Computerized Analysis of a Verbal Fluency Test

James O. Ryan[1], Serguei Pakhomov[1], Susan Marino[1], Charles Bernick[2], and Sarah Banks[2]
[1] College of Pharmacy, University of Minnesota
[2] Lou Ruvo Center for Brain Health, Cleveland Clinic
{ryanx765, pakh0002, marin007}@umn.edu
{bernicc, bankss2}@ccf.org

Abstract

We present a system for automated phonetic clustering analysis of cognitive tests of phonemic verbal fluency, on which one must name words starting with a specific letter (e.g., 'F') for one minute. Test responses are typically subjected to manual phonetic clustering analysis that is labor-intensive and subject to inter-rater variability. Our system provides an automated alternative. In a pilot study, we applied this system to tests of 55 novice and experienced professional fighters (boxers and mixed martial artists) and found that experienced fighters produced significantly longer chains of phonetically similar words, while no differences were found in the total number of words produced. These findings are preliminary, but strongly suggest that our system can be used to detect subtle signs of brain damage due to repetitive head trauma in individuals that are otherwise unimpaired.

1 Introduction

The neuropsychological test of phonemic verbal fluency (PVF) consists of asking the patient to generate as many words as he or she can in a limited time (usually 60 seconds) that begin with a specific letter of the alphabet (Benton et al., 1989). This test has been used extensively as part of larger cognitive test batteries to study cognitive impairment resulting from a number of neurological conditions, including Parkinson's and Huntington's diseases, various forms of dementia, and traumatic brain injury (Troyer et al., 1998a,b; Raskin et al., 1992; Ho et al., 2002). Patients with these disorders tend to generate significantly fewer words on this test than do healthy individuals. Prior studies have also found that clustering (the degree to which patients generate groups of phonetically similar words) and switching (transitioning from one cluster to the next) behaviors are also sensitive to the effects of these neurological conditions.

Contact sports such as boxing, mixed martial arts, football, and hockey are well known for high prevalence of repetitive head trauma. In recent years, the long-term effects of repetitive head trauma in athletes has become the subject of intensive research. In general, repetitive head trauma is a known risk factor for chronic traumatic encephalopathy (CTE), a devastating and untreatable condition that ultimately results in permanent disability and premature death (Omalu et al., 2010; Gavett et al., 2011). However, little is currently known about the relationship between the amount of exposure to head injury and the magnitude of risk for developing these conditions. Furthermore, the development of new behavioral methods aimed at detection of subtle early signs of brain impairment is an active area of research.

The PVF test is an excellent target for this research because it is very easy to administer and has been shown to be sensitive to the effects of acute traumatic brain injury (Raskin and Rearick, 1996). However, a major obstacle to using this test widely for early detection of brain impairment is that clustering and switching analyses needed to detect these subtle changes have to be done manually. These manual approaches are extremely labor-intensive, and are therefore limited in the types of clustering analyses that can be performed. Manual methods are also not scalable to large numbers of tests and are subject to inter-rater variability, making the results difficult to compare across subjects, as well as across different studies. Moreover, traditional manual clustering and switching analyses rely primarily on word orthography to determine phonetic similarity (e.g., by comparing the first two letters of two words), rather than phonetic representations, which would be prohibitively time-

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 884–889,
Sofia, Bulgaria, August 4-9 2013.

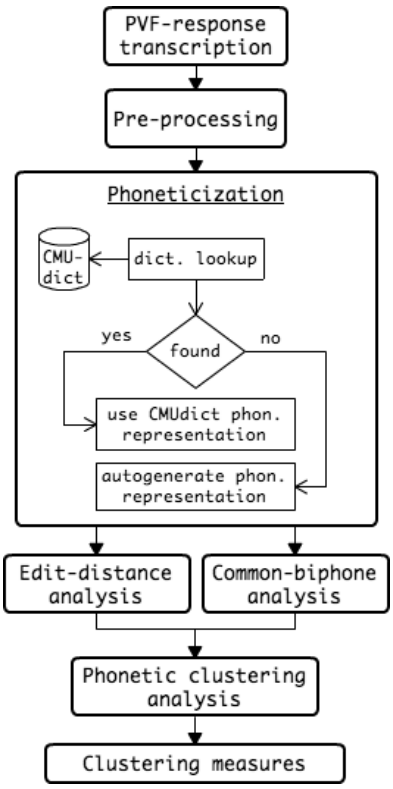


Figure 1: High-level system architecture and workflow.

consuming to obtain by hand.

Phonetic similarity has been investigated in application to a number of research areas, including spelling correction (Toutanova and Moore, 2002), machine translation (Knight and Graehl, 1998; Kondrak et al., 2003), cross-lingual information retrieval (Melamed, 1999; Fujii and Ishikawa, 2001), language acquisition (Somers, 1998), historical linguistics (Raman et al., 1997), and social-media informatics (Liu et al., 2012); we propose a novel clinical application.

Our objective was to develop and pilot-test a relatively simple, but robust, system for automatic identification of word clusters, based on phonetic content, that uses the CMU Pronouncing Dictionary, a decision tree-based algorithm for generating pronunciations for out-of-dictionary words, and two different approaches to calculating phonetic similarity between words.

We first describe the system architecture and our phonetic-similarity computation methods, and then present the results of a pilot study, using data from professional fighters, demonstrating the utility of this system for early detection of subtle signs of brain impairment.

2 Automated Clustering Analysis

Figure 1 shows the high-level architecture and workflow of our system.

2.1 Pronunciation Dictionary

We use a dictionary developed for speech recognition and synthesis applications at the Carnegie Mellon University (CMUdict). CMUdict contains phonetic transcriptions, using a phone set based on ARPABET (Rabiner and Juang, 1993), for North American English word pronunciations (Weide, 1998). We used the latest version, *cmudict.0.7a*, which contains 133,746 entries.

From the full set of entries in CMUdict, we removed alternative pronunciations for each word, leaving a single phonetic representation for each heteronymous set. Additionally, all vowel symbols were stripped of numeric stress markings (e.g., `AH1` → `AH`), and all multicharacter phone symbols were converted to arbitrary single-character symbols, in lowercase to distinguish these symbols from the original single-character ARPABET symbols (e.g., `AH` → `c`). Finally, whitespace between the symbols constituting each phonetic representation was removed, yielding compact phonetic-representation strings suitable for computing our similarity measures.

To illustrate, the CMUdict pronunciation entry for the word *phonetic*, `[F AH0 N EH1 T IH0 K]`, would be represented as `FcNiTmK`.

2.2 Similarity Computation

Our system uses two methods for determining phonetic similarity: edit distance and a common-biphone check. Each of these methods gives a measure of similarity for a pair of phonetic representations, which we respectively call a *phonetic-similarity score* (PSS) and a *common-biphone score* (CBS).

For PSS, we first compute the Levenshtein distance (Levenshtein, 1966) between compact phonetic-representation strings and normalize that to the length of the longer string; then, that value is subtracted from 1. PSS values range from 0 to 1, with higher scores indicating greater similarity. The CBS is binary, with a score of 1 given for two phonetic representations that have a common initial and/or final biphone, and 0 for two strings that have neither in common.

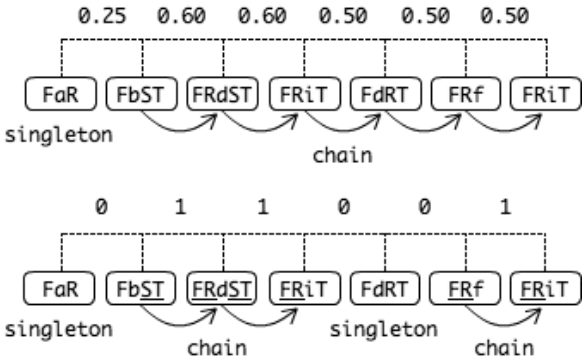


Figure 2: Phonetic chain and common-biphone chain (below) for an example PVF response.

2.3 Phonetic Clustering

We distinguish between two ways of defining phonetic clusters. Traditionally, any sequence of n words in a PVF response is deemed to form a cluster if all pairwise word combinations for that sequence are determined to be phonetically similar by some metric. In addition to this method, we developed a less stringent approach in which we define *chains* instead of clusters.

A chain comprises a sequence for which the phonetic representation of each word is similar to that of the word immediately prior to it in the chain (unless it is chain-initial) and the word subsequent to it (unless it is chain-final). Lone words that do not belong to any cluster constitute *singleton* clusters. We call chains based on the edit-distance method *phonetic chains*, and chains based on the common-biphone method *common-biphone chains*; both are illustrated in Figure 2.

Unlike the binary CBS method, the PSS method produces continuous edit-distance values, and therefore requires a threshold for categorizing a word pair as similar or dissimilar. We determine the threshold empirically for each letter by taking a random sample of 1000 words starting with that letter in CMUdict, computing PSS scores for each pairwise combination ($n = 499,500$), and then setting the threshold as the value separating the upper quintile of these scores. With the common-biphone method, two words are considered phonetically similar simply if their CBS is 1.

2.4 System Overview

Our system is written in Python, and is available online.[1] The system accepts transcriptions of a PVF response for a specific letter and, as a pre-processing step, removes any words that do not begin with that letter. After pre-processing, all words are phoneticized by dictionary lookup in our modified CMUdict. For out-of-dictionary words, we automatically generate a phonetic representation with a decision tree-based grapheme-to-phoneme algorithm trained on the CMUdict (Pagel et al., 1998).

Next, PSSs and CBSs are computed sequentially for each pair of contiguous phonetic representations, and are used in their respective methods to compute the following measures: mean pairwise similarity score (MPSS), mean chain length (MCL), and maximum chain length (MXCL). Singletons are included in these calculations as chains of length 1.

We also calculate equivalent measures for clusters, but do not present these results here due to space limitations, as they are similar to those for chains. In addition to these measures, our system produces a count of the total number of words that start with the letter specified for the PVF test (WCNT), and a count of repeated words (RCNT).

3 Pilot Study

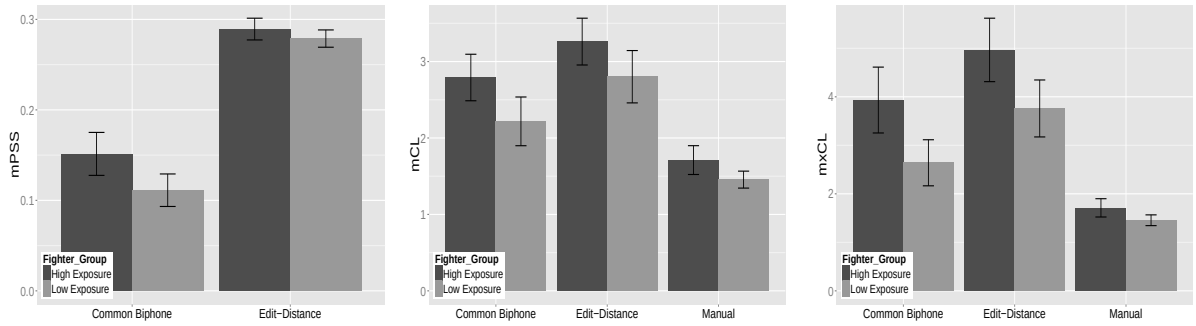
3.1 Participants

We used PVF tests from 55 boxers and mixed martial artists (4 women, 51 men; mean age 27.7 y.o., SD 6.0) that participated in the Professional Fighters Brain Health Study (PFBH). The PFBH is a longitudinal study of unarmed active professional fighters, retired professional fighters, and age/education matched controls (Bernick et al., in press). It is designed to enroll over 400 participants over the next five years. The 55 participants in our pilot represent a sample from the first wave of assessments, conducted in summer of 2012. All 55 participants were fluent speakers of English and were able to read at at least a 4th-grade level. None of these participants fought in a professional or amateur competition within 45 days prior to testing.

3.2 Methods

Each participant's professional fighting history was used to determine his or her total number of pro fights and number of fights per year. These figures were used to construct a composite fight-exposure index as a summary measure of cumulative traumatic exposure, as follows.

[1] http://rxinformatics.umn.edu/downloads.html



(a) Mean pairwise similarity score (b) Mean chain/cluster length (c) Max chain/cluster length

Figure 3: Computation-method and exposure-group comparisons showing significant differences between the low- and high-exposure fighter groups on MPSS, MCL, and MXCL measures. Error bars represent 95% confidence intervals around the means.

Fighters with zero professional fights were assigned a score of 0; fighters with between 1 and 15 total fights, but only one or fewer fights per year, were assigned a score of 1; fighters with 1-15 total fights, and more than one fight per year, got a score of 2; fighters with more than 15 total fights, but only one or fewer fights per year, got a score of 3; remaining fighters, with more than 15 fights and more than one fight per year, were assigned the highest score of 4.

Due to the relatively small sample size in our pilot study, we combined groups with scores of 0 and 1 to constitute the *low-exposure* group ($n = 25$), and the rest were assigned to the *high-exposure* group ($n = 30$).

All participants underwent a cognitive test battery that included the PVF test (letter 'F'). Their responses were processed by our system, and means for our chaining variables of interest, as well as counts of total words and repetitions, were compared across the low- and high-exposure groups. Additionally, all 55 PVF responses were subjected to *manual* phonetic clustering analysis, following the methodology of Troyer et al. (1997). With this approach, clusters are used instead of chains, and two words are considered phonetically similar if they meet any of the following conditions: they begin with the same two orthographic letters; they rhyme; they differ by only a vowel sound (e.g., *flip* and *flop*); or they are homophones.

For each clustering method, the differences in means between the groups were tested for statistical significance using one-way ANOVA adjusted for the effects of age and years of education. Spearman correlation was used to test for associations between continuous variables, due to nonlinearity, and to directly compare manually determined clustering measures with corresponding automatically determined chain measures.

4 Results

The results of comparisons between the clustering methods, as well as between the low- and high-exposure groups, are illustrated in Figure 3.[2]

We found a significant difference ($p < 0.02$) in MPSS between the high- and low-exposure groups using the common-biphone method (0.15 vs. 0.11), while with edit distance the difference was small (0.29 vs. 0.28) and not significant (Figure 3a). Due to infeasibility, MPSS was not calculated manually.

Mean chain sizes determined by the common-biphone method correlated with manually determined cluster sizes more strongly than did chain sizes determined by edit distance (ρ = 0.73, $p < 0.01$ vs. ρ = 0.48, $p < 0.01$). Comparisons of maximum chain and cluster sizes showed a similar pattern (ρ = 0.71, $p < 0.01$ vs. ρ = 0.39, $p < 0.01$).

Both automatic methods showed significant differences ($p < 0.01$) between the two groups in MCL and MXCL, with each finding longer chains in the high-exposure group (Figure 3b, 3c); however, slightly larger differences were observed using the common-biphone method (MCL: 2.79 vs. 2.21 by common-biphone method, 3.23 vs. 2.80 by edit-distance method; MXCL: 3.94 vs. 2.64 by

[2]Clustering measures rely on chains for our automatic methods, and on clusters for manual analysis.

common biphone, 4.94 vs. 3.76 by edit distance). Group differences for manually determined MCL and MXCL were also significant ($p < 0.05$ and $p < 0.02$, respectively), but less so (MCL: 1.71 vs. 1.46; MXCL: 4.0 vs. 3.04).

5 Discussion

While manual phonetic clustering analysis yielded significant differences between the low- and high-exposure fighter groups, our automatic approach, which utilizes phonetic word representations, appears to be more sensitive to these differences; it also appears to produce less variability on clustering measures. Furthermore, as discussed above, automatic analysis is much less labor-intensive, and thus is more scalable to large numbers of tests. Moreover, our system is not prone to human error during analysis, nor to inter-rater variability.

Of the two automatic clustering methods, the common-biphone method, which uses binary similarity values, found greater differences between groups in MPSS, MCL, and MXCL; thus, it appears to be more sensitive than the edit-distance method in detecting group differences. Common-biphone measures were also found to better correlate with manual measures; however, both automated methods disagreed with the manual approach to some extent. The fact that the automated common-biphone method shows significant differences between group means, while having less variability in measurements, suggests that it may be a more suitable measure of phonetic clustering than the traditional manual method.

These results are particularly important in light of the difference in WCNT means between low- and high-exposure groups being small and not significant (WCNT: 17.6, SD 5.1 vs. 18.7, SD 4.7; $p = 0.24$). Other studies that used manual clustering and switching analyses reported significantly more switches for healthy controls than for individuals with neurological conditions (Troyer et al., 1997). These studies also reported differences in the total number of words produced, likely due to investigating already impaired individuals.

Our findings show that the low- and high-exposure groups produced similar numbers of words, but the high-exposure group tended to produce longer sequences of phonetically similar words. The latter phenomenon may be interpreted as a mild form of perseverative (*stuck-in-set*/repetitive) behavior that is characteristic of disorders involving damage to frontal and subcortical brain structures.

To test this interpretation, we correlated MCL and MXCL, the two measures with greatest differences between low- and high-exposure fighters, with the count of repeated words (RCNT). The resulting correlations were 0.41 ($p = 0.01$) and 0.48 ($p < 0.001$), respectively, which supports the perseverative-behavior interpretation of our findings.

Clearly, these findings are preliminary and need to be confirmed in larger samples; however, they plainly demonstrate the utility of our fully automated and quantifiable approach to characterizing and measuring clustering behavior on PVF tests. Pending further clinical validation, this system may be used for large-scale screening for subtle signs of certain types of brain damage or degeneration not only in contact-sports athletes, but also in the general population.

6 Acknowledgements

We thank the anonymous reviewers for their insightful feedback.

References

Atsushi Fujii and Tetsuya Ishikawa. 2001. Japanese/English cross-language information retrieval: Exploration of query translation and transliteration. In *Computers and the Humanities 35.4*.

A.L. Benton, K.D. Hamsher, and A.B. Sivan. 1989. Multilingual aphasia examination.

C. Bernick, S.J. Banks, S. Jones, W. Shin, M. Phillips, M. Lowe, M. Modic. In press. Professional Fighters Brain Health Study: Rationale and methods. In *American Journal of Epidemiology*.

Brandon E. Gavett, Robert A. Stern, and Ann C. McKee. 2011. Chronic traumatic encephalopathy: A potential late effect of sport-related concussive and subconcussive head trauma. In *Clinics in Sports Medicine 30, no. 1*.

Aileen K. Ho, Barbara J. Sahakian, Trevor W. Robbins, Roger A. Barker, Anne E. Rosser, and John R. Hodges. 2002. Verbal fluency in Huntington's disease: A longitudinal analysis of phonemic and semantic clustering and switching. In *Neuropsychologia 40, no. 8*.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady, vol. 10*.

Fei Liu, Fuliang Weng, and Xiao Jiang. 2012. A broad-coverage normalization system for social media language. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics.

Kevin Knight and Jonathan Graehl. 1998. Machine transliteration. In *Computational Linguistics 24.4*.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003*. Association for Computational Linguistics.

I. Dan Melamed. 1999. Bitext maps and alignment via pattern recognition. In *Computational Linguistics 25.1*.

Bennet I. Omalu, Julian Bailes, Jennifer Lynn Hammers, and Robert P. Fitzsimmons. 2010. Chronic traumatic encephalopathy, suicides and parasuicides in professional American athletes: The role of the forensic pathologist. In *The American Journal of Forensic Medicine and Pathology 31, no. 2*.

Vincent Pagel, Kevin Lenzo, and Alan Black. 1998. Letter to sound rules for accented lexicon compression.

Lawrence Rabiner and Biing-Hwang Juang. 1993. Fundamentals of speech recognition.

Anand Raman, John Newman, and Jon Patrick. 1997. A complexity measure for diachronic Chinese phonology. In *Proceedings of the SIGPHON97 Workshop on Computational Linguistics at the ACL97/EACL97*.

Sarah A. Raskin, Martin Sliwinski, and Joan C. Borod. 1992. Clustering strategies on tasks of verbal fluency in Parkinson's disease. In *Neuropsychologia 30, no. 1*.

Sarah A. Raskin and Elizabeth Rearick. 1996. Verbal fluency in individuals with mild traumatic brain injury. In *Neuropsychology 10, no. 3*.

Harold L. Somers. 1998. Similarity metrics for aligning children's articulation data. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 2*. Association for Computational Linguistics.

Kristina Toutanova and Robert C. Moore. 2002. Pronunciation modeling for improved spelling correction. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Angela K. Troyer, Morris Moscovitch, and Gordon Winocur. 1997. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. In *Neuropsychology, 11*.

Angela K. Troyer, Morris Moscovitch, Gordon Winocur, Michael P. Alexander, and Don Stuss. 1998a. Clustering and switching on verbal fluency: The effects of focal frontal- and temporal-lobe lesions. In *Neuropsychologia*.

Angela K. Troyer, Morris Moscovitch, Gordon Winocur, Larry Leach, and Morris Freedman. 1998b. Clustering and switching on verbal fluency tests in Alzheimer's and Parkinson's disease. In *Journal of the International Neuropsychological Society 4, no. 2*.

Robert Weide. 2008. Carnegie Mellon Pronouncing Dictionary, v. 0.7a. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.