INTERSPEECH 2020: Show & Tell Contribution
October 25–29, 2020, Shanghai, China

Toward Remote Patient Monitoring of Speech, Video, Cognitive and Respiratory Biomarkers Using Multimodal Dialog Technology

Vikram Ramanarayanan[*†], Oliver Roesler*[*], Michael Neumann*[*], David Pautler*[*], Doug Habberstad*[*], Andrew Cornish*[*], Hardik Kothare*[*†], Vignesh Murali*[*], Jackson Liscombe*[*], Dirk Schnelle-Walka*[*], Patrick Lange*[*], and David Suendermann-Oeft*[*]

[*] Modality.ai, Inc.
[†] University of California, San Francisco
www.modality.ai

Abstract

We demonstrate a multimodal conversational platform for remote patient diagnosis and monitoring. The platform engages patients in an interactive dialog session and automatically computes metrics relevant to speech acoustics and articulation, oro-motor and oro-facial movement, cognitive function and respiratory function. The dialog session includes a selection of exercises that have been widely used in both speech language pathology research as well as clinical practice – an oral motor exam, sustained phonation, diadochokinesis, read speech, spontaneous speech, spirometry, picture description, emotion elicitation and other cognitive tasks. Finally, the system automatically computes speech, video, cognitive and respiratory biomarkers that have been shown to be useful in capturing various aspects of speech motor function and neurological health and visualizes them in a user-friendly dashboard.

1. Dialog Systems for Remote Patient Monitoring

Diagnosis, detection and monitoring of neurological and mental health in patients remain a critical need today. We developed NEMSI [1] – NEurological and Mental health Screening Instrument – to address this need. NEMSI is a cloud-based multimodal dialog system that conducts automated screening interviews over the phone or web browser to elicit evidence required for detection or progress monitoring of neurological or mental health conditions, such as clinical depression, Amyotrophic Lateral Sclerosis (ALS), Alzheimer's disease, and dementia, among others. It makes use of devices available to everyone everywhere (web browser, mobile app, or regular phone), as opposed to dedicated, locally administered hardware, like cameras, servers, audio devices, etc. The back-end is deployed in an automatically scalable cloud environment allowing it to serve an arbitrary number of end users at a small cost per interaction. It is also natively equipped with real-time speech and video analytics modules that extract a variety of features of direct relevance to clinicians in the neurological and mental spaces.

2. Interaction Flow

Patients and other end users can access the secure conversational platform via a website link using secure login credentials and multi-factor authentication. After completing appropriate microphone and camera checks for quality of the recorded audio and video, users hear the dialog agent's voice and are prompted to start a conversation with the agent, whose virtual image also appears in a web window (Figure 1). Users are also able to see their own video in a small window in the upper right corner of the screen. The virtual agent then engages with users in structured conversational exercises designed to elicit speech and facial behaviors that help assess various aspects of speech acoustics and articulation, oro-motor and oro-facial movement, cognitive function and respiratory function.

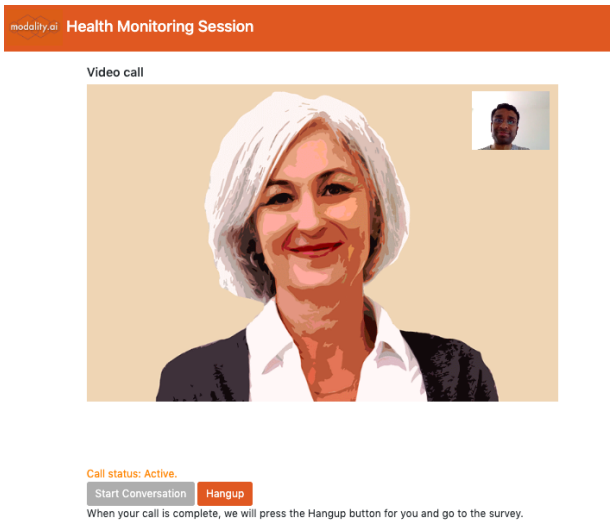

Figure 1: *Screenshot of the calling page during an interactive session.*

3. Conversational Protocol

The clinician can have the virtual agent engage patients in various combinations of exercises that have been widely

used in both speech language pathology research as well as clinical practice, including, but not limited to:

- an oral motor exam to assess the range and speed of motion of various orofacial articulators
- sustained phonation exercises, to investigate source properties such as fundamental frequency of phonation
- diadochokinesis exercises that investigate sequential and alternating motion rates of articulation
- read speech, including isolated words, sentences and read passages such as the Rainbow Passage [2] or the Bamboo Passage [3]
- spontaneous speech prompts
- spirometric exercises to measure respiratory function and lung volume, such as exhalation and coughing
- picture description, emotion elicitation and other cognitive tasks

4. Analytics

Analytics modules extract multiple speech (for instance, speaking rate, duration measures, F0, etc.) and video features (such as range and speed of movement of various facial landmarks) and store them in a database, along with information about the interaction itself such as the captured user responses, call duration, completion status, etc. All this information can be accessed by the clinicians after the interaction is completed through an easy-to-use dashboard (Figure 2) which provides a high-level overview of the various aspects of the interaction (including the video thereof and analytic measures computed), as well as a detailed breakdown of the individual sessions and the underlying interaction turns. The extracted features have been shown to correlate with different neurological conditions, such as depression or amyotrophic lateral sclerosis, by a number of studies, e.g. [4, 5, 6, 7, 8]. Although these studies analyzed data either offline or in controlled laboratory environments, which might differ significantly from the real-world data obtained and analyzed by NEMSI, a recent study showed preliminary evidence that the features extracted through NEMSI show similar correlations [9].

5. Conclusion

We demonstrated NEMSI, a multimodal conversational platform for remote patient diagnosis and monitoring, which extracts a variety of biomarkers while engaging patients in an interactive dialog session. The obtained biomarkers have been shown to be useful for a number of neurological conditions and are visualized in a user-friendly dashboard for further analysis. Further information, including a demo video of the system capabilities, can found at `www.modality.ai`.

Patient Id	Access code	Patient type	Provide access	Session Date (UTC)	ALSFRS-R score	Speaking rate (words/minute including pauses)	Articulation rate (words/minute excluding pauses)
[redacted]	j462w0	Crowdsourced testers (xx2x)	Re-invite	11/27/2019 18:44	44/44	211.99	231.68
[redacted]	w9egv3	Modality tester (xx0x)	Invite	02/08/2020 17:52			
[redacted]	4f03au	ePHI (xx3x)	Re-invite	01/06/2020 20:49	No survey responses (1)	No speech metrics (1)	
[redacted]	t3cx06	Modality tester (xx0x)	Re-invite	12/16/2019 22:46	48/48	222.02	297.17
[redacted]	sr1cyz	Internal tester (xx1x)	Re-invite	01/24/2020 23:50	48/48	214.39	232.08
[redacted]	cc0db8	Modality tester (xx0x)	Re-invite	11/21/2019 23:33	No survey responses (4)	150.56	170.32
[redacted]	xt034d	Modality tester (xx0x)	Re-invite	02/03/2020 23:20	No survey responses (4)	No speech metrics (2)	
[redacted]	o6gjxq	Internal tester (xx1x)	Re-invite	01/21/2020 23:10	48/48	145.63	197.33

Figure 2: *Screenshot of the dashboard showing a number of sessions with corresponding details including analytical measures, e.g. articulation rate.*

6. References

[1] D. Suendermann-Oeft, A. Robinson, A. Cornish, D. Habberstad, D. Pautler, D. Schnelle-Walka, F. Haller, J. Liscombe, M. Neumann, M. Merrill *et al.*, "Nemsi: A multimodal dialog system for screening of neurological or mental conditions," in *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 2019, pp. 245–247.

[2] G. Fairbanks, *Voice and Articulation Drillbook*, 2nd ed. Harper & Row, 1960, pp. 124–139.

[3] J. R. Green, D. R. Beukelman, and L. J. Ball, "Algorithmic estimation of pauses in extended speech samples of dysarthric and typical speech," *J Med Speech Lang Pathol*, vol. 12, no. 4, pp. 149–154, 2004.

[4] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 829–837, 2000.

[5] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on Multimodal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[6] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 87–91.

[7] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 43–50.

[8] A. Bandini, J. R. Green, L. Zinman, and Y. Yunusova, "Classification of bulbar als from kinematic features of the jaw and lips: Towards computer-mediated assessment," in *Interspeech*, Stockholm, Sweden, August 2017.

[9] M. Neumann, O. Roesler, D. Suendermann-Oeft, and V. Ramanarayanan, "On the utility of audiovisual dialog technologies and signal analytics for real-time remote monitoring of depression biomarkers," in *1st Workshop on NLP for Medical Conversations at ACL 2020*, Seattle, US, July 2020.