DAVE: Detecting Agitated Vocal Events

Asif Salekin[1], Hongning Wang[1], Kristine Williams[2], John Stankovic[1]
[1] University of Virginia, [2]University of Kansas

***Abstract*—DAVE is a comprehensive set of event detection techniques to monitor and detect 5 important verbal agitations: asking for help, verbal sexual advances, questions, cursing, and talking with repetitive sentences. The novelty of DAVE includes combining acoustic signal processing with three different text mining paradigms to detect verbal events (asking for help, verbal sexual advances, and questions) which need both lexical content and acoustic variations to produce accurate results. To detect cursing and talking with repetitive sentences we extend word sense disambiguation and sequential pattern mining algorithms. The solutions have applicability to monitoring dementia patients, for online video sharing applications, human computer interaction (HCI) systems, home safety, and other health care applications. A comprehensive performance evaluation across multiple domains includes audio clips collected from 34 real dementia patients, audio data from controlled environments, movies and Youtube clips, online data repositories, and healthy residents in real homes. The results show significant improvement over baselines and high accuracy for all 5 vocal events.**

I. Introduction

According to the National Center for Assisted Living, more than 735,000 people in the USA reside in assisted living facilities [2]. Most of the assisted living facilities rely on the nursing staff and caregivers to monitor and record actions of their patients. One of the most important events to monitor is agitation. A recent study showed that approximately 30% - 50% of patients with cognitive disorder (dementia) suffer from various forms of agitation [11], [19]. The value of accurately identifying agitation is that we can detect dementia early, improve care, and initiate procedures to keep people safe.

The medical community has defined the Cohen-Mansfield Agitation Inventory [10] which specifies approximately 28 agitated behaviors for identifying whether a person is suffering from agitation. In this paper we describe DAVE, an automated comprehensive set of techniques for real time monitoring and recording the 5 most important of the vocal agitation metrics of the Cohen-Mansfield Inventory. This includes cursing, constant unwarranted request for help, making verbal sexual advances, asking constant questions and talking with repetitive sentences.

There are several ways, in which a verbal event can be conveyed. However, in detection of verbal events from speech two factors are important: the choice of words and acoustic variation. When a speaker expresses a verbal event while adhering to an inconspicuous intonation pattern, listeners can nevertheless perceive the information through the lexical content (i.e. words). On the other hand, some verbal event conveying sentence structures share the same lexical representation with other general statements. For example, detecting asking for help or verbal sexual advances using only textual inference or only acoustic features results in high false positives and false negatives. If we try to detect asking for help using only textual features (e.g., using similarity based text analysis and content matching), we can mistakenly identify a story about helping a kid or a discussion about helping others as asking for help. On the other hand, relying only on acoustic signal processing (e.g., temporal pattern mining in the acoustic signal) cannot recognize the situation where people do not depict any specific verbal tone while asking for help, i.e., one might ask for help in a submissive tone or in a dominant tone based on his/her mood. The main contributions of this paper are:

- An automatic and comprehensive set of techniques id developed for detecting 5 verbal agitations based on both extending various algorithms and combining acoustic signal processing with three different text mining paradigms.
- None of the previous state of the art solutions has addressed: asking for help and verbal sexual advances. In this paper we are the first to show that detection of these two vocal events depends both on the acoustic signal processing and the semantics of the speech. To understand the semantics of speech we employ statistical text data mining techniques. Using such a combined feature set we achieve a detection accuracy of 93.45% for asking for help and a detection accuracy of 91.69% for verbal sexual advances.
- Cursing is difficult to detect because many such words have multiple meanings. We have used a modified version of the adapted Lesk algorithm [8] which considers a word's sense, to detect curse words with multiple ambiguous meanings. Using this approach we have detected cursing with 95.6% accuracy.
- We are the first to evaluate a large combination of acoustic, tf-idf and language model features to detect questions from English speech data, and achieved 89.68% accuracy.
- Repetitive sentences from an agitated patient are not precisely repetitive. We have addressed the issue of skipping or adding multiple words in sentences by using a modified version of the prefixSpan algorithm [20] and achieved 100% accuracy.
- We have evaluated DAVE on 34 real agitated elderly (age varies from 63 to 98 years) dementia patients across 16 different nursing homes and achieved 90%, 88.1%, 94% and 100% precision for verbal events: asking for help, questions, cursing and asking repetitive sentences, respectively. Here we solve the challenge that dementia patients mumble, speak in low volume and don't articulate words well. (Section V).
- To show it's generalizability to different domains and for the healthy population, we have evaluated DAVE on movies, *Youtube* clips, the Tatoeba website speech clips [4] with acted and real vocal events, using audio clips from controlled experiments and from real homes. We show accuarcy in the 90-100% range. (Section IV).

II. Related Work

Previous works on questions detection considered both textual features and acoustic features. Since, verbal questions detection is a language specific problem several studies from different languages have explored different acoustic features. Pitch, energy, duration and the fundamental frequency have been explored to detect French questions [18], [22], where energy and fundamental frequency were used as features to detect Arabic questions from speech. A recent study [25] has detected English questions with 87.1% precision using pitch, energy and the fundamental frequency.

To utilize the linguistic content of speech some recent studies used textual features in addition to acoustic features for questions detection. [13] combined acoustic features with key words. Unigram, bigram and trigrams, start and end utterance tags, parse tree representation of syntax, etc. have been used as textual features in addition to acoustic features in questions detection from English, French and Vietnamese utterances [9], [14], [21]. A recent study [18] on French questions detection has combined language model features extracted from speech text with acoustic features (duration, energy and pitch) with 75% accuracy.

Bag-of-word features (tf-idf and language model) has not been evaluated for questions detection from English speech utterances. According to our knowledge we are the first to combine acoustic features with unigram and bigram bag-of-word features from speech content to detect questions from English utterances.

According to our knowledge we are the first to detect verbal events: asking for help and verbal sexual advances. There has been a work on English curse detection from twitter data [26], which detects cursing using predefined key (curse) word matching. Hence, this work can not address the challenge of ambiguous curse word detection with multiple meanings. According to our knowledge we are the first to detect curse words with multiple ambiguous meanings from English speech utterances.

There are noninvasive systems used by physicians, that monitor agitated behaviors in dementia patients [16]. There is research that attempts to detect complex agitated behaviors using video data [12]. Also, research has been done for the diagnosis of mild cognitive impairment (MCI) or early dementia using audio-recorded cognitive screening (ARCS) [24]. There is research to detect agitated physical behavior using the skeleton data collected from the Kinect sensor. The focus of that research was to identify agitated activities such as kicking, punching, and pushing [17]. We are the first to detect 5 agitated verbal behaviors from speech and evaluated on real agitated elderly suffering from dementia.

III. Design of DAVE

DAVE consists of two categories of solutions: detection of *(asking for help, verbal sexual advances and questions)* uses a combination of acoustic and bag-of-word textual features and detection of *(cursing and using repetitive sentences)* uses textual features only, shown in Figure 1.

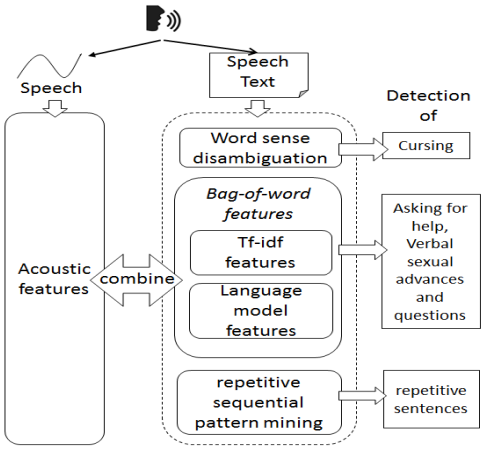

Fig. 1: *Block Diagram of DAVE*

A. Textual and Acoustic Features

Acoustic analysis is significant to detect *asking for help, verbal sexual advances* and *questions*, since studies [29] have shown that, human behaviors are consistent with specific conscious and unconscious emotion concepts. But, relying only on acoustic signal processing might result in inaccuracy since asking for help, verbal sexual advances and questions rely heavily on semantics of speech data. Hence, our solution combines acoustic signal processing with textual inference to detect these three verbal events. In the following subsections we discuss the textual and acoustic features evaluated to detect *asking for help, verbal sexual advances* and *questions*.

1) Text Features: The Bag-of-word representation is widely used in text data analysis. Two of the most widely used bag-of-word representation models are the tf-idf vector model and language models where terms are assumed to be unrelated, in the sense that each term is considered to be an atomic unit of information. DAVE considers converted speech text as a document and extracts bag-of-word features from that document which represent the textual concept of speech. In our solution, a combination of unigram and bigram words are used as terms.

a) Tf-idf Features

tf-idf stands for 'Term Frequency, Inverse Document Frequency' which is a way to score the importance of terms in a text document based on how frequently they appear in that text document and across multiple text documents, where each text document is represented as a text vector and each dimension corresponds to an individual term. The value of a term in this vector shows how important that term is to represent that text document [23].

We represent text portions of our converted speech text using a text vector that captures the relative importance of the terms in the text. The value of a term in our text vector representation is calculated using *tf-idf weighting*. Intuitively, if a term appears frequently in a text document, it is important. Since that relation is not linear, as shown by equation 1 we have used *sublinear tf scaling* to calculate tf where $C(t, d)$ is the frequency of occurrences of term t in text document d. And idf measures how important a term is in an overall sense. While computing tf, all terms are considered equally

important. However certain terms, such as "is", "of", and "that", may appear many times, but have little importance. Thus idf is used to lower the emphasis of frequent terms while scaling up the rare ones, which is computed using equation 2 where N is the total number of documents (sentences) in the training corpus and $DF(t)$ is the total number of documents containing term t. Hence, $tf-idf = tf \times idf$. We represent the text portion from which we want to detect these three verbal events using this text vector representation, and use the term weights from the text vector representation as textual features.

$$tf_t = \begin{cases} 1 + \log C(t,d) & \text{if } C(t,d) > 0 \\ 0 & \text{else} \end{cases} \quad (1)$$

$$idf_t = 1 + \log(\frac{N}{DF(t)}) \quad (2)$$

b) Language Model Features

In utilizing language models, our solution uses the ratio between the log-likelihood of the sentence with respect to the 'verbal event (asking for help, verbal sexual advances or questions) language model' and the log-likelihood of the sentence with respect to the 'non-verbal event language model' as language model features. This log-likelihood ratio (LLR) is computed as:

$$LLR(S) = \log(\frac{P(S|verbaleventLanguageModel)}{P(S|nonverbaleventLanguageModel)}) \quad (3)$$

Here, $P(S|C)$ is the conditional probability of sentence S given class C, where $C \in$ ('verbal event Language Model' or 'non-verbal event language model').

We have explored both the unigram and bigram language models. The language models are computed with maximum likelihood estimation. Equation 4 and 5 show the calculation of unigram model and bigram language model probability where $N(T)$ is the frequency of the term $T \in$ (unigram or bigram) in the training corpus.

$$P^{Uni}(w_i) = \frac{N(w_i)}{\Sigma_{j \in allwords} N(w_j)} \quad (4)$$

$$P^{Bi}(w_i|w_{i-1}) = \frac{N(w_{i-1}, w_i)}{N(w_{i-1})} \quad (5)$$

In the case of the unigram language model, $P(S|C)$ is calculated by equation 6 and the bigram language model is calculated by equation 7, where $S = w_1, w_2 ... w_L$.

$$P^{Uni}(S|C) = \prod_{i=1....L} P(w_i|C) \quad (6)$$

$$P^{Bi}(S|C) = P(w_1|C) \prod_{i=2....L} P(w_i|w_{i-1}C) \quad (7)$$

It is important for language models to attribute a non-zero probability to the words or n-grams that are not seen in a set of training documents (training corpus). To avoid zero probability in calculating the probabilities we used following smoothing methods. These smoothing methods subtract a very small constant from the probability of seen events and distribute it over all seen and unseen events.

Additive Smoothing: Equation 8 shows the computation of additive smoothing for the unigram language model of class C where δ is the smoothing parameter. $N(w_i|C)$ represents the frequency of word w_i in the training corpus. $|C|$ is total word count and $|V|$ is the vocabulary size of the training corpus.

$$P^{AS}(w_i|C) = \frac{N(w_i|C) + \delta}{|C| + \delta|V|} \quad (8)$$

Linear Interpolation Smoothing: This smoothing method use (N-1)gram probabilities to smooth N-gram probabilities. Equation 9 shows the computation of linear interpolation smoothing for the bigram language model of class C where λ is the smoothing parameter to be determined and $P^{AS}(w_i|C)$, $P^{Bi}(w_i|w_{i-1}C)$ are computed using equation 8,5 respectively.

$$\begin{aligned} &P^{LIS}(w_i|w_{i-1}C) = \lambda P^{Bi}(w_i|w_{i-1}C) \\ &+(1-\lambda)P^{AS}(w_i|C) \end{aligned} \quad (9)$$

Absolute Discounting Smoothing: Equation 10 shows the computation of the absolute discounting smoothing for the bigram language model of class C where δ is the smoothing parameter, $N(w_i)$ is the frequency of word w_i, S is the number of seen word types occur after w_{i-1} in the training corpus and $P^{AS}(w_i|C)$ that is computed with equation 8.

$$\begin{aligned} &P^{ADS}(w_i|w_{i-1}C) = \frac{max(N(w_{i-1},w_i)-\delta,0)}{N(w_{i-1})} \\ &+\frac{\delta S}{N(w_{i-1})} P^{AS}(w_i|C) \end{aligned} \quad (10)$$

Hence, we have three language model features (log-likelihood ratios) for each verbal event; One each from the unigram language model with additive smoothing, the bigram language model with linear interpolation smoothing and the bigram language model with absolute discounting smoothing.

2) Acoustic Signal Features: Since, human behaviors remain consistent with specific emotion concepts [29], our goal is to extract acoustic features to represent those emotional concepts that are depicted through their tone of speech. The arousal state of the speaker affects the overall energy, energy distribution across the frequency spectrum, and the frequency and duration of pauses in a speech signal. Hence, the primary continuous acoustic features: *energy and pitch* are used as features in our analysis.

Another important continuous feature is the fundamental frequency *(F0)*, that is produced by the pitch signal, also known as the glottal waveform, which carries speaker tone information because of its dependency on the tension of the vocal folds and the subglottal air pressure.

The harmonics to noise ratio (HNR) in speech provides an indication of the overall aperiodicity of the speech signal. Breathing and roughness are used as parameters for speech analysis and they are estimated by HNR. There is significantly higher HNR in the sentences expressed with anger than the neutral expressions. *Zero crossing rate* is a measure of number of times in a given time interval/frame that the amplitude of the speech signals passes through a value of zero. There is a strong correlation between zero crossing rate and energy distribution with frequency and a reasonable generalization is that if the zero crossing rate is high, the speech signal is unvoiced. Also, *the voicing probability* computed from the *ACF* indicates an acoustic signal is from speech or non-speech. *MFCC* features are the means by which spectral information in the sound can be represented. Here the changes within each coefficient

across the range of the sound are examined. These features take human perception sensitivity with respect to frequencies into consideration.

Hence, in our acoustic analysis the low-level descriptors extracted from small frames are: *zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) frame energy, pitch frequency (normalised to 500 Hz), harmonics-to-noise ratio (HNR) by the autocorrelation function, the fundamental frequency computed from the Cepstrum and mel-frequency cepstral coefficients (MFCC) 1-12 in full accordance to HTK-based computation.* To each of these, the *delta coefficients* are additionally computed. Next the 12 functionals: mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their mean square error (MSE) are applied on a chunk of small frames.

3) Combination of Features Used in Solution: Various combinations of features from sections III-A1 & III-A2 are evaluated for each of the verbal events: asking for help, verbal sexual advances and questions. Through our extensive evaluation we conclude to use a combination of acoustic and all 3 language model features as input to detect verbal events: verbal sexual advances and questions. Also, asking for help is detected using a combination of all the bag-of-word features with acoustic features. Our solution is shown in Figure 2.

4) Detection classifier: Features extracted from both acoustic signals and converted textual data are used as input for a detection classifier. We have used three separate binary classifiers to detect each of these 3 verbal events, where class labels are positives and negatives. We have explored the NaiveBayes, K-nearest neighbor and SVM classifiers as a detection classifier for each of these verbal events (see section IV for the evaluation).

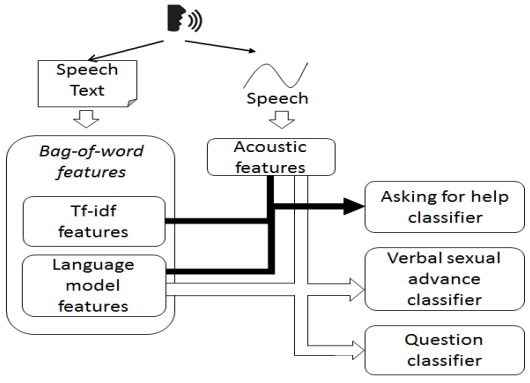

Fig. 2: *Feature combinations to detect verbal events*

B. Using Text Only

1) Detecting Agitated Event: Cursing: To detect *cursing*, we have built a word dataset which contains 165 most used curse words. The words in the converted text are matched with the words of the curse word dataset. If a match is found, it is considered that the curse might have occurred, but we must check the sense in which the word was used. Within these curse words some of them have multiple meanings or word senses. Such as word: 'dog' can be used to describe a pet, also, it can be used as a curse word. Since, linguistic content

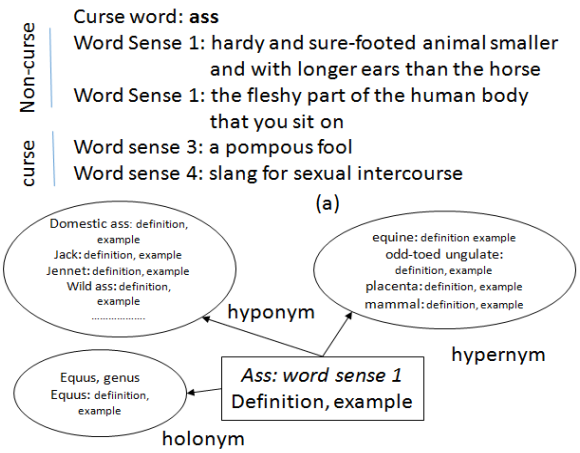

Fig. 3: *WordNet wordsenses for word:ass and relation of one wordsense of word:ass with synset words*

of cursing does not contain compound concepts, to address this challenge we performed word sense identification analysis to detect the latent semantic meaning or word sense of these curse words (instead of text document representation features in section III-A1).

There are several word sense analysis approaches, such as knowledge based methods, supervised methods, and semi-supervised methods. Supervised and semi-supervised approaches need a large dataset and they identify semantic meanings of a word in a specific domain. Our goal is to identify if a word is used as a curse word or not in a generic context. Also, due to the lack of large curse word datasets we have developed a knowledge based approach for our cursing detection. We have used a modified version of the Adapted Lesk Algorithm [8], which uses Wordnet [15] to detect a word sense using the context of neighbor words. WordNet [15] is a lexical database for the English language that groups English words into sets of synonyms called synsets. It provides short definitions and usage examples, and stores relations such as *hypernyms, hyponyms, meronyms, troponyms* etc. among these synonym sets.

To detect the latent word sense of each of the curse words with multiple meanings from the converted text, we define K neighbor context words around the target curse word. For each word in the selected context, our algorithm looks up and lists all the possible senses of two parts of speech: noun and verb. For each word sense our algorithm takes into account its own gloss or definition and examples provided by WordNet [15], and the gloss and examples of the *synsets* that are connected to it through *hypernym, hyponym, meronym or troponym* relations to build an enlarged context for that word sense. All the enlarged contexts for each word sense of all these context words are compared with the enlarged context of each of the word senses of the targeted curse word. The enlarged word sense contexts that overlap most with the enlarge context of all the word senses of neighbor context words of the targeted curse word is the word sense of the targeted curse word.

As an example, suppose a curse word:'ass' occurs in a converted speech text. Figure 3 (a) shows the 4 word senses of the curse word:'ass' extracted from WordNet, where word

sense 1 & 2 are categorized as 'non curse senses' and word sense 3 & 4 are categorized as 'curse senses'. DAVE builds a enlarged context set for all the word senses considering the *hypernyms, hyponyms, meronyms, troponyms* relationship in synset. For example, Figure 3 (b) shows the relationships of *'wordsense 1'* of the curse word 'ass' in synset provided by WordNet. All the words in the definition and examples of word sense 1 and of related words shown in figure 3 (b) are included in the enlarged context set for word sense 1. Hence, the enlarged context set for *'non curse category'* and *'non curse category'* is the union of the enlarged context sets of all the word senses included in that respective category.

To detect cursing we have used WordNet instead of a dictionary since, while traditional dictionaries are arranged alphabetically, WordNet is arranged semantically where each word is connected with words in its synset based on various semantic relations.

2) Detecting Repetitive Sentences: To detect repetitive sentences or questions from text we performed indexing and give unique IDs to each of the words in the text data. Then we convert the words to their corresponding IDs in the text. We modified prefixSpan [20] to find the repetitive subsequences which occurred a minimum of T times in the converted word ID sequence. Since, repetitive sentences from an agitated patient may not be exactly the same; we identify that the repetition of sequence of words has occurred if word sequences match with a maximum of s_n number of words skipped. That means, if s_n is 2, "I eat chocolate" matches with "I eat too many chocolate" but it is not matched with "I eat too too many chocolate". Our prefixSpan [20] modification limits the expansion of search space into further branches using the knowledge of minimum number of repetitions required and maximum number of allowed word skips. Suppose a converted sequential word ID representation of a text is $< W_1W_2W_3W_1W_2W_4W_3W_5 >$ where each W_i is the unique word ID of the ith unique word of the text. If we consider T as 2 and s_n as 1, the search space of our algorithm is shown in figure 4. Here, the search space is divided into 5 branches, one for each of the unique word IDs. This growth-based approach of finding sequential words grows larger by dividing the search space and focusing only on the subspace potentially supporting further growth. Unlike traditional apriori based approaches which perform candidate generation and test, this approach does not generate any useless candidate sequences. Also, two word sequences extracted from the left-most branch in figure 4 are combined since, sequence $< W_1W_2 >$ is a subset of sequence $< W_1W_2W_3 >$. Hence, our resultant repetitive word sequences are: $< W_1W_2W_3 >, < W_2W_3 >$. Studies [20] have shown that, in the average case the growth based approaches for sequential pattern mining perform up to 40% faster and uses about 0.1 times the memory for computation, compared to other apriori based approaches. For DAVE we use $s_n = 3$.

IV. Evaluation on Healthy People

Section IV-A describes the experimental setup and datasets used from movies, Youtube, Taboeba website and our own data collection. Using this data, the next three sections show the evaluation for vocal events that require both bag-of-word textual features and acoustic information (section IV-B), that require word sense disambiguation from text information (section IV-C), and that require repetitive sequential pattern mining from text information (section IV-D), respectively. Through these evaluations we find which combinations of features and approaches provide higher detection accuracy for respective verbal events in generic domains.

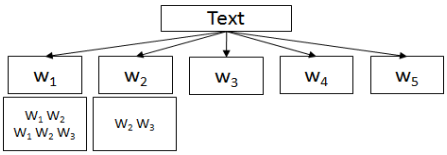

Fig. 4: *Search space of modified version of prefixSpan*

A. Experimental Setup - Preliminaries

1) Acoustic Pre-processing: For completeness, this section describes the noise filtering and the conversion of audio to text.

Filtering and Removing Noise: The first step for pre-processing is to remove unvoiced audio segments using zero crossing rate (ZCR) [7]. To capture the pause in spoken sentences, the detected voiced segments are lengthened by 1 second on both sides. If another consecutive voiced segment starts within the lengthened 1 second segment portion, both the voice segments are merged into one.

Noises which are out of human voice frequency range were removed using a bandpass filter with a low frequency of 80Hz and a high frequency of 3000Hz. Hiss, hum or other steady noises were reduced using a spectral noise gating algorithm [3].

Converting Audio to Text: Since all of our solutions require text, we require audio to text conversion of the sound clips. We have used Dragon NaturallySpeaking [1] which is a speech recognition and transcription system.

2) Textual Data for Training Lexical Models: We used separate training corpuses to compute 'Document Frequency' of unigram and bigram words and language models for each of the verbal events: asking for help, verbal sexual advances and questions. We learned two language models corresponding to each of the verbal events (asking for help, verbal sexual advances and questions) while one detects the presence of that verbal event, other detects the absence. These language models have the purpose of representing the main word sequences that occur in an utterance from respective verbal events rather than other events.

The wiki talk pages [6] consist of threaded posts by different authors about a particular wikipedia entry. While the sentences from these posts lack certain properties of spontaneous speech, they are more conversational than articles. Tatoeba website [4] contains a large collection of human spoken sentences in text and audio. Also, the urbandictionary website [5] has a large collection of human spoken sentences performing verbal events: verbal sexual advances and cursing. We labeled sentences from wiki talk posts, tatoeba and urbandictionary websites to include them in our training corpuses.

To achieve lexical characteristics of spontaneous verbal event utterances we conducted a survey of 21 volunteers where participates were asked what they will say to perform our targeted verbal events in different random scenarios and included the responses in respective training corpuses. Table I shows the number of sentences in training corpuses for each of the verbal events: asking for help, verbal sexual advances and questions.

Training corpus	Verbal events		non-verbal events
	From survey	From other sources	
Asking for help	144	856	3000
Verbal sexual advances	98	601	3000
Questions	335	1165	5000

TABLE I: *Number of sentences in training corpuses*

3) Data for Verbal Event Detection Evaluation: Since there is no existing available dataset for asking for help, verbal sexual advances, questions, cursing and talking with repetitive sentences we had to create our own dataset for training and evaluation. We have collected verbal speech data from 6 individuals, whose ages varies from 21 to 30. There were 4 females and 2 males. We also, collected human spoken sentences audio clips from Tatoeba website [4]. To enrich our dataset we have included audio clips from movies and real *Youtube* videos where people are performing our targeted verbal events. Table II shows the number of audio clips for evaluation for each of the 5 verbal events. These clips have lengths varying from 2 to 20 seconds, containing 1 to 23 words. To evaluate curse detection from audio we have collected 260 clips from movies, among 137 of them people used 'cursing' in their conversation. Among these 137 'cursing' events, 91 of them have curse words which have a single meaning, and 46 of them have multiple meanings. Also, 50 audio clips have multiple meaning 'curse' words with non-curse meanings.

Verbal event	Number of clips
Asking for help	260
Verbal sexual advances	165
Questions	400
Cursing	260
Repetitive sentences	80
Others	500

TABLE II: *Number of clips for evaluation of verbal event detection*

4) Pre-processing of Textual Data: Stop words are usually the most frequent words including articles, auxiliary verbs, prepositions, conjunctions and they do not provide additional improvement for textual similarity analysis. We have created a customized stop-word list for verbal events: asking for help, verbal sexual advances and questions, and created our vocabulary set with corresponding *idf* values from the training converted text set. We used Porter steaming to reduce inflected words to their base form and normalization to remove punctuation marks, and converted words to lower case in our process of vocabulary building. After this pre-processing the vocabulary size for asking for help, verbal sexual advances, and questions are 214, 178, 658, respectively.

B. Verbal Events: Combination of Acoustic and Text Data

For each of the verbal events: asking for help, verbal sexual advances and questions we have trained a binary class classifier. Since, binary classifiers do not work well when trained with imbalanced data sets: new instances are likely to be classified as the class that has more training samples. In order to avoid this over-fitting problem, we chose to resample the dataset by keeping all clips for respective verbal events and randomly extracting subsets of clips of the same size (sampled from other 5 categories of table II). In the following section we evaluate how a verbal event detection classifier performs using only acoustic features (section IV-B1), then we evaluate how the detection accuracy changes by adding textual features into the classifier (section IV-B2). All the evaluations were done using 10-fold cross validation with 33.33% of the data as test data. We used accuracy, precision, and recall as our detection performance evaluation metrics.

1) Acoustic Features: We tested the NaiveBayes, K-nearest neighbor and SVM classifiers using the acoustic features discussed in section III-A2. Table III shows the evaluation of these three verbal events using only acoustic features. As we can see, the accuracy is ranging between 66% to 82.5%.

Event	Classifier	Accuracy	Precision	Recall
Asking for help	NaiveBayes	71.72	0.707	0.718
	K-nearest neighbor	80.314	0.803	0.805
	SVM	79.031	0.811	0.79
Verbal sexual advances	NaiveBayes	73.35	0.78	0.734
	K-nearest neighbor	81.43	0.813	0.814
	SVM	82.54	0.824	0.826
Questions	NaiveBayes	66.093	0.719	0.661
	K-nearest neighbor	81.97	0.821	0.82
	SVM	82.22	0.829	0.822

TABLE III: *Evaluation with acoustic features*

2) Combination of Acoustic-Textual Features: According to Table III the best detection accuracy using acoustic features alone are 82.54%, 80.31% and 82.22% for verbal events: verbal sexual advances, asking for help and questions. To achieve higher accuracy, we introduce textual features which represent the semantic understanding of speech while expressing these verbal events.

The studies related to automatic speech recognition systems have to additionally take into account the speech recognition errors which get more frequent for poor sound qualities and on spontaneous speech, and can highly decrease the classification performance. Hence, the classifier evaluations are carried out using features stemming from:

- manual transcriptions - to study the classifier's maximum performance, obtainable only in ideal conditions (i.e. with perfect transcripts)
- automatic transcriptions (obtained with Dragon speech recognizer) - to study the performance under real conditions

Sections IV-B2a, IV-B2b,and IV-B2c are devoted to the evaluation of verbal event detection using a combination of acoustic and tf-idf features, using a combination of acoustic and language model features and a combination of acoustic and all bag-of-word features.

a) Combination: Acoustic and Tf-idf Features

In this work two types of tf-idf features: unigram and bigram were extracted from transcribed speech text. For all the evaluations shown in this section, the SVM classifier gave

higher accuracy, hence only the results obtained with the SVM classifier are presented here. Figure 5 and 6 shows the

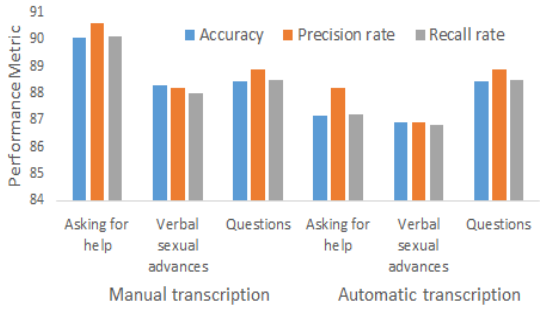

Fig. 5: *Evaluation with acoustic and unigram tf-idf features*

evaluation combining acoustic with unigram and all (unigram and bigram) tf-idf features, respectively with both manual and automatic transcription. According to these evaluations detection accuracy increases up to 91.88%, 89.44% and 88.92% for verbal events: asking for help, verbal sexual advances and questions with both unigram and bigram textual features extracted from manual transcription. Accuracy decreases by 0.28%, 1.56% and 1.1% when tf-idf features are extracted from automatic speech transcription. According to our evaluation most important tf-idf terms (higher tf-idf values) for questions and asking for help event detection are *wh-words* (why, who, which, what, where, when, and how) and 'help', 'please', 'need', 'can you', 'will you' , etc. respectively. Transcription error rate for these simple words are low for Dragon NaturallySpeaking software, hence decrease of detection accuracy for: asking for help and questions detection is lower compared to verbal sexual advances detection.

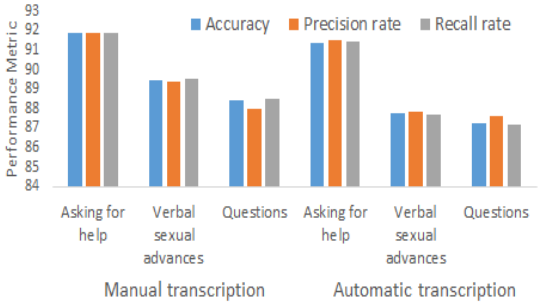

Fig. 6: *Evaluation with acoustic, unigram and bigram tf-idf features*

As shown in figure 5, with manual transcription questions detection accuracy (88.45%) is higher using combination of acoustic and unigram tf-idf features. Hence, we conclude that questions detection perform better using a combination of acoustic and unigram tf-idf features, where other two event detections perform better using acoustic features in addition to both unigram and bigram tf-idf features.

b) Combination: Acoustic and Language Model Features

In our evaluation three language model features: 'unigram log-likelihood ratio', 'log-likelihood ratio of bigram language models with linear interpolation smoothing' and 'log-likelihood ratio of bigram language models with absolute discount smoothing' are extracted from speech text. For all the evaluations shown in this section, the SVM classifier gave higher accuracy, hence only the results obtained with the SVM classifier are presented here.

Table IV shows the evaluation using various combinations of language model features in addition to acoustic features to detect verbal events: asking for help, verbal sexual advances and questions using manual transcription. According to this evaluation, all 3 language model features in addition to acoustic features used as input provide higher accuracy for all 3 verbal events. Hence, in the later sections of this paper the term 'using language model features' will be referred to as using all 3 of the language model features.

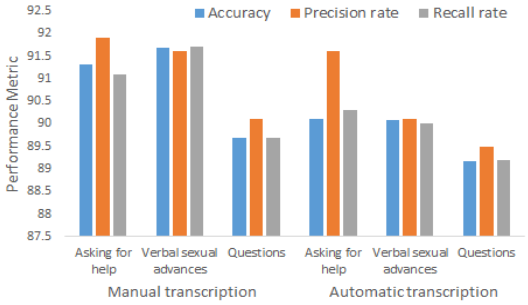

Fig. 7: *Evaluation with acoustic and language model features*

Figure 7 shows the evaluation metrics using acoustic and language model features extracted from manual and automatic transcription. Using manual transcription, the highest accuracy for asking for help is 91.36%, which is lower compare to detection using acoustic and tf-idf features as shown in Figure 6. On the contrary, accuracy for verbal sexual advances and questions detection increases to 91.69% and 89.68%, respectively using a combination of acoustic and language model features as input for the classifier. Detection accuracy decreases by 0.8%, 1.74% and 0.05% for verbal events: asking for help, verbal sexual advances and questions, respectively when textual features are extracted from automatic speech transcription, which complies with our evaluation in section IV-B2a.

c) Combination: Acoustic and Textual Features

Tables V and VI show the evaluation metrics of verbal events: asking for help, verbal sexual advances and questions detection using acoustic features in addition to both bag-of-word (tf-idf and language model) textual features extracted from manual and automatic transcriptions, respectively. As shown in Tables V and VI combining all textual and acoustic features we achieve up to 93.45% and 91.36% accuracy for asking for help detection using manual and automatic transcription. According to Tables V & VI, and Figure 7 the highest achieved accuracy of verbal sexual advances and questions detection using a combination of all textual and acoustic features are similar to detection using acoustic and language model features only, for both manual and automatic transcription.

Hence we conclude that, among all the combinations of features we have evaluated, a combination of acoustic and language model features are sufficient to detect verbal events: verbal sexual advances and questions, where to achieve higher

Event	Language model features	Acoustic features	Accuracy	Precision	Recall
Asking for help	Additive smoothing	All	87.95	0.894	0.88
	Linear interpolation smoothing	All	87.958	0.895	0.885
	Absolute discounting smoothing	All	87.96	0.9	0.88
	All 3 features	All	91.1	0.919	0.911
Verbal sexual advances	Additive smoothing	All	90.943	0.909	0.91
	Linear interpolation smoothing	All	90.944	0.91	0.91
	Absolute discounting smoothing	All	90.944	0.91	0.91
	All 3 features	All	91.6981	0.916	0.917
Questions	Additive smoothing	All	88.206	0.885	0.882
	Linear interpolation smoothing	All	88.24	0.886	0.884
	Absolute discounting smoothing	All	87.96	0.882	0.88
	All 3 features	All	89.68	0.901	0.897

TABLE IV: *Evaluation of various combinations of language model features in addition to acoustic features with manual transcription*

Event	Classifier	Accuracy	Precision	Recall
Asking for help	NaiveBayes	80.36	0.801	0.804
	KNN	91.36	0.915	0.914
	SVM	93.45	0.934	0.935
Verbal Sexual Advances	NaiveBayes	81.13	0.846	0.811
	KNN	87.15	0.874	0.872
	SVM	91.69	0.916	0.917
Questions	NaiveBayes	71.74	0.755	0.717
	KNN	87.96	0.88	0.89
	SVM	89.697	0.9	0.897

TABLE V: *Evaluation with combined features with manual transcription*

Event	Classifier	Accuracy	Precision	Recall
Asking for help	NaiveBayes	78.79	0.784	0.788
	KNN	89.79	0.9	0.898
	SVM	91.36	0.915	0.914
Verbal Sexual Advances	NaiveBayes	80.32	0.843	0.801
	KNN	84.15	0.854	0.842
	SVM	90.154	0.902	0.91
Questions	NaiveBayes	72.48	0.758	0.725
	KNN	89.6	0.897	0.893
	SVM	88.68	0.89	0.88

TABLE VI: *Evaluation with combined features with automatic transcription*

accuracy for asking for help we need a combination of all the bag-of-word features with acoustic features.

C. Detecting Cursing

As shown in Table VII we evaluate cursing detection using only acoustic features (from section III-A2), our cursing detection approach (shown in section III-B1) and a combination of both, where output of our cursing detection approach is used as a binary feature. We have used the SVM classifier as a detection classifier for this evaluation. As shown in Table VII, cursing detection using only acoustic features results in a low precision and recall rate of 75.1% and 77.4%, respectively. When manual transcription data is used for evaluation, our cursing detection approach detects all of the 91 single sense curse words and 41 of the multiple sense curse words. As shown in table VII the precision rate for the multiple sense curse word detection is 87.23% and the recall rate is 89.13% and for overall curse detection the precision rate is 95.6% and the recall rate is 96.35%. The multiple sense curse words have word senses that varied from 2 to 9. If we try to detect the specific word sense for the curse words with multiples senses, the detection evaluation, precision rate goes down to 72.7% which shows that our binary word sense adaptation of Adapted Lesk algorithm [8] improves the curse word detection performance.

Transcription	Features	Precision rate	Recall rate
	Acoustic	75.1	77.4
Manual	Textual	95.6	96.35
	Acoustic and textual	95.6	96.35
Automatic	Textual	93.9	91.2
	Acoustic and Textual	93.9	91.2

TABLE VII: *Evaluation of cursing detection*

Since, many of the curse words are complex and uncommon in general English vocabularies, the transcription error rate for software like Dragon is higher for them. After a short training of curse words transcription accuracy improves significantly. After short training, with automatic transcription by Dragon we achieved 93.9% precision and 91.2% recall for overall curse detection using our cursing detection approach (shown in section III-B1).

According to our evaluation as shown in Table VII, combining acoustic inference from speech with our cursing detection approach from transcribed speech does not improve accuracy. Hence, we conclude that our cursing detection approach (as shown in section III-B1) is sufficient to detect cursing from speech.

D. Detecting Repetitive Sentences

A study [27] on agitated demented elderly patients has shown that $50 - 80\%$ of them suffer from palilalia, which is a speech disorder characterized by the involuntary repetition of syllables, words, or phrases. Hence, it is highly unlikely of them will repeat large sentences with many words skipped. We have evaluated our algorithm on 80 speech samples collected in controlled experiments. Each of the converted text of these speech samples contains sentence repetition with at most 3 words skipped. Using manual and automatic transcription detection, the accuracy was 100% and 98.7%, respectively.

V. Real Patient Evaluation

We have also evaluated verbal event detections approaches on real agitated dementia patients using audio clips collected in realistic settings from elderly suffering from dementia. The clips (N=107) were collected for an NIH-funded randomized clinical trial (ClinicalTrials.gov Identifier: NCT01324219) that tested whether improved nursing home staff communication

reduces challenging behaviors in persons with Alzheimer's disease and other dementias [28]. The clips were collected during morning care activities in 16 midwestern nursing homes. Duration of the audio clips vary between one to 30 minutes. The clips contain examples of agitated verbal events as well as periods without agitated verbal events. In total, 34 residents were included in the clips ranging in age from 63 to 98 years old (Mean=88, Standard Deviation = 7.2) and were 70% female, 97% Caucasian non-Hispanics, and 67% were prescribed psychotropic medications. Table VIII shows the number of agitated verbal events that appeared in those audio clips. The distribution of verbal events shows that verbal event questions are more common in agitated elderly suffering from dementia. There were no examples of verbal sexual advances. This data is representative of the fact that these events are not common for agitated elderly in nursing homes suffering from dementia.

Verbal events	Total (136 events in 107 clips)
Asking for help	11
Verbal sexual advances	0
Questions	52
Cursing	10
Repetitive sentences	14

TABLE VIII: *Verbal events from real patients.*

Speech from agitated elderly suffering from dementia may vary from generic cases, but we have a relatively small dataset of 107 audio clips from 34 read agitated dementia patients. Hence, we apply the solutions from section IV where we performed a leave one out cross validation on real patient data, additionally incorporating our collected data (as shown in section IV-A3) from previous steps in the training set. Note that audio clips from real dementia patients contain extensive amount of hiss, hum, and other steady noises as well as music, coughing, door and window movements, beep sounds from air conditioners, etc. We used a spectral noise gating algorithm to reduce steady noises like hiss or hum sounds. To eliminate the effect of environmental noise, we have included environmental sounds (music, coughing, door and window movements, beep sounds, etc.) in the training set. For example, to evaluate the performance of asking for help detection using a binary classifier with one clip collected from the real patient data being tested, we included all the data from real patients and our other collected data (as shown in section IV-A3) in the training set. Also, examples for environmental noise are included as negative examples for this binary classifier in the training set. The evaluation results of verbal event detection

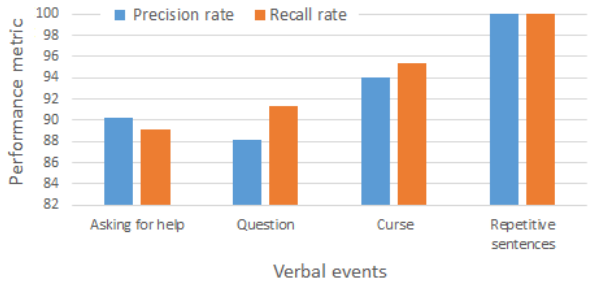

Fig. 8: *Real patients evaluation with manual transcription*

using real dementia patient data with manual transcription is shown in Figure 8 where performance metrics for verbal event detection remain approximately similar to our evaluation with our collected data (as shown in section IV-A3). Figure 9 shows the evaluation of verbal events with automatic transcription on real patient data. Due to presence of noise, transcription error rate for real patient audio data was higher compare to our previous evaluations, which reduce the event detection accuracy. In this evaluation the SVM classifier is used as a detection classifier.

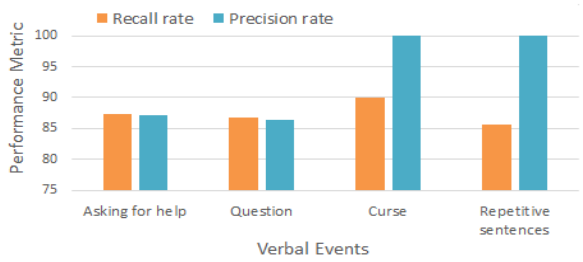

Fig. 9: *Real patients evaluation with automatic transcription*

VI. Discussion

A major challenge that we solved was detecting verbal agitation for dementia patients who mumble, speak in low volume, and don't articulate words very well. The value of this detection is clear from the medical community which uses them in their Cohen-Mansfield metrics to help in treatment. Significantly, we also showed that our solutions generalizes to the healthy population. The value for the healthy population is less obvious. However, applications of our solution include online video sharing sites such as *Youtube* and movies, where providers and users are able to detect objectionable content such as cursing, sexual advances, etc. to impose restrictions (e.g., for children). Detection of asking for help and questions can improve several human computer interaction (HCI) systems such as: automated customer service interaction systems, smart classrooms, etc. Also, some of the vocal events such as: asking for help, verbal sexual advances, and cursing are important for home safety.

While the focus of this paper is to provide the details of the algorithmic solutions and their evaluation, it is possible to incorporate the solutions into a working system. In fact, we have implemented the solutions on a Kinect system and deployed it in three homes with healthy people. Figure 10 shows the evaluation of the home deployments using automatic transcription. According to this evaluation, the performance of the classifiers (SVM classifiers) for all the verbal events are approximately similar to the evaluation using our collected data (as shown in section IV-A3). Using this system with dementia patients will require a multi-year pilot study which is beyond the scope of this paper.

DAVE also addresses one key aspect of privacy. It keeps the recorded acoustic data private and only presents the type and time of occurrences of agitated vocal events through its interface. For example, a graphical representation of the change of frequency of agitated behavior can be displayed and then used by the caregiver to help diagnose the state of the disease of a patient.

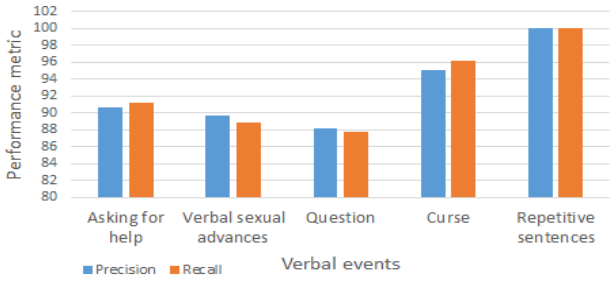

Fig. 10: *Detection of verbal events in homes*

VII. Conclusions

DAVE is the first system which accurately and automatically detects the 5 vocal events of the Cohen-Mansfield inventory. To our knowledge, the automatic detection of verbal events asking for help, sexual verbal advances, cursing with word sense, and repetitive sentence have not been studied. Our solution of questions detection improves the accuracy above the state of art. To solve the detection problems for asking for help, verbal sexual aggression, and questions we use a novel combination of text mining and signal processing. For cursing we apply a word sense disambiguation technique. For repetitive sentences we employ a sequential pattern mining approach. We have provided an extensive evaluation of DAVE that includes audio clips collected from real agitated elderly patients suffering from dementia, *Youtube*, movies, online data repositories, controlled experiments, and home deployments. In this study, the use of human subjects was approved by an IRB.

VIII. Acknowledgement

This paper was supported, in part, by DGIST Research and Development Program (CPS Global center) funded by the Ministry of Science, ICT and Future Planning, NSF CNS-1319302 and, the National Institute of Nursing Research of the NIH under Award Number R01NR011455.

References

[1] 2016a. Dragon NaturallySpeaking. http://tinyurl.com/26rcknk. (1 Jan 2016).

[2] 2016b. Resident profile. http://tinyurl.com/ohesoqh. (1 Jan 2016).

[3] 2016c. spectral noise gating algorithm. http://tinyurl.com/yard8oe. (1 Jan 2016).

[4] 2016d. Tatoeba. https://goo.gl/sr54d0. (1 Jan 2016).

[5] 2016e. Urban dictionary. http://www.urbandictionary.com/. (1 Jan 2016).

[6] 2016f. The Wiki talk pages. https://en.wikipedia.org/wiki/Help:Using_talk_pages. (1 July 2016).

[7] RG Bachu, S Kopparthi, B Adapa, and BD Barkana. 2008. Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) Zone Conference Proceedings*. 1–7.

[8] Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Computational linguistics and intelligent text processing*. Springer, 136–145.

[9] Kofi Boakye, Benoit Favre, and Dilek Hakkani-Tür. 2009. Any questions? Automatic question detection in meetings. In *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 485–489.

[10] Jiska Cohen-Mansfield. 1991. Instruction Manual for the Cohen-Mansfield Agitation Inventory (CMAI). *Research Institute of the Hebrew Home of Greater Washington* (1991).

[11] Jiska Cohen-Mansfield. 1997. Conceptualization of agitation: results based on the Cohen-Mansfield agitation inventory and the agitation behavior mapping instrument. *International Psychogeriatrics* 8, S3 (1997), 309–315.

[12] Victor Foo Siang Fook, Pham Viet Thang, That Mon, Qiang Qiu Htwe, Aung Aung Phyo Phyo, Biswas Jit Jayachandran, and Philip Yap. 2007. Automated Recognition of Complex Agitation Behavior of Demented Patient Using Video Camera. (2007).

[13] Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. 1997. Automatic detection of discourse structure for speech recognition and understanding. In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 88–95.

[14] Anna Margolis and Mari Ostendorf. 2011. Question detection in spoken conversations using textual conversations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 118–124.

[15] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[16] Torin Monahan and Tyler Wall. 2002. Somatic surveillance: Corporeal control through information networks. *Surveillance & Society* 4, 3 (2002).

[17] Shahriar Nirjon, Chris Greenwood, Carlos Torres, Stefanie Zhou, John A Stankovic, Hee Jung Yoon, Ho-Kyeong Ra, Can Basaran, Taejoon Park, and Sang H Son. 2014. Kintense: A robust, accurate, real-time and evolving system for detecting aggressive actions from streaming 3D skeleton data. In *Pervasive Computing and Communications (PerCom), 2014 IEEE International Conference on*. IEEE, 2–10.

[18] Luiza Orosanu and Denis Jouvet. 2015. Combining lexical and prosodic features for automatic detection of sentence modality in French. In *International Conference on Statistical Language and Speech Processing*. Springer, 207–218.

[19] Vikram Patel and RA Hope. 1992. A rating scale for aggressive behaviour in the elderly–the RAGE. *Psychological medicine* 22, 01 (1992), 211–221.

[20] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. 2004. Mining sequential patterns by pattern-growth: The prefixspan approach. *Knowledge and Data Engineering, IEEE Transactions on* 16, 11 (2004), 1424–1440.

[21] V Minh Quang, Laurent Besacier, and Eric Castelli. 2007. Automatic question detection: prosodic-lexical features and cross-lingual experiments. In *Proc. Interspeech*, Vol. 2007. 2257–2260.

[22] Vũ Minh Quang, Eric Castelli, and Phm Ngc Yên. 2006. A decision tree-based method for speech processing: question sentence detection. In *International Conference on Fuzzy Systems and Knowledge Discovery*. Springer, 1205–1212.

[23] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[24] Margaret C Sewell, Xiaodong Luo, Judith Neugroschl, and Mary Sano. 2013. Detection of mild cognitive impairment and early stage dementia with an audio-recorded cognitive scale. *International Psychogeriatrics* 25, 08 (2013), 1325–1333.

[25] Yaodong Tang, Yuchen Huang, Zhiyong Wu, Helen Meng, Mingxing Xu, and Lianhong Cai. 2016. Question detection from acoustic features using recurrent neural network with gated recurrent unit. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6125–6129.

[26] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. 2014. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 415–425.

[27] Pauline K Wiener, Dimitris N Kiosses, Sibel Klimstra, Christopher Murphy, and George S Alexopoulos. 2001. A short-term inpatient program for agitated demented nursing home residents. *International journal of geriatric psychiatry* 16, 9 (2001), 866–872.

[28] Herman R Bossen A Williams K, Perkhounkova Y. A Communication Intervention to Reduce Resistiveness in Dementia Care: A Cluster Randomized Controlled Trial.. In *The Gerontologist*.

[29] Yael Zemack-Rugar, James R Bettman, and Gavan J Fitzsimons. 2007. The effects of nonconsciously priming emotion concepts on behavior. *Journal of Personality and Social Psychology* 93, 6 (2007), 927.