SENSAY ANALYTICS™: A REAL-TIME SPEAKER-STATE PLATFORM

A. Tsiartas[1], C. Albright[2], N. Bassiou[1], M. Frandsen[2], I. Miller[2], E. Shriberg[1], J. Smith[1], L. Voss[2], V. Wagner[2]

[1]Speech Technology and Research (STAR) Laboratory
[2]Center for Software Engineering
SRI International

andreas.tsiartas@sri.com

ABSTRACT

Growth in voice-based applications and personalized systems has led to increasing demand for speech- analytics technologies that estimate the state of a speaker from speech. Such systems support a wide range of applications, from more traditional call-center monitoring, to health monitoring, to human-robot interactions, and more. To work seamlessly in real-world contexts, such systems must meet certain requirements, including for speed, customizability, ease of use, robustness, and live integration of both acoustic and lexical cues. This demo introduces SenSay Analytics™, a platform that performs real-time speaker-state classification from spoken audio. SenSay is easily configured and is customizable to new domains, while its underlying architecture offers extensibility and scalability.

Index Terms— speech analytics, emotion detection, speaker-state analysis, affective computing, social signal processing

1. INTRODUCTION

Speech-analytics systems provide estimates of speaker states and traits by modeling information from both what people say and how they say it. The applications of speech analytics are broad and diverse, and are growing rapidly (along with image-based analytics) as personalization and affective computing [1] become increasingly important in today's technologies. Sample applications for speech analytics include call-center monitoring [2]; automatic emotion assessment [3, 4, 5, 6, 7] for personal assistants or human-robot interactions [8]; mental-health monitoring for depression or post-traumatic stress [9, 10, 11, 12, 13, 14, 15, 16]; automatic assessment of collaboration [17]; and many others. Such applications are often computationally demanding for a speech-analytics system; hence, an ideal speech-analytics setup must be resource efficient. Moreover, many applications, such as continuous monitoring in the mental-health domain, require the system to run continuously, and thus, system reliability, stability and scalability are critical. The diversity of applications also requires such systems to be configurable, customizable, and adaptable.

2. SENSAY ANALYTICS™

This demo presents SenSay Analytics™, a new speech technology from SRI International that performs real-time speaker-state classification from spoken audio[1]. Current applications, as shown in Fig. 1, include detection and assessment of emotion, sentiment, cognition, health, mental health, and communication quality in a range of end-use domains.

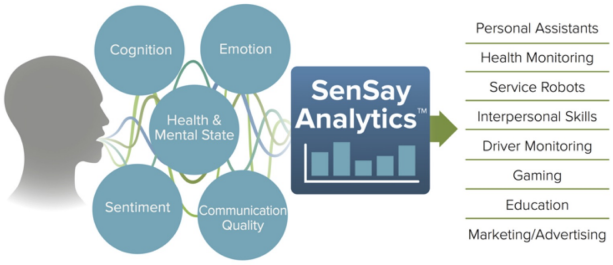

Fig. 1. Application areas

The platform performs not only real-time feature extraction but also real-time classification, updating both features and class estimates at sub-second intervals. It uses advanced signal features that capture spectral, prosodic, articulatory, auditory, discourse, and fluency characteristics, as well as features designed specifically for robustness to noise and reverberation (e.g., [9]). The platform analyzes the features from the signal alone or can be combined with automatic speech recognition (from either SRI[2] or a third-party) to model lexical information. The extracted features (acoustic and lexical) are modeled by state-of-the-art machine learning approaches.

[1]http://www.sensay-analytics.com/
[2]https://dev-portal.api.sri.com/#/dynaspeak

3. PERFORMANCE

The system is currently being benchmarked for a range of applications in both research and industry. Early analysis of results using the live system on a small set of standard data-community evaluations shows performance approximately on par with literature reports for offline systems. Results for emotion classification on a proprietary dataset with a large and diverse set of speakers and recording setups suggest performance about half-way between chance performance and perfect performance, where perfect performance overestimates human inter-annotator agreement. Some of the system features appear particularly strong for robustness to noise and reverberation [9], both of which are crucially relevant to real-world contexts. Further work is underway and will be presented at the demo.

4. ADVANTAGES

The platform is specifically designed to provide the robustness and speed required for supporting continuous monitoring and handling of big data. It operates at speeds crucial for safety applications (such as driving), Interactive Voice Response (IVR) systems, or personal assistant technologies, and it requires no pre-existing segmentation. The platform is deployable to cloud environments, personal computers, and client-hosted machines. This offers flexibility to customers who want to run the system locally on their own premises due to privacy concerns. Another important feature of the platform is its customizability: it can be adapted for specific tasks, domains, languages and for use with single- or multi-party conversations. Additionally, the platform's architecture supports the processing of simultaneous live streams.

5. REFERENCES

[1] R.W. Picard, *Affective Computing*, MIT Press, 1997.

[2] C. Vaudable and L. Devillers, "Negative emotions detection as an indicator of dialogs quality in call centers," in *Proc. ICASSP*, March 25-30 2012, pp. 5109–5112.

[3] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing realism in spontaneous emotion recognition from speech: feature enhancement by autoencoder with lstm neural networks," in *Proc. INTERSPEECH*, 2016, pp. 3593–3597.

[4] S.M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Proc. ACII*, 2015, pp. 125–131.

[5] Z. Yang and S. Narayanan, "Analyzing temporal dynamics of dyadic synchrony in affective interactions," in *Proc. INTERSPEECH*, 2016, pp. 42–46.

[6] B. Schuller and A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing*, Chichester, UK: Wiley, 2014.

[7] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[8] T. Chaspari and J.F. Lehman, "An acoustic analysis of child-child and child-robot interactions for understanding engagement during speech-controlled computer games," in *Proc. INTERSPEECH*, September 8-12 2016, pp. 595–599.

[9] V. Mitra, A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," in *Proc. ICASSP*, 2016, pp. 5795–5799.

[10] J. Gideon and M. McInnis E. Mower Provost, "Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder," in *Proc. ICASSP*, 2016, pp. 2359–2363.

[11] D. Le, K. Licata, C. Persad, and E. Mower Provost, "Automatic assessment of speech intelligibility for individuals with aphasia," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2187–2199, 2016.

[12] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T.F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, July 2015.

[13] D. Vergyri, B. Knoth, E. Shriberg, V. Mitra, M. McLaren, L. Ferrer, P. Garcia, and C. Marmar, "Speech-based assessment of PTSD in a military population using diverse feature classes," in *Proc. INTERSPEECH*, 2015, pp. 3729–3733.

[14] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling spectral variability for the classification of depressed speech," in *Proc. INTERSPEECH*, 2013, pp. 857–861.

[15] D.E. Sturim, P.A. Torres-Carrasquillo, T.F. Quatieri, and A. McCree, "Automatic detection of depression in speech using gaussian mixture modeling with factor analysis," in *Proc. INTERSPEECH*, August 2011, pp. 27–31.

[16] J.R. Orozco-Arroyave, J.C. Vásquez-Correa, F. Hönig, J.D. Arias-Londono, J.F. Vargas-Bonilla, S. Skodda, J. Rusz, and E. Nöth, "Towards an automatic monitoring of the neurological state of parkinson's patients from speech," in *Proc. ICASSP*, 2016, pp. 6490–6494.

[17] N. Bassiou, A. Tsiartas, J. Smith, H. Bratt, and C. Richey, "Privacy-preserving speech analytics for automatic assessment of student collaboration," in *Proc. INTERSPEECH*, September 8-12 2016, pp. 888–892.