USING ISOLATED VOWEL SOUNDS FOR CLASSIFICATION OF MILD TRAUMATIC BRAIN INJURY

Michael Falcone

Youngstown State University
Computer Science and Information Systems
Youngstown, OH, USA
mrfalcone@student.ysu.edu

Nikhil Yadav, Christian Poellabauer, Patrick Flynn

University of Notre Dame
Computer Science and Engineering
Notre Dame, IN, USA
nyadav@nd.edu, cpoellab@nd.edu, flynn@nd.edu

ABSTRACT

Concussions are Mild Traumatic Brain Injuries (mTBI) that are common in contact sports and are often difficult to diagnose due to the delayed appearance of symptoms. This paper explores the feasibility of using speech analysis for detecting mTBI. Recordings are taken on a mobile device from athletes participating in a boxing tournament following each match. Vowel sounds are isolated from the recordings and acoustic features are extracted and used to train several one-class machine learning algorithms in order to predict whether an athlete is concussed. Prediction results are verified against the diagnoses made by a ringside medical team at the time of recording and performance evaluation shows prediction accuracies of up to 98%.

Index Terms— Speech analysis, predictive models, health and safety, concussion

1. INTRODUCTION

Mild Traumatic Brain Injury (mTBI) is a serious problem for many athletes in the United States. In 2008, there were approximately 44,000 emergency department visits for sports-related mTBI [1]. Repeated concussions can cause risks such as dementia and Parkinson's disease. In the U.S., TBI accounts for an estimated 1.6-3.8 million sports injuries every year [2] and nearly 300,000 concussions are being diagnosed among young athletes every year [3]. Athletes in sports such as football, hockey, and boxing are at a particularly large risk, e.g., six out of ten NFL athletes have suffered concussions, according to a study conducted by the American Academy of Neurology in 2000. However, TBI is also very frequent among soldiers, and is often called the "signature wound" of the Iraq and Afghanistan wars. Recent insights that the neuropsychiatric symptoms and long-term cognitive impacts of blast or concussive injury of U.S. military veterans are similar to the ones exposed by young amateur American football players [4] have led to collaborative efforts between athletics and the military, e.g., the United Service Organizations Inc. recently announced that it will partner with the NFL to address the significant challenges in effectively detecting and treating TBI. The importance of new and novel ways to assess mTBI has become increasingly important as a consequence. Tests which are easy to administer, accurate, and not prone to unfair manipulation are required to assess mTBI. In this paper, the feasibility of using speech analysis for detection and assessment of mTBI is studied. Vowel sounds are isolated from speech recordings and the best acoustic features which are most successful at assessing concussions are identified. The remainder of this paper is structured as follows. In Section 2, we describe prior efforts in studying brain injuries and their effects on speech. In Section 3, we outline the nature of the speech recordings and the different vowel sounds they are comprised of. Section 4 describes the vowel extraction and classification procedure with results. Finally, Section 5 concludes the paper and describes future work.

2. RELATED WORK

There have been several previous studies related to motor speech disorders and their effects on speech acoustics. Theodoros et al. conducted a study of the speech characteristics of 20 individuals with closed head injuries (CHI) [5]. Their main result was that the CHI subjects were found to be significantly less intelligible than normal non-neurologically impaired individuals, and exhibited deficits in the prosodic, resonatory, articulatory, respiratory, and phonatory aspects of speech production. Ziegler and von Cramon discovered an increase in vowel formant frequencies as well as duration of vowel sounds in persons with spastic dysarthria resulting from brain injury [6]. In [7], a variation of the Paced Auditory Serial Addition Task (PASAT) test, which increases the demand on the speech processing ability with each subtest, is used to detect the impact of TBI on both auditory and visual facilities of the test takers. Hinton et al. [8] illustrated that tests on speech processing speed were affected by post-acute mTBI on a group of rugby players. Recently, Tsanas et al. used acoustic features of sustained vowels to classify Parkinson's Disease with Support Vector Machines (SVM) and Random Forests (RF), and showed that SVM outperformed RF [9]. Studies have also been conducted on the accommodation phenomenon, where test takers tend to adapt or adjust to unfamiliar speech patterns over time. Research has shown that accommodation is fairly rapid for healthy adults [10, 11], and it is studied as a speed based phenomenon in [12]. To the best of our knowledge, work investigating the effects of concussion on specific speech features like formant frequencies, pitch, jitter, and shimmer, has not been researched extensively using real-world speech data before. This is also the first study to address the feasibility of using the relationship between TBI and speech to develop a more scientific and novel concussion assessment technique.

3. DATA

Speech recordings were acquired under a protocol approved by the Institutional Review Board (IRB) at the University of Notre Dame. Speech data consisted of recordings taken from 105 male athletes

before, during, and after participation in several matches of a boxing tournament. Subjects were recorded speaking a fixed sequence of digits that appeared on screen every 1.5 seconds for 30 seconds. Subjects spoke digit words in the following sequence: "two", "five", "eight", "three", "nine", "four", "six", "seven", "four", "six", "seven", "two", "one", "five", "three", "nine", "eight", "five", "one", "two".

Each subject was recorded on a mobile tablet by a directional microphone. Several of the recordings contained background noise or background speakers. Speech was sampled at 44.1 kHz with 16 bits per sample in two channels and later mixed down to mono-channel for analysis. The recordings were split into training/test data and grouped into three classes: *baseline* (training), *post-healthy* (test), and *post-mTBI* (test). Table 1 summarizes these classes and gives the number of recordings in each class. A few speakers have recordings in both the *post-healthy* class and the *post-mTBI* class if they were diagnosed with mTBI in a match following acquisition of the *post-healthy* recordings. In such cases, the recordings were taken in separate matches of the tournament. Thus, the number of test recordings is greater than the number of training recordings but both sets of data are mutually exclusive.

Table 1. Classes of speech recordings.

Class of Speech	n	Description
baseline	105	Recorded prior to tournament; all subjects healthy
post-healthy	101	Recorded following preliminary match; subjects not diagnosed with mTBI and assumed healthy
post-mTBI	7	Recorded at subject's final match of participation; subjects diagnosed with mTBI

4. METHODS

4.1. Isolating vowel segments

Vowel segments were isolated from each speech recording by first locating vowel onsets and then extracting 140 ms of speech for each vowel sound, following each onset. Onsets were detected using an adaptation of the method described by Hermes for onset detection in isolated words [13]. This process yielded a total of 3786 vowel sounds among each of the three classes of recordings. Table 2 shows the number of segments isolated from each class of recordings. Note that each class contains a different number of vowel sounds. This is because the number of whole recordings differs for each class and occasionally vowel onsets are missed during the isolation process.

4.2. Extracting features from vowels

Eight speech features were investigated in this study: *pitch, formant frequencies* F_1*-*F_4*, jitter*, *shimmer*, and *harmonics-to-noise ratio (HNR)*. While jitter and shimmer are typically measured over long sustained vowel sounds, recently the use of jitter over short-term time intervals has shown promise in analyzing pathological speech [14].

Pitch was estimated using autocorrelation and formants were estimated via FFT. LPC was avoided for formant estimation since it can be prone to errors and because its accuracy depends heavily on proper choice of pole order [15].

Table 2. Number of vowel sound instances isolated from each class of speech recordings.

Sound	*baseline*	*post-healthy*	*post-mTBI*
/i/$_{\text{three}}$	150	160	10
/I/$_{\text{six}}$	190	188	12
/e/$_{\text{eight}}$	162	160	10
/ɛ/$_{\text{seven}}$	207	200	14
/ʌ/$_{\text{one}}$	205	189	13
/u/$_{\text{two}}$	212	224	18
/o/$_{\text{four}}$	204	202	14
/ai/$_{\text{five}}$	313	302	21
/ai/$_{\text{nine}}$	205	190	11

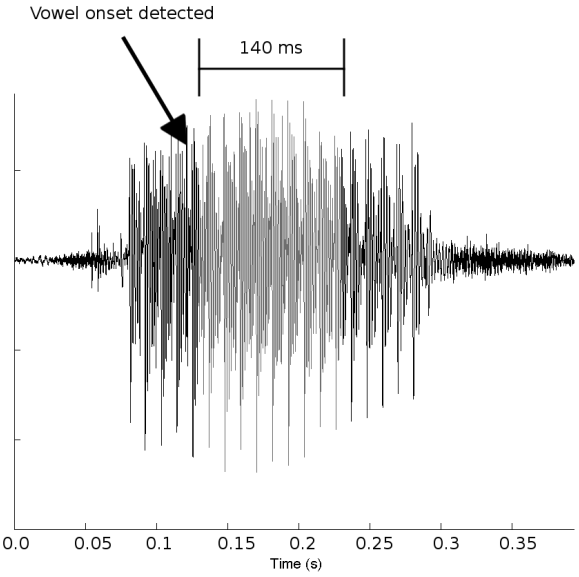

Fig. 1. The extraction of the /ai/ vowel sound from the recording of a subject speaking "five".

Jitter is a measure of the average variation in pitch between consecutive cycles, and is given by

$$Jitter = \frac{\sum_{i=2}^{N} |T_i - T_{i-1}|}{N-1}$$

where N is the total number of pitch periods and T_i is the duration of the ith pitch period.

Shimmer is a measure of the average variation in amplitude between consecutive cycles, given by

$$Shimmer = \frac{\sum_{i=2}^{N} |A_i - A_{i-1}|}{N-1}$$

where N is the total number of pitch periods and A_i is the amplitude of the ith pitch period.

4.3. One-class SVM for mTBI classification

Combinations of extracted features were selected as inputs to several one-class SVM classifiers. The LIBSVM implementation [17] was used. A one-class classifier was chosen because the baseline

data did not include any mTBI speech and the number of recordings in the *post-mTBI* class was significantly lower than the number of recordings in *post-healthy*. Features were scaled to the ranges 0-1 by dividing each feature by the maximum value of that feature in the training set. In order to find the optimal combination of features for each vowel sound, each possible combination of at least three features was used to train and test the classifier for each vowel sound.

4.4. Classification of individual vowel sounds

An individual classifier was trained for each vowel sound in the *baseline* class. The /ai/ sound in the word "five" was treated separately from the /ai/ sound in "nine", because the consonantal context differs between these words, i.e., the /ai/ sound in "five" occurs between two fricatives while the /ai/ sound in "nine" occurs between two nasal consonants. Each sound in the *post-healthy* and *post-mTBI* classes was tested and the prediction results were used to compute three standard performance measures: *recall*, *precision*, and *accuracy*. Recall gives the percentage of correctly predicted mTBI segments and is defined as

$$\text{Recall} = \frac{\text{\# of segments correctly classified mTBI}}{\text{Total \# of true mTBI segments}}$$

Precision is the rate at which the mTBI predictions were correct, and is defined as

$$\text{Precision} = \frac{\text{\# of segments correctly classified mTBI}}{\text{Total \# of segments classified mTBI}}$$

Finally, accuracy is the percentage of segments that were classified correctly (either mTBI or healthy), and is defined as

$$\text{Accuracy} = \frac{\text{\# correctly classified segments}}{\text{Total \# of segments}}$$

The classifier achieved accuracies approaching 70% for some feature combinations and recall rates as high as 92% for other combinations. Table 3 shows the features that achieved maximum *accuracy* for each vowel sound. In any case where equal accuracies were achieved for more than one feature combination, the combination yielding the best recall is listed.

Table 3. Vowel sounds and features achieving maximum accuracy.

Vowel	Recall	Prec.	**Acc.**	Features*
/i/	0.4 (4/10)	0.069	0.65	F_3,F_4,J,H,P
/I/	0.5 (6/12)	0.11	0.71	F_1,F_4,S,H
/e/	0.6 (6/10)	0.083	0.59	F_4,J,H
/ɛ/	0.5 (7/14)	0.089	0.63	F_3,S,H,P
/ʌ/	0.54 (7/13)	0.095	0.64	F_4,S,H,P
/u/	0.61 (11/18)	0.11	0.59	F_3,F_4,J
/o/	0.79 (11/14)	0.14	0.67	F_1,F_4,S
/ai/$_{\text{five}}$	0.76 (16/21)	0.13	0.66	F_1,F_3,J,S,H,P
/ai/$_{\text{nine}}$	0.64 (7/11)	0.097	0.66	F_2,F_3,F_4

* F_n = frequency of formant n, J = jitter, S = shimmer, H = harmonics-to-noise ratio, P = pitch frequency

Table 4 shows the feature combinations that achieved maximum *recall* for each vowel sound. In any case where an equal recall was achieved for more than one combination of features, the combination yielding the best accuracy is shown. In any case where multiple feature combinations yielded equal maximum recalls and equal accuracies, the combination with the fewest number of features was chosen. In the case of the /e/ sound, two combinations yielded recalls of 80% and accuracies of 56%. In this case, all features from both combinations were used despite a reduction in accuracy for that sound by 3%.

Table 4. Vowel sounds and features achieving maximum recall.

Vowel	**Recall**	Prec.	Acc.	Features*
/i/	0.9 (9/10)	0.11	0.55	F_1,F_3,S
/I/	0.92 (11/12)	0.1	0.51	F_1,F_2,P
/e/	0.8 (8/10)	0.093	0.53	F_2,F_4,S,P
/ɛ/	0.79 (11/14)	0.11	0.57	F_2,J,S
/ʌ/	0.77 (10/13)	0.1	0.55	F_1,F_4,P
/u/	0.89 (16/18)	0.13	0.55	F_2,F_3,J,S,P
/o/	0.79 (11/14)	0.14	0.67	F_1,F_4,S
/ai/$_{\text{five}}$	0.81 (17/21)	0.14	0.66	F_1,F_2,F_3,J,S,H,P
/ai/$_{\text{nine}}$	0.82 (9/11)	0.12	0.65	F_1,F_2,F_3

* F_n = frequency of formant n, J = jitter, S = shimmer, H = harmonics-to-noise ratio, P = pitch frequency

4.5. Classification of whole recordings

The classification of boxers' speech recordings by using each vowel can now be elaborated. A tradeoff between accuracy and recall can be seen from Table 3 and Table 4 for most vowel sounds. In order to keep false negatives to a minimum, a higher importance was placed on recall of mTBI vowel sounds. Similarly to individual vowel sound segments, performance of whole recording classification was evaluated by measuring recall, precision, and accuracy measures.

Using the feature combinations that achieved maximum recall for individual vowel sound segments (Table 4), individual one-class SVM classifiers were again trained for each vowel sound in the *baseline* class of recordings. Next, each speech recording in *post-healthy* and *post-mTBI* was classified as a whole by classifying each instance of a specific vowel sound from the recording. A threshold was defined, δ, such that the speech recording was classified as mTBI speech if the following relationship holds true:

$$\delta \leq \frac{N(v)}{M(v)}$$

where N gives the number of instances of the vowel sound v classified as mTBI in the recording and M gives the total number of instances of v that could be isolated in the recording. Several trials were performed in which each recording was classified and performance was measured with v as a different vowel sound for each trial, i.e., each unique vowel sound corresponds to a single trial. For each trial, δ was adjusted until recall of mTBI recordings reached 100%. The corresponding value of δ is shown in Figure 2.

A final classification trial was performed in which all vowel sounds were aggregated such that a recording is classified as mTBI speech if

$$\delta \leq \frac{\sum_{v \in V} N(v)}{\sum_{v \in V} M(v)}$$

holds true, where V is the set of all vowel sounds isolated from that recording. Figure 2 compares performance measurements and shows the minimum δ for each trial that resulted in recall of all seven mTBI recordings. The "All" trial in Figure 2 shows the performance

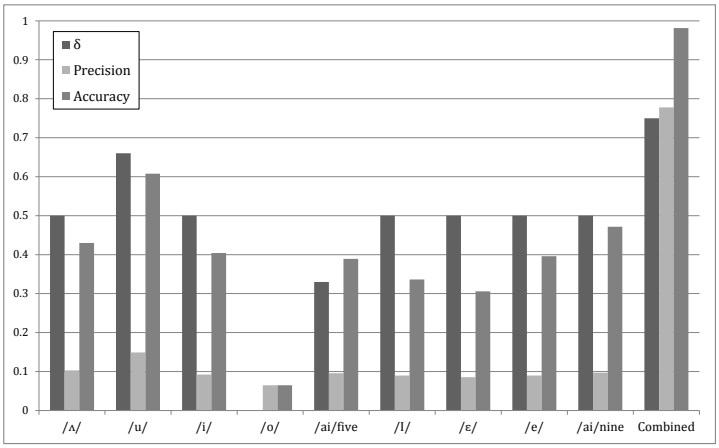

Fig. 2. Performance measurements for each classification trial and minimum δ yielding 100% mTBI recall.

measures for the aggregate trial along with the corresponding δ that achieved 100% recall of mTBI recordings.

Figures 3 - 5 show the recall, precision, and accuracy measurements, respectively, as the value of δ was adjusted in the aggregate trial. It can be seen that as δ increases, recall decreases while precision and accuracy tend to increase.

For the aggregate trial, $\delta = 0.75$ resulted in best accuracy while still recalling all mTBI recordings. A value of $\delta = 0.75$ means that when the classification system encounters a speech recording in which more than 75% of all isolated vowel sound segments are classified mTBI, the entire recording is classified mTBI. This δ was able to recall all seven mTBI recordings with an accuracy of 0.982 and precision of 0.778.

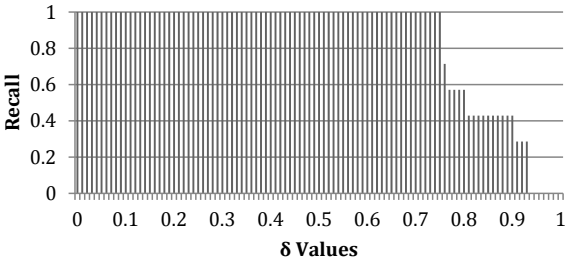

Fig. 3. Recall measurements for increasing values of δ in aggregate vowel sounds trial.

5. CONCLUSIONS AND FUTURE WORK

By using speech analysis on isolated vowel sounds extracted from a mobile application test, the vowel acoustic features that give the best recall and accuracy measures in identifying concussed athletes are identified. In future work various combinations of vowel sounds and acoustic features will be studied to select the most effective δ values. Further noise reduction techniques will be studied and applied to the recordings to give samples that are ideal for extraction of the vowel sounds and features. An implementation of vowel sounds analysis for concussion classification in real time using a cloud-based feedback approach is anticipated in the future. This would help sideline physicians at contact sports to instantly identify suspected concussion cases. A newer test using monosyllabic and multisyllabic words rather than numbers is being developed for this purpose. This test will emphasize words with the vowel sounds and their acoustic features identified as the most successful in assessing concussive behavior in our research.

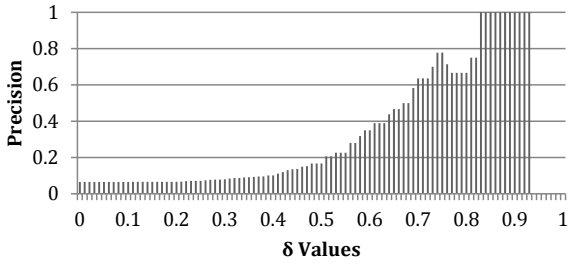

Fig. 4. Precision measurements for increasing values of δ in aggregate vowel sounds trial.

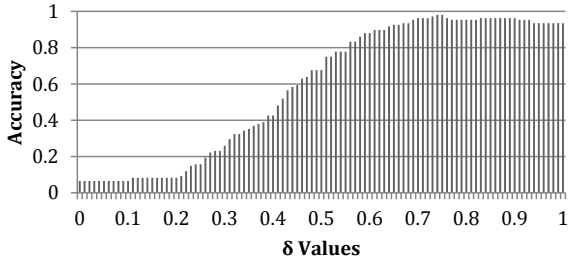

Fig. 5. Accuracy measurements for increasing values of δ in aggregate vowel sounds trial.

6. ACKNOWLEDGMENT

The authors thank Dr. James M. Moriarity, M.D. for providing his medical expertise and support to us in the test administration and data collection effort. We also thank the reviewers for their helpful comments. This work was funded in part by the National Science Foundation with Grant No. CNS-1062743.

7. REFERENCES

[1] L. Zhao, W. Han, and C. Steiner, "Sports related concussions, 2008," Agency for Healthcare Research and Quality (AHRQ), Statistical Report 114, May 2011.

[2] J. Gilchrist, "Nonfatal traumatic brain injuries from sports and recreation activities, United States, 2001–2005," *MMWR MorbMortal Wkly Rep*, vol. 56, no. 29, pp. 733 – 737, 2007.

[3] J. Kelly,"Sports related recurrent brain injuries," *MorbMortal Weekly Report (MMWR)*, vol. 46, pp. 224 – 227, 1997.

[4] New research shows brain injuries from blasts similar to football impacts, 2012. [Online]. Available: http://www.eurekalert.org/features/doe/2012-05/dlnl-nrs052312.php

[5] D. G. Theodoros, B. E. Murdoch, and H. J. Chenery, "Perceptual speech characteristics of dysarthric speakers following severe closed head injury," *Brain Injury*, vol. 8, no. 2, pp. 101 –124, February-March 1994.

[6] W. Ziegler and D. v. Cramon, "Spastic dysarthria after acquired brain injury: An acoustic study," *International Journal of Language & Communication Disorders*, vol. 21, no. 2, pp. 173–187, 1986.

[7] N. K. Madigan, J. DeLuca, B. J. Diamond, G. Tramontano, and A. Averill "Speed of information processing in traumatic brain injury: modality-specific factors," *Journal of Head Trauma Rehabilitation*, vol. 15, pp. 943 – 956, 2000.

[8] A. D. Hinton-Bayre, G. Geffen, and K. McFarland "Mild head injury and speed of information processing: a prospective study of professional rugby league players," *Journal of Clinical and Experimental Neuropsychology*, vol. 19, pp. 275 – 289, 1997.

[9] A. Tsanas, M.A. Little, P.E. McSharry, J. Spielman, and L.O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's Disease," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 5, pp. 1264–1271, 2012.

[10] C. M. Clarke and M. F. Garett, "Mild head injury and speed of information processing: a prospective study of professional rugby league players," *Journal of Clinical and Experimental Neuropsychology*, vol. 116, pp. 3647 – 3658, 2004.

[11] M. H. Davis, I. S.Johnsrude, A. Hervais-Adelman, K. Taylor, and C. McGettigan "Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences," *Journal of Experimental Psychology*, vol. 134, pp. 222 – 240, 2005.

[12] M. O. Krause, "The effects of brain injury and talker characteristics on speech processing in a single-talker interference task," Ph.D. dissertation, University of Minnesota, July 2011.

[13] D. J. Hermes, "Vowel-onset detection," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 866 – 873, 1990.

[14] M. Vasilakis and Y. Stylianou, "Voice pathology detection based eon short-term jitter estimations in running speech," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 3, pp. 153 – 170, 2009.

[15] G. K. Vallabha and B. Tuller, "Systematic errors in the formant analysis of steady-state vowels," *Speech Communication*, vol. 38, no. 1-2, pp. 141 – 160, 2002.

[16] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[17] F. L. Darley, A. E. Aronson and J. R. Brown, " Clusters of deviant speech dimensions in the dysarthrias," *Journal of Speech and Hearing Research*, vol. 12, pp. 462-496, 1969

[18] C. Middag, J. P. Martens, G. van Nuffelen and M. de Bodt, "Automated intelligibility assessment of pathological speech using phonological features EURASIP." *Journal on Applied Signal Processing*, 2009

[19] S. K. Fager, D. R. Beukelman, T. Jakobs and J. P. Hosom, "Evaluation of a speech recognition prototype for speakers with moderate and severe dysarthria: A preliminary report. *Augmentative and Alternative Communication*, vol. 26(4), pp. 267-277, 2010

[20] F. Corinne and P. Gilles "Automatic detection of abnormal zones in pathological speech." *ICPhS XVII*, 2011