HHS Public Access

Author manuscript

*Speech Commun*. Author manuscript; available in PMC 2016 December 01.

Published in final edited form as:

*Speech Commun*. 2015 December 1; 75: 14–26. doi:10.1016/j.specom.2015.09.010.

# Using Automatic Speech Recognition to Assess Spoken Responses to Cognitive Tests of Semantic Verbal Fluency

**Serguei V.S. Pakhomov**[1], **Susan E. Marino**[1], **Sarah Banks**[2], and **Charles Bernick**[2]

[1] Center for Clinical and Cognitive Neuropharmacology, University of Minnesota, Minneapolis, USA

[2] Lou Ruvo Center for Brain Health, Cleveland Clinic

## Abstract

Cognitive tests of verbal fluency (VF) consist of verbalizing as many words as possible in one minute that either start with a specific letter of the alphabet or belong to a specific semantic category. These tests are widely used in neurological, psychiatric, mental health, and school settings and their validity for clinical applications has been extensively demonstrated. However, VF tests are currently administered and scored manually making them too cumbersome to use, particularly for longitudinal cognitive monitoring in large populations. The objective of the current study was to determine if automatic speech recognition (ASR) could be used for computerized administration and scoring of VF tests. We examined established techniques for constraining language modeling to a predefined vocabulary from a specific semantic category (e.g., animals). We also experimented with post-processing ASR output with confidence scoring, as well as with using speaker adaptation to improve automated VF scoring. Audio responses to a VF task were collected from 38 novice and experienced professional fighters (boxing and mixed martial arts) participating in a longitudinal study of effects of repetitive head trauma on brain function. Word error rate, correlation with manual word count and distance from manual word count were used to compare ASR-based approaches to scoring to each other and to the manually scored reference standard. Our study's results show that responses to the VF task contain a large number of extraneous utterances and noise that lead to relatively poor baseline ASR performance. However, we also found that speaker adaptation combined with confidence scoring significantly improves all three metrics and can enable use of ASR for reliable estimates of the traditional manual VF scores.

### Keywords

speech analysis; cognitive testing; automatic speech recognition; verbal fluency; speaker adaptation; confidence scoring

**Corresponding Author:** Serguei Pakhomov, PhD, 308 Harvard St. SE, Minneapolis, MN 55455, Tel: 612-624-1198, pakh0002@umn.edu.

## 1. Introduction

Tests of verbal fluency (VF) (Benton and Hamsher, 1994) are widely used in neurological, psychiatric, mental health, and school settings. There are two main types of VF tests - phonemic and semantic. The phonemic test (PVF) consists of naming as many words as possible in one minute that begin with a letter of the alphabet (e.g., A, F, or S). The semantic test (SVF) consists of naming words belonging to a semantic category (e.g., animals). These tests have been demonstrated to be useful for characterization of cognitive impairment due to a number of conditions including neurodegenerative disease (Henry et al., 2005, 2004; Henry and Crawford, 2004a), psychiatric diagnoses (Henry and Crawford, 2005), developmental disorders (Spek et al., 2009), drug toxicity or metabolic effects (Marino et al., 2012; Witt et al., 2013), as well as impairment due to traumatic brain injury or cardiovascular accidents (Henry and Crawford, 2004b, 2004c). In particular, contact sports such as boxing, mixed martial arts, football, and hockey are particularly well-known for high prevalence of repetitive head trauma which is a major risk factor for chronic traumatic encephalopathy (CTE), a devastating and untreatable condition that ultimately results in permanent disability and premature death (McKee et al., 2013). Athletes with prior exposure to head trauma show significant declines in verbal fluency performance among other types of cognitive impairment (Tremblay et al., 2013).

While clinically useful, VF tests are currently administered manually and are too cumbersome for wide adoption on a large scale in fast-paced and overburdened healthcare systems. Furthermore, manual VF testing is also prone to scoring subjectivity and variability, and cannot be easily self-administered, which limits its applications in large and/or longitudinal investigations of cognitive biomarkers of neurodegenerative disease. One of the goals of developing automated VF testing is to enable easy and non-threatening long-term monitoring of cognitive performance in an attempt to detect early subtle cognitive changes that may warrant a more in-depth clinical assessment. Following the desiderata for computerized VF testing (Kemper and McDowd, 2008), we propose to address existing limitations of VF tests by automating their administration and scoring. Our approach consists of using automatic speech recognition (ASR) technology applied to the speech collected during VF testing to estimate an approximate count of “legitimate” words produced during the VF task.

A typical administration of VF testing includes instructing the subject to restrict his or her responses to the specified stimulus (letter or category) and avoid using proper names. A subject's test-taking behavior is commonly evaluated both by the total number of correct responses produced during the task, and by an analysis of the subject's errors or incorrect responses. Under ideal circumstances, given all correct responses, an individual's speech produced on these tasks would consist of 1-, 2-, or 3-word phrases denoting the concepts relevant to the test (e.g., animals) separated by silent pauses without any repetitions, disfluencies, comments, or other extraneous utterances and noise. However, in reality, response errors and noise are common and subjects’ speech often contains events that should not be included in the calculation of the total test score, which presents a significant challenge for automation (Miller et al., 2013).

Computerized administration and scoring of VF tests is a promising area in which computational linguistic and computerized speech processing approaches can make a significant contribution. While ample evidence exists to show clinical usefulness of manually administered VF tests, in order to extend this body of evidence to automated approaches, it is necessary to demonstrate that automated approaches provide acceptable estimates of the manual VF assessments in a variety of populations and environments (Bauer et al., 2012). Miller and colleagues successfully experimented with using an ASR-driven interactive voice response system to administer and score VF tests (Miller et al., 2013). However, using ASR for automatic scoring of VF tests remains a largely unexplored area and needs to be investigated further. To our knowledge, the use of various ASR techniques such as acoustic speaker adaptation and confidence scoring of ASR output have not been investigated in relation to optimizing automatic VF scoring.

Previous work shows that unsupervised speaker adaptation of acoustic models used in ASR can significantly improve recognition accuracy even with very limited (as little as 11 seconds) amounts of available adaptation data (Leggetter and Woodland, 1995; Wang et al., 2007) and is useful in a number of specific applications including dialogue act segmentation (Kolář et al., 2010) and language tutoring systems (Ohkawa et al., 2009). Similarly to speaker adaptation, the use of confidence scoring in general-purpose ASR applications has also been well documented (see (Jiang, 2005) for a review), although the findings with respect to confidence scoring are mixed. Confidence scoring approaches tend to be highly application specific (Zeljkovic, 1996) and lack a systematic way of determining the most optimal confidence threshold (Bouwman et al., 1999). In the current study, we apply confidence scoring to post-process ASR output from VF responses in the context of using a language model with a highly constrained vocabulary in an attempt to leverage the fact that any speech that does not contain an item from the semantic category of interest would get a lower confidence score and thus may be reliably filtered out. The objective of the current study was to experiment with speaker adaptation and confidence scoring as applied to the specific task of automated VF assessment and to validate these approaches on an ‘animal’ verbal fluency test in a sample of cognitively normal individuals.

## 2. Materials and Methods

### 2.1 Animal name recognition system

Standard ASR approaches (described in the next section in more detail) were used to create a system designed specifically to yield a score that represents how many animal names the speaker was able to produce in response to the VF task. The resulting animal name recognition system consisted of the ASR decoder with a specially trained animal fluency language model and a speaker adapted acoustic model, as well as a set of post-processing filters. The post-processing filters were designed to account for the spurious words that may appear in the raw ASR output. Some of these spurious words represent errors produced by the ASR decoder; however, many of these words represent animal naming errors– non-animal names that were recognized correctly by the ASR decoder. The latter type of errors should not be counted against the accuracy of the ASR system but should be counted against the accuracy of the animal name recognition system. Thus the SVF score estimate produced

by the animal naming recognition system represents the number of words that were both most likely to be correctly recognized by the ASR decoder and most likely to be names of animals. The post-processing filters use confidence scores produced by the ASR decoder and a set of additional output characteristics to produce an estimate of the actual SVF animal fluency score from the raw ASR output, as described in the next section.

### 2.2 Animal name recognition system components

To implement an ASR system designed specifically to process VF responses, we used KALDI, an open source automatic speech recognition toolkit (Povey et al., 2012). Our KALDI-based ASR system relies on an acoustic model and a language model in order to automatically convert the input speech signal to a textual representation. For the current implementation of the system, we trained a speaker-independent acoustic model and a bi-gram statistical language model as described in detail in the following sub-sections.

**Acoustic model—**We trained a speaker-independent acoustic model that consists of a set of Hidden Markov Models (HMMs) that represent 88 base phones occurring in multiple acoustic contexts collected from a large corpus of general English speech. The phone set of 88 phones was derived from the Carnegie Mellon University dictionary (CMU dictionary) of pronunciations and included 84 consonants and vowels with preserved stress marking, a special silence phone ("SIL") and special phones to represent speech noise ("SPN"), non-speech noise ("NSN") and filled pauses ("ah" – FPU, and "um" – FPM). Each of the 84 consonants and vowels were modeled as consisting of three parts: an onset, a nucleus, and a coda, represented by a 4-state HMM network consisting of one non-emitting state and three emitting states, each having a transition to itself and the following state. The other phones (SIL, SPN, NSN, FPU and FPM) were modeled with a 6-state network in which each state could transition to itself or the following two states. The corpus of audio recordings with corresponding verbatim transcriptions used to estimate the parameters of the Gaussian mixtures for the phones appearing in various acoustic contexts consisted of the Wall Street Journal corpus (CSR-II – approximately 78,000 utterances from read and dictated speech of over 240 different speakers) (Paul and Baker, 1992), augmented with spontaneous speech from the TRAINS corpus (approximately 98 dialogs between 34 different speakers – 55,000 words) (Allen et al., 1995).

**Language Model—**We trained a bi-gram language model that represents the distribution of two-word sequences obtained from a corpus of prior clinical and research studies that included verbal fluency testing. This model partially captures the fact that when people respond to the animal verbal fluency task they tend to group animals into categories (Troyer, 2000) thus making some two-word sequences more likely to appear than others. For example, this modeling process captures the fact that the word ALLIGATOR is much more likely to be followed by the word CROCODILE (probability of 0.166 in our data) than the word GOOSE (probability of 0.003). However, prior work has demonstrated that the mean size of semantic clusters is around 1.5, which means that most clusters consist of one or two words. Based on this observation, we limited language modeling to bi-grams. To train the bi-gram model we used a corpus of responses to SVF tests provided by 1,367 participants in a Mayo Clinic Study of Aging (Roberts et al., 2008) that repeated this test up to 7 times (at

the time of training the model) resulting in a total of 6,453 responses (125,601 words). The resulting model consisted of n-gram probability estimates for 990 unigrams, 18,325 bigrams with the Good-Turing discounting smoothing method, as implemented in the open-source CMU Language Modeling Toolkit (Rosenfeld and Clarkson, 1997).

The data used to train the language model consisted of handwritten notes taken by the psychometrists during SVF testing and subsequently converted to electronic format. Thus, these notes do not represent verbatim transcriptions of the speech. Therefore, the language model constructed from these notes does not capture bi-gram distributions of filled pauses or the speech and non-speech noise events. We manually introduced filled pauses and speech and non-speech noise events as unigrams into the model after it was trained by taking some of the probability mass from word tokens and assigning it to these events.

**Speech-to-text Conversion (Decoding)—**The speech signal was pre-processed by splitting it into 25 millisecond frames shifted by 10 milliseconds. Each frame was coded as a set of 13 Mel-spectrum Frequency Cepstral Coefficients (MFCCs) with added delta coefficients, resulting in a vector of 26 coefficients. For each set of MFCC vectors representing the speech input frames, KALDI ASR decoder was used to find the highest likelihood path through the lattice of hypotheses constructed based on the language and acoustic models described above. The output of the decoding process consists of the most likely sequence of phones and words (tokens) corresponding to the input audio signal with the start and end time information for each phone and word.

**Confidence scoring—**In order to obtain confidence scores for the output of the ASR decoder, we used the *lattice-to-ctm* tool in the KALDI toolkit with default parameters to generate maximum a-posteriori scores for each token in the top best hypothesis from the decoder lattices. For the current study, we used a threshold to filter out tokens with lower confidence scores from the decoder output prior to estimating the SVF score. The optimal threshold value was determined with respect to minimizing ASR word error rate with 5-fold cross-validation. Minimum word error rates were found at the threshold value of 0.7 on both training and testing folds.

**Speaker Adaptation—**As part of the cognitive test battery, the participants were asked to read aloud a short paragraph consisting of the first 6 sentences of “The Rainbow Passage” (Fairbanks, 1960), repeated here for convenience:

### The Rainbow Passage

When the sunlight strikes raindrops in the air, they act as a prism and from a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow.

The audio recordings of the participants reading this passage were force-aligned with the text of the passage and subsequently used to perform feature-space Maximum Likelihood

Linear Regression (fMLLR) adaptation of the speaker-independent acoustic model. Consequently, we were able to process each audio sample with a speaker-independent as well as a speaker-adapted set of acoustic models.

**Reliability scoring—**In addition to using confidence scores to filter the ASR output prior to counting words in the speaker's SVF response, we also used the ratio of the number of words with low confidence scores to the total number of unique words produced by the ASR decoder to identify speakers that were likely to produce poor SVF score estimates with the automated approach. Our rationale was that if a large proportion of words in the ASR output were recognized with low confidence, or if the total number of words produced on this test was too high, then we may treat these results with caution and, possibly, exclude these suspect samples from automatic analysis altogether. For the current study we used a reliability threshold of 0.5 (not to be confused with the confidence threshold) as the minimum acceptable ratio of low confidence words, and one standard deviation above the mean SVF score as the threshold for the total number of unique words. Thus, we marked a sample as suspect if more than 50% of the ASR output consisted of low-confidence words (those with confidence scores below the empirically established confidence threshold of 0.7). We also marked samples as suspect if they contained more than 30 words, a number that is substantially higher than what is produced on this test on average by healthy speakers. The intended use of this reliability scoring in a clinical practice scenario is to identify audio samples that may need to be verified by hand. In this study, we evaluated the effect of this strategy on the accuracy of automated scoring for those participants that passed this reliability scoring filter.

### 2.3 Study Participants

The participants in this study were 38 (mean age 28.4, SD 13.5; mean years of education 13.5, SD 2.15; 4 women and 34 men) novice and experienced professional fighters (boxing and mixed martial arts) participating in a longitudinal study of effects of repetitive head trauma on brain function. For the current study, we used baseline recordings of cognitive tests only for those participants that were considered to be cognitively normal at baseline.

### 2.4 Study Design

All participants underwent a series of cognitive tests that included semantic verbal fluency as part of the neuropsychological test battery. Following informed consent, each participant was given a brief description of the tasks he or she was expected to perform during testing. The participants were then seated in a quiet room equipped with a standard hand-held audio recorder. The examiner initiated the test sequence that included reading aloud of “The Rainbow Passage” and the verbal fluency task. For the verbal fluency task, the participants were asked to name all animals they could think of as fast as they could in 60 seconds. The responses were audio recorded and subjected to subsequent automated as well as manual scoring. Automatic scoring was performed using several methods for subsequent comparison. In this study, we experimented with several ways of enhancing ASR accuracy and automatically estimating the SVF scores summarized in Table 1. The baseline ASR system consisted of KALDI decoder with the speaker independent acoustic model and the “animal” language model described above. Enhancements to the baseline system consisted

of using confidence scoring with and without speaker adaptation, and reliability score filtering.

### 2.5 Manual assessment of verbal fluency

All verbal fluency tests collected in this study were manually examined by one scorer who followed standard scoring guidelines to calculate the total number of legitimate animal names excluding repetitions and non-animal words. The guidelines used for scoring instruct the scorer to count all animals, including birds, fish, reptiles, insects, humans, extinct and mythical animals, but not made up animals. Credit may be given for general category terms (e.g., dog) and for specific instances (e.g., terriers) when both are given. However, only one item is to be counted when people name the same animal at different developmental stages (e.g., sheep, lamb). Questionable cases (e.g., counting mythical creatures as valid responses) were discussed and resolved with the first author (SP) to ensure consistency in scoring.

In addition to standard manual scoring of verbal fluency tests, all responses were transcribed verbatim including all erroneous words, repetitions, word fragments, filled pauses (“um's” and “ah's”), noise (e.g., breaths, coughs, lipsmacks, laughter), and comments (i.e., “oh I already said cat”, “I can't think of anything else”, etc.). These verbatim transcripts were used to investigate the rate of disfluent events that could present a potential challenge for automated scoring.

As a result of manual assessments, we were able to measure the performance of both the ASR decoder component and the animal naming recognition system as a whole. The verbatim manual transcripts were used to evaluate the performance of the ASR decoder component separately from the overall system performance, as detailed in the next section. The standard manual SVF scoring was used to evaluate the performance of the animal naming recognition system that included post-processing in addition to the ASR decoder component.

### 2.6 Automatic assessment of verbal fluency

The raw ASR output produced by the system described in Section 2.1 was used to estimate speaker SVF performance as follows. We counted all tokens in the ASR output that were not labeled as silent/filled pauses, noise, and utterance boundaries (SIL, FPU, FPM, NSN, SPN, <s>, </s>). Since the language model was constrained to a closed vocabulary consisting of animal names, the system's output was consequently also constrained to animal names only. Constraining the language model to animal names also had the effect of producing lower confidence scores on input that does not correspond to naming an animal and the confidence score filter leverages that to exclude, to some degree, extraneous speech not relevant to SVF scoring. The variations of the systems that used confidence scoring also excluded all tokens with confidence scores below the threshold of 0.7. The scoring approach is illustrated in Figure 1 showing a side-by-side comparison between a manual verbatim transcript and raw ASR output with confidence scores for a 15 second segment of a VF test audio sample. This example shows minimal differences between automatic and manual scoring despite a number of errors produced by the recognizer (e.g. ELEPHANT recognized as COYOTE and WREN; a filled pause (FPU) recognized as CAT; extraneous words SAME THING

recognized as animal names OXEN and SNAKE). This example also shows that using the confidence scoring threshold can help avoid erroneously counting CAT, OXEN, and SNAKE but it can also mistakenly leave out legitimate words (e.g., ELEPHANT) that happened to be recognized as two words (COYOTE and WREN) with confidence below the threshold.

### 2.7 Statistical Analysis

To compare the performance of various approaches to ASR-based fluency scoring we used the following measures: the standard word error rate (WER), the distance between manual and automatic SVF scores, the correlation between manual and automatic SVF scores (CORR), and the naming error rate (NER). WER was computed with the NIST *sctk* (http://www.itl.nist.gov/iad/mig//tools/) package based on the alignment between the manual transcription (reference) of each audio sample and the ASR output (hypothesis), and was defined as the proportion of all substitutions, deletions, and insertions to the total number of words in the reference. Prior to performing alignment, the words in both the reference and the hypothesis were stemmed to remove plural forms using regular expressions, which resulted in requiring that only the stems of the words be identical to be considered as a correct match. Differences between manual and automatic SVF scores can be positive or negative, depending on whether the automated scoring approach under- or over-estimated true SVF performance. In order to account for this, we report two types of differences – absolute (aDIFF) and signed (sDIFF). Absolute differences were computed as the average of absolute values, whereas signed differences rely on signed positive and negative values. The correlation between manual and automatic scores was estimated by calculating the Spearman rank correlation coefficient. NER was defined as the number of insertions and deletions divided by the total number of words in the raw ASR output. This measure is specific to evaluating category naming recognition performance and is based on the observation that only insertion and deletion errors contribute to measuring naming performance in terms of the total number of valid words that belong to a category (e.g., animals in this case). Differences between various experimental approaches were tested for statistical significance by using the Student's t-test for paired observations. The statistical significance threshold was set at $p < 0.05$. All statistical computations were performed using the R package (R Core Team, 2014).

## 3. Results

In this section, we present the results of comparisons between manual assessments of SVF responses and various combinations of the base ASR decoder components with and without adaptation and confidence scoring. Two types of evaluations are presented. First, we evaluated the animal naming recognition system on verbatim transcripts of the audio samples to get a sense of the ASR decoder performance in general using WER. Subsequently, we also evaluated the system specifically with respect to its ability to correctly estimate the number of valid animal names produced by the speaker on the task using NER.

### 3.1 Word and animal naming accuracy comparisons

Table 2 and Figure 2 show the comparison between the various experimental approaches in terms of the ASR WERs. The boxplot in Figure 2 shows that using speaker adaptation combined with confidence scoring results in a significant reduction in WER as compared to the baseline and to the adaptation alone. It also shows that confidence scoring without adaptation results in a significant improvement in the WER over the baseline.

Table 3 and Figure 3 show the comparison between the experimental approaches in terms of aDIFF and sDIFF in manually and automatically estimated SVF scores. Figure 3 juxtaposes only the sDIFFs between manually and automatically obtained scores with the differences in scores obtained by 8 human raters that participated in an independent study of verbal fluency scoring reliability (Passos et al., 2011).

The smallest aDIFF of 3.1 (95% CI: 2.1, 4.0) between the manually and the automatically computed scores was a result of using a speaker-adapted acoustic model and filtering the ASR output with a confidence threshold. The application of reliability score filtering to identify potentially suspect samples to the results produced with the approach that combined speaker adaptation with confidence score filtering further reduced aDIFF between manual and automatic scores to 2.14 (95% CI: 1.02, 3.26). However, the reliability score filtering also reduced the number of participants that could be scored automatically from 38 to 14 (37%). For the purposes of comparing our study's results with other studies, we also calculated sDIFF between manual and automatically computed scores after adaptation and confidence scoring, resulting in sDIFF of 0.46 (95% CI: 0.32, 0.62).

Correlations (CORR) between the best performing combination of approaches (speaker adaption together with confidence scoring) with manual SVF scores as well as the total number of unique words spoken by the participants are shown in Figure 4. We observe higher correlation between automatically computed SVF scores and the total number of words ($r = 0.81$, $n = 38$) than between automatically computed SVF scores and the manually estimated scores ($r = 0.80$, $n = 38$). The correlation between automatic and manual scores after reliability filtering improved to $r = 0.86$ ($n = 14$).

### 3.2 Extraneous speech and non-speech events

The speech produced on PVF and SVF tests is not completely natural and continuous; however, it also does not consist of only isolated words. In addition to the words spoken in response to the task, the speech samples also contain disfluencies including filled pauses, non-speech and speech noise, word fragments, as well as repetitions, comments, and intrusions (e.g., errors - words spoken on the animal fluency test that do not denote an animal or proper nouns on phonemic fluency tests). In addition to these events generated by the participant, in a typical testing situation, the audio recordings may also contain the speech of the examiner. All of these events can present a challenge to the approach described in this study testing the assumption that the traditional total verbal fluency score can be estimated by simply counting utterances produced by the person taking the test without knowing the content of these utterances. Therefore, we assessed the rate of occurrence of these potentially problematic events in the samples obtained in this study. The

results are illustrated in Figure 5 that shows the variability in the amount of various extraneous speech and non-speech events. For example, speakers 3, 16, and 31 show fairly large differences between the total number of unique words that they uttered on the SVF test and their SVF score. These speakers produced a large number of conversational comments, requests for clarification and other "asides" (e.g., "ok that's it I am done", "can it be variations", " what's another one ... UM dang I'm trying to get my animal- animal planet brain going"). At the other end of the spectrum, there are speakers 10-12, 15, 17, 20, 22-24, 26, 30, 33 that produced hardly any conversational comments.

## 4. Discussion

We have developed and evaluated a fully automated approach to the assessment of performance on standard neuropsychological tests of verbal fluency. Our approach offers significant advantages over traditional manual assessments including objectivity and reproducibility, ease of administration, and the possibility of self-administration. Automation of verbal fluency tests can facilitate wider adoption and standardization of these tests in research and clinical evaluation settings including remote administration over the telephone or the Internet.

### 4.1 Effects of speaker adaptation and confidence scoring on ASR accuracy

We found that both speaker adaptation using "The Rainbow Passage" reading task and filtering raw ASR output with a confidence score threshold improves the accuracy of ASR by the same amount (from 90% WER for baseline to 70% WER with either speaker adaptation or confidence scoring alone). However, the improvement in WER for these two approaches happens for different reasons. Speaker adaptation results in more words that are correctly recognized (see Table 2) but also produces more insertions – words produced by the recognizer that are not found in the manual transcription of the speech. The confidence scoring approach reduces the number of insertions; however, the trade-off is an increased number of deletions – words that are in the manual transcription but were either filtered out from the ASR output or were not in the output initially. An interesting finding of this study is that, when combined, speaker adaption and confidence scoring result in a synergistic interaction – the ASR system that combines these approaches is able to take advantage of the improved accuracy that comes with speaker adaption and at the same time compensate for the additional spurious words by removing them with confidence filtering. This combination also leads to fewer substitutions, which is likely to be due to the better match between the acoustic model and the speaker as a result of speaker adaptation.

### 4.2 Comparison between manual and automatic scores

The results of this study suggest that it is feasible to closely approximate traditional manually obtained verbal fluency scores on the SVF task by using ASR. This study also demonstrates that even very limited speaker adaptation combined with filtering of ASR output with confidence scores leads to significant improvements in the accuracy of automatic measurement of performance on the SVF test. In order to perform speaker adaptation for this study, we had the participants read only 6 sentences (~ 30 seconds). A reading task like this one would be easy to implement on a computer (e.g., a mobile tablet,

laptop, or desktop). The reading task can be administered immediately prior to administering the SVF task without substantially adding to the testing time or imposing a significant additional burden on the person taking or administering the test. In a repeated testing scenario, either the speaker's prior audio samples from the reading task may be used for adaptation or a new reading sample may be collected in order to account for possible seasonal or age-related changes in voice quality.

While we found relatively high correlation between manual and automatic scores, there clearly are individual samples in which there were differences between the scores. In order to determine the causes of these differences, we calculated the frequency of various events unrelated to the SVF performance that may have been erroneously counted towards the automated score estimate, resulting in decreased accuracy. Not surprisingly, the extraneous comments were the biggest contributor to the discrepancies between the manually and automatically estimated SVF scores. We experimented with one way of accounting for the extraneous comments by taking advantage of the fact that the language model constructed for this task consisted of animal names and did not include most of the words used by the speakers to comment on the task itself (unless they were homophonous – e.g., mite vs. might). Thus, non-animal words spoken by the participants that are phonetically distinct from animal names included in the language model are likely to result in lower overall confidence scores produced by the ASR decoder and can subsequently be filtered out.

Manual scoring of verbal fluency tests responses has been previously demonstrated to have very high inter-rater reliability (Moms et al., 1989; Norris et al., 1995); however, we were only able to find one study by Passos et al. (2011) that reported inter-rater agreement in terms of the actual differences in scores in addition to the standard summary agreement statistics (intra-class correlation coefficient, kappa or Pearson correlation). This study was conducted to assess the reliability of manual verbal fluency scoring on 120 test samples scored by 8 individuals. While Passos and colleagues found very high agreement between the scorers (intra-class correlation coefficient ~ 0.98), they also found that the 95% confidence interval included deviations of approximately up to 2 words in both positive and negative directions from perfect agreement. Passos and colleagues also found deviations of up to 6 words outside of the 95% confidence interval. Although the Passos et al. (2011) study was conducted in Portuguese, both English and Portuguese are Indo-European languages and there is no reason to believe that the scoring of SVF responses in these two languages would present significantly different challenges, and thus can be treated as similar for the purposes of estimating the variability in SVF scoring. In the current study, we found that after using speaker adaptation combined with confidence score filtering, the automated approach presented in this paper was able to significantly reduce the discrepancy between manually and automatically estimated SVF scores (sDIFF) from −6.0 at baseline to 0.47 words on average after adaptation and confidence filtering. The finding of this fairly small mean difference between manual and automatic assessments is particularly encouraging in light of the data on inter-rater variability in manual assessments of SVF tests presented by Passos et al. (2011). The comparison of our study results to the Passos et al. study (Passos et al., 2011) shows that we can expect the differences in SVF score estimates obtained by ASR enhanced with acoustic adaptation and confidence scoring to be well within the limits on

variability of human raters manually scoring the test, despite the lower correlation between manual and automatic scores.

Furthermore, the results of the current study also show that if greater accuracy is desired, it can be achieved by performing reliability filtering based on simple heuristics. Reliability filtering does reduce the number of samples that can be automatically scored; however, the scores on these samples are very close to manual scores. While the samples that do not pass the reliability filter would need to be manually verified, this approach would still be useful in a practical setting (e.g., a clinic) by significantly reducing the workload required to score SVF tests.

It is also worth mentioning that some of the practical (e.g., clinical) applications of verbal fluency testing can tolerate some variability in scoring. For example, the mean number of legitimate words produced on the “animal” verbal fluency test by healthy individuals range from 21 for younger persons (50-60 years old) to 18 for older persons (70-80 years old) (Lezak, 2004). The mean score for cognitively impaired individuals with mild Alzheimer's disease dementia is 8.8 (SD 3.9) and 6.8 (SD 3.8) for individuals with moderate Alzheimer's disease dementia. Thus, even with the larger mean aDIFF of 3.1 words from the manual score, the ASR-based approach to verbal fluency scoring may be used as a screening tool; however, its validity for any clinical application would need to the further established on larger samples containing patients from relevant diagnostic categories.

### 4.3 Comparison to previous studies of automatic verbal fluency scoring

A number of studies have demonstrated the utility of using ASR for the assessment of various types of fluency. For example, several automated approaches have been developed towards automated assessment of reading fluency in children and language learners (Bolaños et al., 2013; Cucchiarini et al., 2000); however, automated assessment of generative verbal fluency (semantic and phonemic) has not been extensively studied despite the recognition that computerized methods such as ASR would lend themselves well to this task (Kemper and McDowd, 2008).

The study by Miller et al. (Miller et al., 2013) is one of the few studies we were able to find that is most directly relevant to the work presented in this paper. Miller and colleagues evaluated a telephone-based interactive voice response (IVR) system that was designed to administer and score several common cognitive tests including the SVF test (fruits category). The evaluation was performed on a large set of participants (n = 158) between 65 and 92 years old. The findings of this study with respect to the SVF testing are consistent with the findings of the current study. The mean discrepancy between the IVR-generated scores and those assigned by a clinician in the Miller et al. (Miller et al., 2013) study was −1.26 (95% CI: −1.38, −0.88). In the current study, we found a similarly small discrepancy of 0.47 (95% CI: 0.32, 0.62), despite the differences in modality (telephone vs. hand-held recorder) and category (animals vs. fruits). Both studies also show the negative impact of extraneous speech and non-speech events. Our approach to handling the extraneous events focuses on using confidence scoring to identify speech and non-speech segments that are less likely to represent words produced in response to a specific verbal fluency task (e.g., animals). Another approach is to instruct the participants to refrain from making comments

or from verbal interaction with the examiner; however, this is less desirable because it may result in making the test more awkward and unnatural. Furthermore, having the participants remember to refrain from comments may also alter the psychometric properties of the test by imposing an additional cognitive burden.

In another study, Jimison et al. (Jimison et al., 2008) developed a computer-game based approach to indirectly approximate verbal fluency scores in a home-based cognitive monitoring system. Their approach consists of having the participants compose words from sets of letters presented as part of a computer game. They found that the complexity of the words composed using their approach correlated with traditionally measured verbal fluency performance; however, the word complexity measure accounted for 42% ($r = 0.65$) of the variability in the traditional verbal fluency scores. Fitting a regression model with several other performance characteristics in addition to word complexity improved the correlation with verbal fluency scores but only marginally to account for 46% of the variability ($r = 0.68$). It is difficult to compare our study's results to those reported by Jimison and colleagues due to fundamental differences in the measurements produced; however, a comparison in terms of correlations between manual and automatic scores shows that the ASR-based approach presented in this paper estimates verbal fluency scores more directly and thus is able to account for 64% of the variability ($r = 0.80$) in the traditional verbal fluency scores. Furthermore, after implementing reliability filtering we were able to automatically compute scores for a subset of individuals in which our approach was able to account for 74% ($r = 0.86$) of the variability in the traditional verbal fluency scores.

### 4.4 Limitations

This study has a number of limitations that should be considered in the interpretation of the results. The current study sample consists of younger individuals with no known cognitive impairment but are routinely exposed to repetitive head trauma. These participants are at a higher risk than the general population for developing CTE (McKee et al., 2009; Omalu et al., 2010). However, at the time of testing for this study, none of the participants were diagnosed with CTE or any other neurodegenerative disorder. In order to determine the feasibility of using this technology for the evaluation of more demographically varied and/or clinical populations, further investigation is necessary for its use in the appropriate similar samples.

The collection of audio recordings for this study was not optimized to produce the highest quality audio. While we believe that better ASR results may be achieved with higher quality equipment optimally positioned in front of the speaker's mouth (e.g., head-word Sennheiser ME-3 microphone), the current study represents a more “realistic” scenario in which cognitive testing is typically performed. Nonetheless, further experimentation is necessary to determine the effect of equipment position and quality on the accuracy of automatic scoring because any such impact will be mediated by the necessity to filter out erroneous words and comments regardless of how well they are recognized by the ASR decoder.

We used a very basic approach to language modeling for this particular study. We created a bi-gram language with Good-Turing discounting as a way of smoothing n-gram counts. Other smoothing methods may yield better results. It would also be interesting to experiment

with class-based modeling to represent various semantic subcategories of animals to better reflect clustering and switching behavior during the performance of verbal fluency tests. It may also be beneficial to use more sophisticated confidence-scoring methods that, for example, rely on differences in scores obtained from multiple ASR passes with different acoustic models. Furthermore, due to relatively small sample size, we did not attempt to optimize the confidence threshold on an independent dataset and instead used cross-validation. Thus, the results obtained with the current threshold of 0.7 may not readily generalize to other datasets and may need to be further verified.

The participants in this study were all native or near-native speakers of English and were asked to perform the test in English. Thus our study's results may not readily generalize to English language learners performing this test in English or speakers of other languages performing the test in their native language. Our current findings as well as prior research suggest that use of speaker adaptation is going to be beneficial in multi-lingual applications and for accented speech; however, further validation of these assumptions is required.

The current study only examined one semantic category – animals; therefore, our study's results may not generalize to other categories typically used in verbal fluency testing. However, the animal category is used most frequently, which is why we focused on this category first.

Our approach to automating VF testing and administration assumes that the person being tested is being cooperative and is not consciously trying to cheat. The only defense mechanism that the current approach has against possible deception is the reliability filtering that can help flag samples with suspiciously high number of words (possibly indicating reading) or low mean confidence scores (possibly indicating use of nonsense words). We do not consider this to be an effective mechanism to identify deception, which limits the usefulness of this approach in law enforcement, as well as certain psychiatric and educational performance testing settings. Similarly, the technology presented here would need to be further validated on other clinical groups including those with different forms of aphasias that may negatively influence the accuracy of the ASR component.

## 5. Future Directions

In this paper, we reported the results of a basic approach to automated scoring of verbal fluency tests with a focus on producing an estimate that is as close as possible to the manually calculated score. However, this is just one of the possibilities that computerized cognitive assessment can offer. ASR technology could potentially be used for more detailed semantic-level analyses such as measuring the size of semantic clusters and the frequency of switching between semantic clusters. However, the current implementation of the ASR-based system is not designed for this task and, thus, the relatively high WER of 56%, even with the best-performing combination of acoustic adaptation and confidence scoring, precludes any semantic-level analysis. We believe that we can significantly reduce the WER by implementing a combination of approaches consisting of language modeling, recording equipment, and speaker behavior modification. We have conducted another study with healthy volunteers in which we used the Sennheiser ME-3 microphone to collect higher

quality audio combined with automated audio instructions for the participants. The latter approach effectively removes the test administrator from the immediate testing environment and thus eliminates the opportunity for participants to engage in a dialogue with the administrator, which in turn results in greatly reducing the amount of extraneous speech. In this subsequent study, we also plan to introduce a background language model consisting of typical patterns representative of extraneous speech in order to better separate countable responses from speech that needs to be ignored during scoring. Furthermore, it may be beneficial to apply standard keyword spotting approaches to this task in an attempt to eliminate extraneous speech as well as capture some of the relevant test characteristics, such as repeated words.

## 6. Conclusions

This study shows that using ASR in conjunction with minimal speaker adaptation and confidence filtering produces VF score estimates that are very close to human assessments. We found that speaker adaptation and confidence filtering need to be used in tandem in order to maximize each other's benefits – neither technique by itself appears to produce most optimal results. We also found that additional classification of ASR output in terms of how likely it is to be accurate based on simple heuristics can help in identifying speakers that can be scored with the automated system without supervision. These findings have practical implications for implementing automated VF scoring in clinical and educational settings. Implementing “The Rainbow Passage” reading task as a way to perform speaker enrollment for acoustic adaptation would require minimal effort and additional time – a total of about 1 minute. Both the reading task and the VF task can be performed as part of an automated sequence on a tablet computer with no additional time required on the part of the staff administering the test. Further investigation is required, however, in order to validate automated VF testing in larger cross-sectional and longitudinal samples of healthy individuals, as well as, various clinical populations.

## ACKNOWLEDGEMENTS

Work on the automated verbal fluency assessment system was supported in part by grants from the National Institutes of Health (NINDS - 5R01NS076665) and the Alzheimer's Association (DNCFI-12-242985). We also would like to thank James Ryan for helping with manual scoring of the audio responses on the VF task.

## REFERENCES

Allen, J.; Heeman, PA. Linguistic Data Consortium. TRAINS Spoken Dialog Corpus. University of Rochester, Department of Computer Science; 1995.

Bauer RM, Iverson GL, Cernich AN, Binder LM, Ruff RM, Naugle RI. Computerized Neuropsychological Assessment Devices: Joint Position Paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. Archives of Clinical Neuropsychology. 2012; 27:362–373. doi:10.1093/arclin/acs027. [PubMed: 22382386]

Benton, A.; Hamsher, K. Multilingual Aphasia Examination. 3rd ed.. Iowa City, IA.: 1994.

Bolaños D, Cole RA, Ward WH, Tindal GA, Hasbrouck J, Schwanenflugel PJ. Human and automated assessment of oral reading fluency. Journal of Educational Psychology. 2013; 105:1142–1151. doi: 10.1037/a0031479.

Bouwman G, Sturm J, Boves L. Incorporating confidence measures in the Dutch train timetable information system developed in the ARISE project. IEEE. 1999; 1:493–496. doi:10.1109/ICASSP.1999.758170.

Cucchiarini C, Strik H, Boves L. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. The Journal of the Acoustical Society of America. 2000; 107:989. doi:10.1121/1.428279. [PubMed: 10687708]

Henry JD, Crawford J. A meta-analytic review of verbal fluency deficits in schizophrenia relative to other neurocognitive deficits. Cognitive Neuropsychiatry. 2005; 10:1–33. doi:10.1080/13546800344000309. [PubMed: 16571449]

Henry JD, Crawford JR. Verbal fluency deficits in Parkinson's disease: A meta-analysis. Journal of the International Neuropsychological Society. 2004a; 10:608–622. doi:10.1017/S1355617704104141. [PubMed: 15327739]

Henry JD, Crawford JR. A Meta-Analytic Review of Verbal Fluency Performance Following Focal Cortical Lesions. Neuropsychology. 2004b; 18:284–295. doi:10.1037/0894-4105.18.2.284. [PubMed: 15099151]

Henry JD, Crawford JR. A Meta-Analytic Review of Verbal Fluency Performance in Patients With Traumatic Brain Injury. Neuropsychology. 2004c; 18:621–628. doi:10.1037/0894-4105.18.4.621. [PubMed: 15506829]

Henry JD, Crawford JR, Phillips LH. A Meta-Analytic Review of Verbal Fluency Deficits in Huntington's Disease. Neuropsychology. 2005; 19:243–252. doi:10.1037/0894-4105.19.2.243. [PubMed: 15769208]

Henry JD, Crawford JR, Phillips LH. Verbal fluency performance in dementia of the Alzheimer's type: a meta-analysis. Neuropsychologia. 2004; 42:1212–22. doi:10.1016/j.neuropsychologia.2004.02.001 S0028393204000296 [pii]. [PubMed: 15178173]

Jiang H. Confidence measures for speech recognition: A survey. Speech Communication. 2005; 45:455–470. doi:10.1016/j.specom.2004.12.004.

Jimison H, Pavel M, Le T. Home-based cognitive monitoring using embedded measures of verbal fluency in a computer word game. IEEE. 2008:3312–3315. doi:10.1109/IEMBS.2008.4649913.

Kemper, S.; McDowd, JM. Handbook of Cognitive Aging: Interdisciplinary Perspectives. SAGE Publications, Inc.; 2455 Teller Road, Thousand Oaks California 91320 United States: 2008. Dimensions of Cognitive Aging: Executive Function and Verbal Fluency; p. 181-192.

Kolář J, Liu Y, Shriberg E. Speaker adaptation of language and prosodic models for automatic dialog act segmentation of speech. Speech Communication. 2010; 52:236–245. doi:10.1016/j.specom.2009.10.005.

Leggetter CJ, Woodland PC. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech & Language. 1995; 9:171–185. doi:10.1006/csla.1995.0010.

Lezak, MD. Neuropsychological Assessment. 4th ed.. Oxford University Press; Oxford, England: 2004.

Marino SE, Pakhomov SVS, Han S, Anderson KL, Ding M, Eberly LE, Loring DW, Hawkins-Taylor C, Rarick JO, Leppik IE, Cibula JE, Birnbaum AK. The effect of topiramate plasma concentration on linguistic behavior, verbal recall and working memory. Epilepsy & Behavior. 2012; 24:365–372. doi:10.1016/j.yebeh.2012.04.120. [PubMed: 22658432]

McKee AC, Cantu RC, Nowinski CJ, Hedley-Whyte ET, Gavett BE, Budson AE, Santini VE, Lee H-S, Kubilus CA, Stern RA. Chronic Traumatic Encephalopathy in Athletes: Progressive Tauopathy After Repetitive Head Injury. Journal of Neuropathology and Experimental Neurology. 2009; 68:709–735. doi:10.1097/NEN.0b013e3181a9d503. [PubMed: 19535999]

McKee AC, Stein TD, Nowinski CJ, Stern RA, Daneshvar DH, Alvarez VE, Lee H-S, Hall G, Wojtowicz SM, Baugh CM, Riley DO, Kubilus CA, Cormier KA, Jacobs MA, Martin BR, Abraham CR, Ikezu T, Reichard RR, Wolozin BL, Budson AE, Goldstein LE, Kowall NW, Cantu RC. The spectrum of disease in chronic traumatic encephalopathy. Brain. 2013; 136:43–64. doi:10.1093/brain/aws307. [PubMed: 23208308]

Miller DI, Talbot V, Gagnon M, Messier C. Administration of Neuropsychological Tests Using Interactive Voice Response Technology in the Elderly: Validation and Limitations. Frontiers in Neurology 4. 2013 doi:10.3389/fneur.2013.00107.

Moms JC, Heyman A, Mohs RC, Hughes JP, van Belle G, Fillenbaum G, Mellits ED, Clark C. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. Neurology. 1989; 39:1159–1159. doi: 10.1212/WNL.39.9.1159. [PubMed: 2771064]

Norris M, Blankenship-Reuter L, Snow-Turek L, Finch J. Influence of depression on verbal fluency performance. Aging, Neuropsychology, and Cognition. 1995; 2:206–215.

Ohkawa Y, Suzuki M, Ogasawara H, Ito A, Makino S. A speaker adaptation method for non-native speech using learners' native utterances for computer-assisted language learning systems. Speech Communication. 2009; 51:875–882. doi:10.1016/j.specom.2009.05.005.

Omalu BI, Bailes J, Hammers JL, Fitzsimmons RP. Chronic Traumatic Encephalopathy, Suicides and Parasuicides in Professional American Athletes: The Role of the Forensic Pathologist. The American Journal of Forensic Medicine and Pathology. 2010; 31:130–132. doi:10.1097/PAF. 0b013e3181ca7f35. [PubMed: 20032774]

Passos V, Giatti L, Barret S, Figueiredo R, Caramelli P, Bensenor I, Fonseca M, Cade N, Goulart A, Nunes M, Alves M, Trindade A. Verbal fluency tests reliability in a Brazilian multicentric study, ELSA-Brasil. Arquivos de Neuro-Psiquiatria. 2011; 69:814–816. [PubMed: 22042187]

Paul DB, Baker JM. The design for the wall street journal-based CSR corpus. Association for Computational Linguistics. 1992:357. doi:10.3115/1075527.1075614.

Povey, D.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannermann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; Silovsky, J.; Stemmer, G.; Vesely, K. The Kaldi Speech Recognition Toolkit.. Presented at the IEEE Automatic Speech Recognition and Understanding Workshop; Honolulu, HI.. 2012.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2014.

Roberts RO, Geda YE, Knopman DS, Cha RH, Pankratz VS, Boeve BF, Ivnik RJ, Tangalos EG, Petersen RC, Rocca WA. The Mayo Clinic Study of Aging: Design and Sampling, Participation, Baseline Measures and Sample Characteristics. Neuroepidemiology. 2008; 30:58–69. doi: 10.1159/000115751. [PubMed: 18259084]

Rosenfeld, R.; Clarkson, P. Statistical Language Modeling Using the CMU-Cambridge Toolkit. ESCA Eurospeech. Presented at the Eurospeech; RHODES, GREECE. 1997.

Spek A, Schatorjé T, Scholte E, van Berckelaer-Onnes I. Verbal fluency in adults with high functioning autism or Asperger syndrome. Neuropsychologia. 2009; 47:652–656. doi:10.1016/ j.neuropsychologia.2008.11.015. [PubMed: 19084028]

Tremblay S, De Beaumont L, Henry LC, Boulanger Y, Evans AC, Bourgouin P, Poirier J, Theoret H, Lassonde M. Sports Concussions and Aging: A Neuroimaging Investigation. Cerebral Cortex. 2013; 23:1159–1166. doi:10.1093/cercor/bhs102. [PubMed: 22581847]

Troyer AK. Normative data for clustering and switching on verbal fluency tasks. J Clin Exp Neuropsychol. 2000; 22:370–8. [PubMed: 10855044]

Wang S, Cui X, Alwan A. Speaker Adaptation With Limited Data Using Regression-Tree-Based Spectral Peak Alignment. IEEE Transactions on Audio, Speech and Language Processing. 2007; 15:2454–2464. doi:10.1109/TASL.2007.906740.

Witt J-A, Elger CE, Helmstaedter C. Impaired verbal fluency under topiramate - evidence for synergistic negative effects of epilepsy, topiramate, and polytherapy. European Journal of Neurology. 2013; 20:130–137. doi:10.1111/j.1468-1331.2012.03814.x. [PubMed: 22827489]

Zeljkovic I. Decoding optimal state sequence with smooth state likelihoods. IEEE. 1996:129–132. doi: 10.1109/ICASSP.1996.540307.

## Highlights

- We evaluated an ASR-based system for automatic scoring of verbal fluency tests
- Combination of confidence scoring and speaker adaptation result in improved scoring over baseline
- Automated scoring is comparable to manual scoring

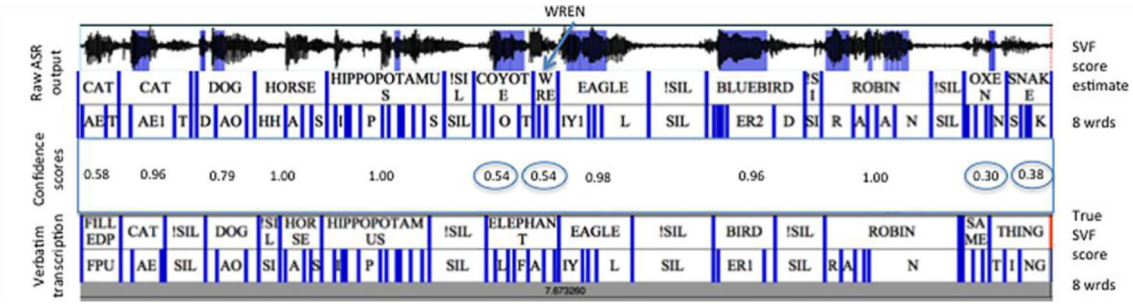
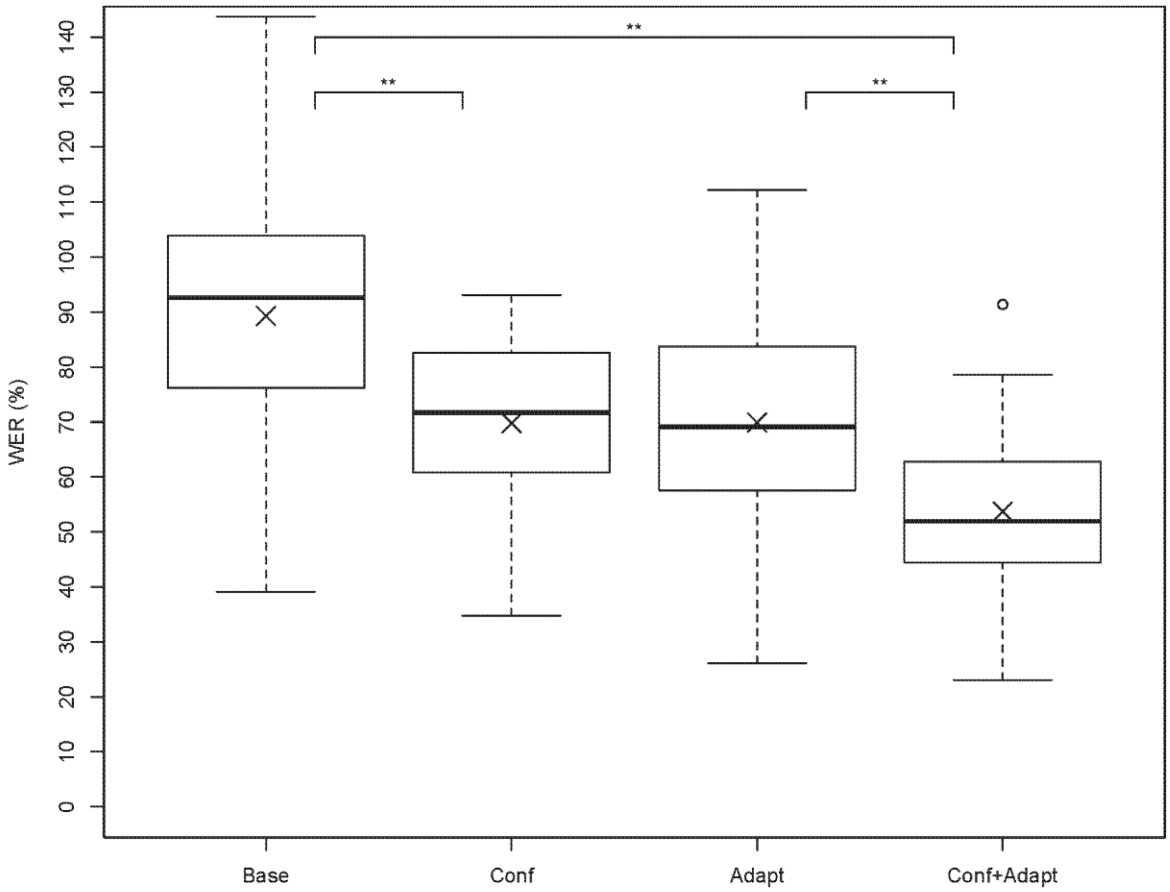



**Figure 1.**

An example showing a comparison between manual and automatic SVF scoring. Raw ASR output from a system using speaker adaptation and confidence scoring (upper panel) is compared to manual verbatim transcription (lower panel). Confidence scores shown inside oval shapes are below the threshold of 0.7. The rendering was generated using Praat.




** indicates a difference significant at alpha < 0.01

x indicates group means

- indicates group medians

**Figure 2.**
Boxplot of differences in WER means between experimental approaches.

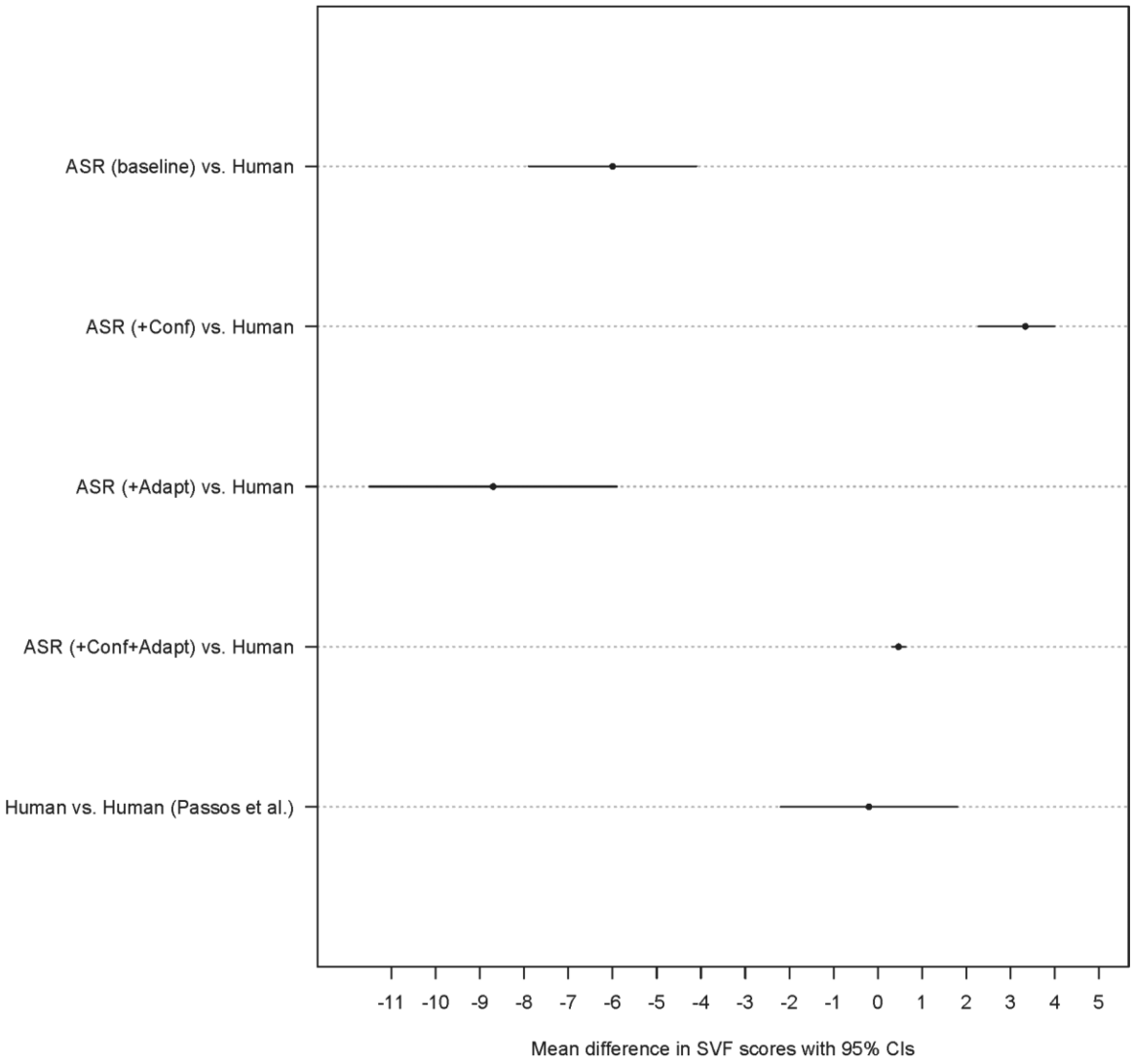



**Figure 3.**

Mean signed differences (sDIFF) between manual and automatic SVF scores produced by different experimental approaches as compared to differences between scores determined by human raters that participated in the Passos et al. (2011) study.

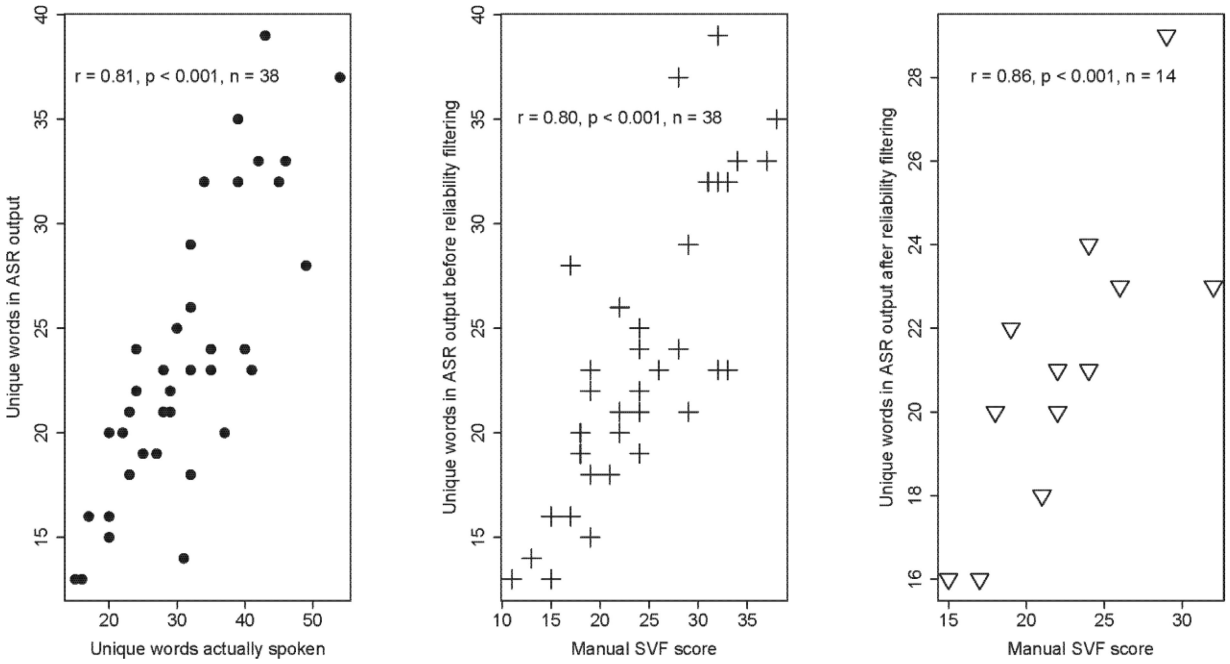



**Figure 4.**

Comparison between SVF score estimates based on ASR output after adaptation and confidence filtering with manual counts of permissible words (manual SVF score).

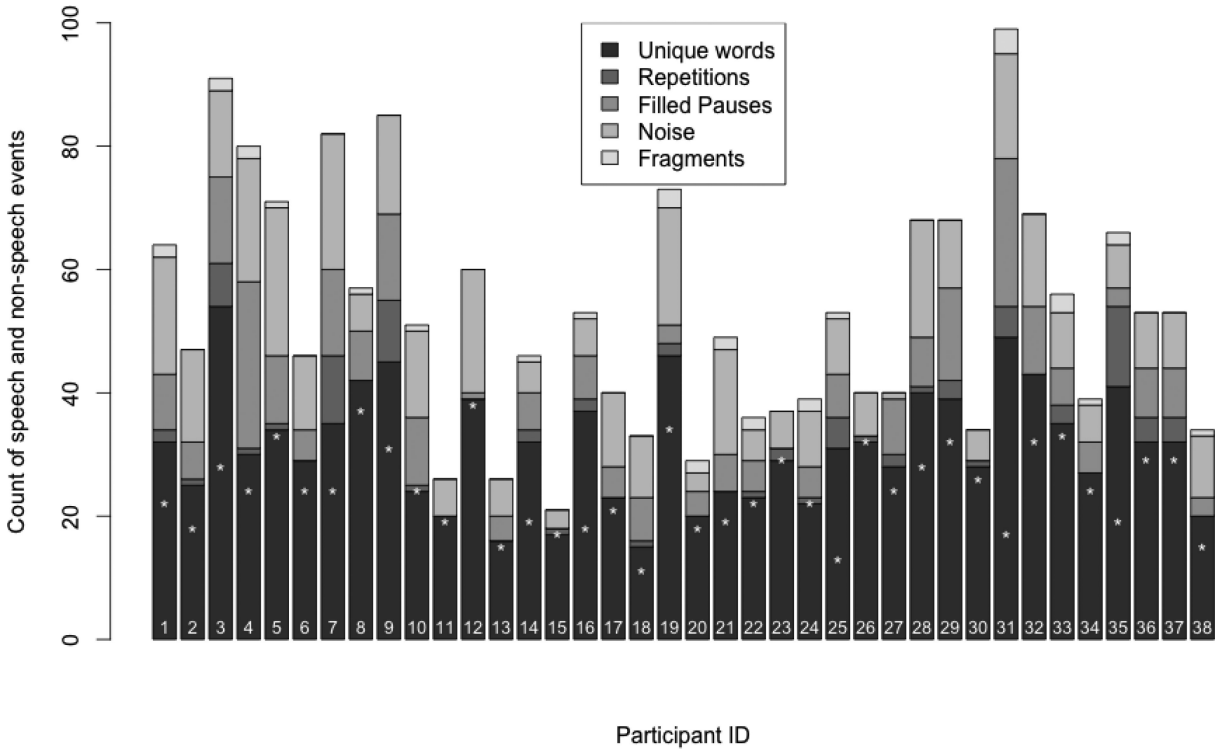



**Figure 5.**

Distribution of various speech and non-speech events in the manual verbatim transcription of each participant's SVF response relative to the manually determined SVF score value (marked with the white asterisk symbol).

**Table 1**

Summary of experimental approaches to estimating the raw SVF scores

| Approach | Accoustic Model | Speaker Adaptation | Confidence Scoring |
|---|---|---|---|
| **Baseline** | SI* | NO | NO |
| **+Confidence** | SI | NO | YES |
| **+Adaptation** | SA | YES | NO |
| **+Confidence+Adaptation** | SA | YES | YES |

*SI – speaker independent; SA – speaker adapted

## Table 2

Comparison of errors in the ASR output across various methods for automatically estimating the SVF scores averaged across speakers.

| | Corr.* (N words) | Ins. (N words) | Del. (N words) | Sub. (N words) | Word Error Rate (WER %) | Name Error Rate (NER %) |
|---|---|---|---|---|---|---|
| **N=38** | **Mean (SD)** | **Mean (SD)** | **Mean (SD)** | **Mean (SD)** | **Mean (SD)** | **Mean (SD)** |
| **Baseline** | 12.7 (5.6) | 9.2 (5.9) | 3.2 (4.9) | 18.6 (11.4) | 89 (21) | 38 (19) |
| **+Confidence** | 11.9 (5.8) | 2.0 (2.7) | 10.7 (7.7) | 11.9 (7.9) | 70 (16) | 36 (12) |
| **+Adaptation** | 17.5 (5.0) | 8.2 (4.9) | 2.6 (3.5) | 14.5 (10.0) | 70 (17) | 32 (13) |
| **+Confidence +Adaptation** | 16.5 (5.1) | 1.7 (2.1) | 9.1 (7.0) | 8.9 (6.8) | 53 (16) | 30 (12) |

*Corr. – count of correctly recognized words averaged across all amples; Ins. – insertions; Del. – deletions; Sub. – substitutions

**Table 3**

Comparison of differences between manually and automatically estimated SVF scores averaged across speakers.

| | Manual SVF score (N words) | Automatic SVF score (N words) | Manual-Auto absolute difference (aDIFF) | Manual-Auto signed difference (sDIFF) |
|---|---|---|---|---|
| | Mean (SD) | Mean (SD) | Mean (95% CI) | Mean (95% CI) |
| **Baseline** | 24.2 (6.9) | 30.2 (8.1) | 7.3 (5.0, 9.7) | −6.0 (−4.1, −7.9) |
| **+Confidence** | 24.2 (6.9) | 20.9 (6.8) | 5.5 (3.8, 7.3) | 3.3 (2.3, 4.4) |
| **+Adaptation** | 24.2 (6.9) | 32.9 (9.5) | 8.8 (6.0, 11.6) | −8.7 (−5.9, −11.5) |
| **+Confidence+Adaptation** | 24.2 (6.9) | 23.7 (6.8) | 3.1 (2.1, 4.0) | 0.47 (0.32, 0.62) |