Demo Abstract: KinVocal: Detecting Agitated Vocal Events

Asif Salekin[1], Hongning Wang[2], John Stankovic[3]
University of Virginia
[1]as3df@virginia.edu, [2]hw5x@virginia.edu, [3]stankovic@virginia.edu

ABSTRACT

Many elderly who are suffering from dementia are also suffering from agitation. While most assisted living facilities and home health care situations rely upon caregivers to monitor and record agitation of their patients, the accuracy is limited because the caregiver must be present during the agitation and must record the event properly. Accurate 24-7 data would help physicians with improved diagnoses and care. To solve this problem we developed KinVocal, a system that continuously monitors and detects agitated vocal events and can be used for the elderly population suffering from dementia. KinVocal, using a novel combination of acoustic signal processing and multiple text mining techniques, automatically detects and records the 8 major vocal agitations for dementia patients as defined by the medical community. This includes: constant unwarranted request for attention or help, making verbal sexual advances, crying, screaming, laughing, cursing, speaking in repetitive sentences, and negativism. The novelty of KinVocal includes the comprehensiveness of addressing all 8 vocal events, using the text of the vocalizations only when accurate, combining text and acoustic features when necessary, and employing text mining and feature identification. A comprehensive performance evaluation includes using data from *Youtube* and movies, controlled experiments, and real in-home deployments. The results show high accuracy for all 8 vocal events.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: User Profiles and Alert; H.5.5 [**Sound and Music Computing**]: Systems

General Terms

Algorithm, Design, Experimentation

Keywords

Vocal agitation, text mining, dementia, word sense

SenSys '15, November 01-04, 2015, Seoul, Republic of Korea
ACM 978-1-4503-3631-4/15/11.
http://dx.doi.org/10.1145/2809695.2817853

1. INTRODUCTION

The medical community has defined the the Cohen Mansfield Agitation Inventory [2] which specifies approximately 30 agitated behaviors for identifying whether a person is suffering from agitation or not. Many of these behaviors are physical such as punching or kicking. An automated system [3] already exists for monitoring and recording physical agitation behaviors. However, there is no such a system for the vocal agitation metrics of Cohen-Mansfield Inventory. In this abstract we describe KinVocal, an automated system for monitoring and recording all the vocal agitation metrics of the Cohen-Mansfield Inventory. This includes crying, screaming, laughing, cursing or verbal aggression, constant unwarranted request for attention or help, negativism, making verbal sexual advances, and talking with repetitive sentences.

Most work in acoustic signal processing for human vocal data uses either features from the signal such as pitch, energy, etc., or convert the speech to text and then use dictionaries to interpret the content of the speech. Our solution, KinVocal, divides the 8 acoustic events of Cohen-Mansfield into three equivalence classes based on the processing technique required to achieve accurate recognition of that particular event. The three classes are: use both the features of the signal and text, use only the acoustic signal features, and use only the text. Further, instead of only using dictionaries for the text we employ 3 different types of text mining techniques depending on the event being detected. This includes textual similarity, text word sense disambiguation, and repetitive sequential pattern mining. Doing this divide and conquer approach results in high accuracy at the lowest cost of processing. 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 will mistakenly falsely 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. In this case the acoustic signal alone will be insufficient due to variation of mood or tone although they correspond to the same verbal event. In another case, for crying, it is sufficient to use only acoustic features (and, in fact, there is usually no text).

The main contributions of KinVocal are:

- An automatic and comprehensive system for detecting verbal agitation 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 the verbal events: asking for help and verbal sexual advances. KinVocal have shown that detection of these two vocal events depends both on the acoustic signal processing and the semantics understanding of the speech. To understand the semantics of speech we employ text data mining techniques exploring text similarity.
- Cursing is difficult to detect because many such words have multiple meanings. We have used a modified version of the adapted Lesk algorithm [1] which considers a word's sense, to detect curse words with multiple ambiguous meanings.

2. OVERVIEW OF KINVOCAL

KinVocal is an automated monitoring and recording system that detects the 8 verbal agitations of the Cohen Mansfield agitation inventory. It uses the Microsoft Kinect sensor to collect raw acoustic data. As the signal arrives, it is important to note that some noise elimination is already handled by the Kinect system. Then, once this *cleaned* data is collected, it proceeds through various stages of processing, see Figure 1, divided into three categories based on two types of signals: text and features of the raw signal. One module converts the raw signal into text. It then supplies that text to 4 other modules, i.e., those modules that are used to detect negativism, cursing, repetitive sentences which only require the text (category 1), and a combined module that addresses asking for help and verbal sexual advances which requires both the text and features of the raw signal (category 2). The raw signal only (category 3) is also sent to the module that detects crying, laughing and screaming.

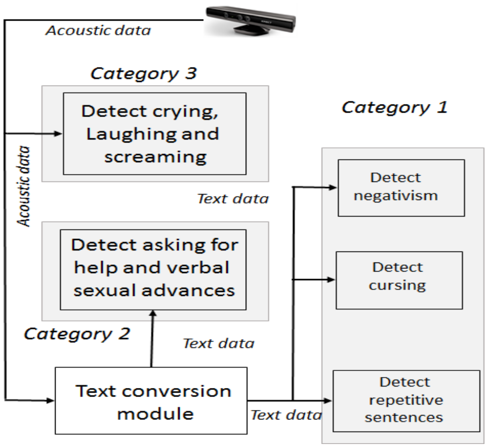

Figure 1: Block Diagram of KinVocal

3. EVALUATION

We have performed an extensive evaluation of KinVocal that includes using *Youtube*, movies, controlled experiments, and home deployments. Using a combined feature set of acoustic and textual features, KinVocal has achieved a detection accuracy of 86.62% for asking for help and a detection accuracy of 89.17% for verbal sexual advances. Also, KinVocal have detected laughing, crying and screaming using an acoustic time series approach resulting in greater than 96% accuracy. Using our modified adapted Lesk algorithm [1] we have detected cursing with 95.6% accuracy. In the future we will apply our solutions to the verbal utterances of actual dementia patients.

4. DEMO SCRIPT

The objective of the demo is to show how KinVocal monitors and detects each of the 8 verbal agitations of the Cohen-Mansfield agitation inventory. The demo will proceed in phases. First, we will list audio clips collected from *Youtube* videos or movies. This clips will then be passed through Kinect for noise removal. After noise removal, acoustic and textual features will be extracted from these audio clips. The set of clips will cover the 8 verbal agitations and a 9th situation where none of these agitations are present. Then, in the final phase of the demo, for each audio clip the 8 separate modules for each of these verbal events will take the features as input and detect whether that targeted verbal event has occurred or not.

5. ACKNOWLEDGMENTS

This work was supported, in part, by NSF grant CNS-1319302.

References

[1] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational linguistics and intelligent text processing*, pages 136–145. Springer, 2002.

[2] J. Cohen-Mansfield. Instruction manual for the cohen-mansfield agitation inventory (cmai). *Research Institute of the Hebrew Home of Greater Washington*, 1991.

[3] S. Nirjon, C. Greenwood, C. Torres, S. Zhou, J. A. Stankovic, H. J. Yoon, H.-K. Ra, C. Basaran, T. Park, and S. H. Son. 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*, pages 2–10. IEEE, 2014.