. This accuracy is within acceptable limits for the driver alertness detection methods.

Real-Time Evaluation

The algorithm was also tested in real-time mode with the previously described hardware configuration. The core of the system is the dSPACE Autobox, which enables the creation of a real-time prototype of the algorithm and the evaluation of performance. Only the speech processor functionality was transferred to real-time and was successfully tested. Figure 4 outlines the results.

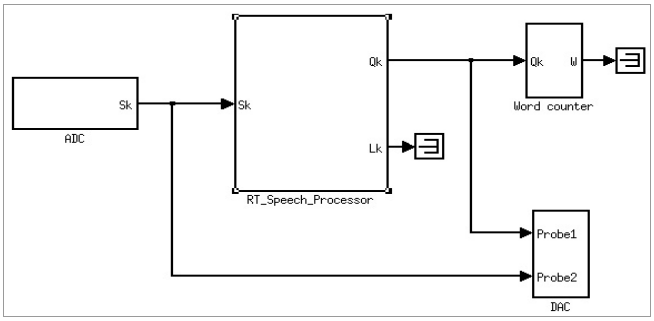

Figure 4. Real-Time System

The real-time system includes a digital-to-analog converter for probing internal signals as analog outputs, as well as an analog-to-digital converter for the input microphone. The blocks represent the connectors to the dSPACE system and provide a maximum of 16 inputs to the system as well as a maximum of six outputs. However, the speech processor is exactly the same as the one used in the batch mode system. For testing purposes, a word counter has been included in the algorithm.

Five different passages from a newspaper article were read to the system. The passages in the article were different in length ranging from 30 to 60 seconds. To further increase the validity of the results, a second speaker reread one of the passages in the article.

Although it is possible for a human to listen to the recorded speech with a sound editing tool such as SoundStudio in order to extract the temporal word boundaries, the sheer quantity of data makes this task very difficult. An alternative validation test is simply to count the total number of words in each spoken passage from the real-time system. Evaluating the word count metric is fundamental to further metrics such as word rate, words per phrase, and word intensity.

Table 2. Experimental Validation of Real-Time Performance

Duration (sec)	Speaker	Estimated Words	Actual Words
30	A	64	66
45	A	73	78
60	A	138	148
70	A	123	125
61	B	141	148

To quantify the accuracy of the proposed algorithm, the RMS value of the error between the actual number of words and the estimated number of words was calculated. The result of this calculation was an RMS error value of 4.92 percent.

Figure 5 is a 4-second plot of the speech signal with the temporal word boundary signal extracted from one segment of the real-time speech testing. The word boundary signal envelops the speech signal without any bounces and performs clear transitions between on and off state. This graphical validation is yet another indicator of the correct functionality and accuracy of the algorithm.

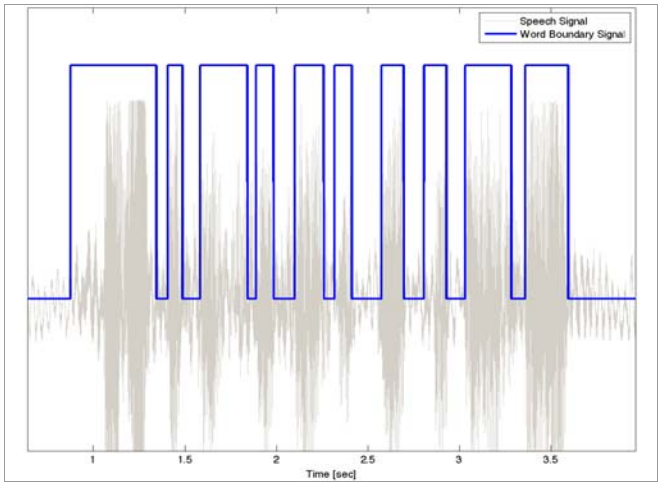

Figure 5. Recording of Real-Time Data Including the Speech and the Word Boundary Signal

CONCLUSION

An algorithm has been proposed for estimating temporal word boundaries in speech signals without the need for linguistic interpretation. This is the most challenging component of the overall design of a system for processing speech for use in monitoring driver alertness. This algorithm has been evaluated with batch data as well as in real-time mode, and the performance was found to be quite acceptable for the intended usage in developing a driver alertness monitoring system.

The success of the validation study for the batch mode temporal word boundary detection algorithm justified an extension of this work to the real-time implementation, which was further validated with experimental testing. Building on the temporal estimation of word boundaries and the number of words, this algorithm can easily be extended to estimate additional real-time speech metrics such as temporal phrase boundaries, word weighting, pitch, and intensity, which are known to be strong indicators of alertness.

REFERENCES

1. Dellaert, F., T. Polzin, and A. Waibel. 1996. "Recognizing emotion in speech," *IEEE*, vol. 3.
2. Lippmann, R.P. 1997. "Speech recognition by machines and humans." *Speech Communication*. 22(1): p. 1-15.
3. Shimp III, S.K., S.C. Southward, and M. Ahmadian. 2007. "Detecting crew alertness with processed speech." *American Society of Mechanical Engineers*, New York, NY.
4. Sharma, M. and R. Mammone. 1996. " 'Blind' speech segmentation: Automatic segmentation of speech without linguistic knowledge." *IEEE*, vol. 2. Piscataway, NJ.
5. Agaiby, H. and T.J. Moir. 1997. "Robust word boundary detection algorithm with application to speech recognition," *IEEE*, vol. 2. Piscataway, NJ,
6. Marzinzik, M. and B. Kollmeier, 2002. "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics." *IEEE Transactions on Speech and Audio Processing.* 10(2): p. 109-118.
7. Shenoi, B.A. 2006. "Introduction to digital signal processing and filter design." Hoboken, N.J.: Wiley-Interscience. Vol. 423 p.

Acknowledgement

Coauthors Parham Shahidi, Steve C. Southward, and Mehdi Ahmadian of Railway Technologies Laboratory, College of Engineering, Virginia Tech, Blacksburg, Virginia, thank the Association of American Railroads and Transportation Technology Center, Inc. for supporting this project.