Real-time Robust Recognition of Speakers' Emotions and Characteristics on Mobile Platforms

Florian Eyben, Bernd Huber, Erik Marchi, Dagmar Schuller, Björn Schuller
audEERING UG, Gilching, Germany
Email: fe@audeering.com

Abstract—**We demonstrate audEERING's sensAI technology running natively on low-resource mobile devices applied to emotion analytics and speaker characterisation tasks. A showcase application for the Android platform is provided, where audEERING's highly noise robust voice activity detection based on Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) is combined with our core emotion recognition and speaker characterisation engine natively on the mobile device. This eliminates the need for network connectivity and allows to perform robust speaker state and trait recognition efficiently in real-time without network transmission lags. Real-time factors are benchmarked for a popular mobile device to demonstrate the efficiency, and average response times are compared to a server based approach. The output of the emotion analysis is visualized graphically in the arousal and valence space alongside the emotion category and further speaker characteristics.**

openSMILE; Android; Mobile computing; Paralinguistics; Emotion; Affect -

I. Introduction

It is broadly agreed upon that state and trait recognition from speech bears interest for mobile services, albeit being challenging, e. g., due to noise and low-quality audio [1], [2], [3], [4], [5]. Few advanced products already exist and can be used through nowadays smartphones such as Beyond Verbal's app "Moodies" or EI Technologies's app "Xpression". Both applications analyse speech and indicate changes to allow training and improvement for its users. With such mobile services, manifold applications of computational paralinguistics, i. e., automatic recognition of speakers' states and traits, can be enhanced due to the mobile and ever-present character of off-line-enabled smart-phone analysis. To illustrate this, let us give but three examples:

First, a job applicant could use such an app not only to train for a job interview and improve the aspects of self-assurance and determination in order to have a higher chance of securing the job [6], but could even secretly use such monitoring during the interview. Second, giving online and in situ rhetorical feedback for different purposes becomes a reality. Third, health is another sector for utilising the possibilities of mobile off-line enabled speech monitoring. El Technologies Xpression app was designed to record mood changes throughout the day for people with anxiety, depression or stress to support the psychologists for their patient's treatment fine-tuning. It records five emotional stages throughout the day and emails the list to the psychologist at the end of the day. With off-line processing ability, a new possibility could be triggering help or warning patients in extreme stress situations [7] by detecting their emotional stage and reacting based on it. Also, a mobile coach for weight-loss that reacts when eating condition is detected from sound seems an interesting field of a product application [8].

In the remainder of this paper, we introduce our implementation (Section II), and show experimental benchmarking results (Section III) before drawing conclusions (Section IV) from these.

II. Implementation

Most current mobile speech processing applications, such as ASR (e. g., Google's voice search), rely on a frontend-backend architecture and require internet access for the services to operate [9], [10], [11]. This has the great advantage of flexibility, as the frontend can be kept light-weight and generic, and all the specific, heavy weight processing can be done on the backend. The frontend typically includes functionality to record, optionally compress, and finally transmit (or stream while recording) audio to the backend over a network/internet connection. The backend receives the recorded audio snippets or streams, decodes them, and runs the requested analysis, e. g., feature extraction and classification of speech emotions and speaker characteristics in our case. Such a distributed approach has been shown feasible including feature vector compression for computational paralinguistics in [12], [13].

In this demo-showcase we show the world's first working *industrial* prototype of a novel embedded audio analysis engine, where all steps are run entirely on a mobile phone and no connectivity to a backend is required. In fact, only related academic prototypes have previously been introduced in the literature, e. g., [14], where speakers' emotions are detected locally on off-the-shelf mobile phones.

For on device feature extraction we use a modified version of the popular feature extractor openSMILE [15], which we have compiled natively for the Android platform (ARM processor).

The demonstrator Android App is capable of recording speech, either in a hands-free mode or in a push-to-talk mode. In case of the hands-free mode, a robust voice activity detector based on memory enhanced neural networks [16] is used to detect segments which contain speech (separated by speech pauses). The detector described in [16] (also included in the latest version of openSMILE, see [17]) which has been optimised computationally to run well faster than real-time on

978-1-4799-9953-8/15/$31.00 ©2015 IEEE

the mobile device. Acoustic features are then extracted only for speech segments and speaker's characteristics and emotional state are classified via Support-Vector Machines (SVMs) and Support Vector Regression (SVR).

The observed vocal emotional state of the speaker as well as its history over the previous speech segments are shown graphically as dots in the upper half of the phone screen in a 2D activation valence coordinate system. Speaker characteristics such as gender are shown in the bottom half. Internally, the output can be further encoded in EmotionML [18] ready for usage, e. g., in other applications on the phone.

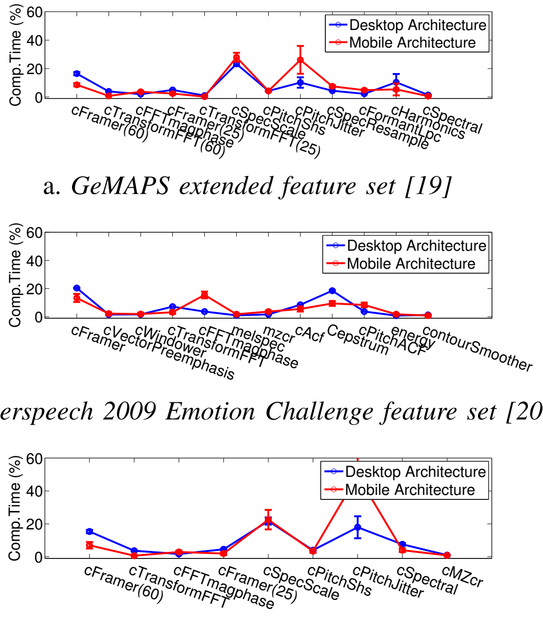
III. Experiments and Results

In order to verify the performance of the on-device feature extraction and classification, run-time benchmarks were conducted for three different acoustic feature sets which are part of the openSMILE package: the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [19], the Interspeech 2009 Emotion Challenge baseline feature set [20], and the Interspeech 2013 Computational Paralinguistics Challenge (ComParE) baseline feature set [21]. The motivation of their choice is a) their free-availability and well-defined standard, b) their increasing size from below 100 to several 1 000 features (cf. Table I), and c) the fact that with these sets, one can reproduce the same benchmark results in terms of recognition rates as were reached in the baselines and above of the challenge events the Interspeech sets have been designed for: These tasks include speaker's cognitive load, physical load, social signals such as laughter and breathing, conflict, up to 12-class emotion, eating condition, degree of nativeness, and for health applications recognition of autism manifestations, and Parkinsons disease [21], [22], [23]. In addition, the former Interspeech challenge task results can be reproduced with improved results due to the larger feature spaces, namely speaker's age, gender, five dimensional personality profile, degree of likability, intelligibility, intoxication, sleepiness, and level of interest [24].

Table I lists average real-time factors (over 50 runs) for extracting features from a 120 second audio file (uncompressed), both on the mobile phone (ARM processor) and a standard pc platform (Intel processor). The results show that the Interspeech 2009 Emotion Challenge feature set can be extracted easily in real-time, which is visible in the demo app by very low response times. Extracting the other two features sets on the device is also feasible; however, the response time (after the end of a speech segment) will be approximately 1.5 times the length of the original recording. Preliminary benchmarking of the three feature sets on a recent high-end phone (e. g., HTC One M9) demonstrated that all three feature sets can be extracted with a real-time factor smaller than 1.0 on this device, enabling low-lag real-time demonstration.

TABLE I
Benchmarks (real-time-factor – RTF) of feature extraction (from uncompressed waveform audio) on a Samsung Galaxy S3 Android (4.3) phone (mobile) and an Intel Core-i3 (2.1 GHz) laptop (pc).

Feature set	# feats.	RTF (mobile)	RTF (pc)
GeMAPS extended	88	2.63	0.11
Interspeech Emotion Challenge	384	0.43	0.04
Intespeech ComParE	6 373	2.81	0.12



a. *GeMAPS extended feature set [19]*

b. *Interspeech 2009 Emotion Challenge feature set [20]*

c. *Interspeech ComParE feature set [21]*

Fig. 1. Relative run-time (of total run-time) of feature extractor components on a desktop architecture (pc) and a mobile device (Galaxy S3).

Figure 1 (a-c) shows that various feature extraction components of three standard acoustic feature sets perform differently well (in relation to others) on the mobile phone. Computation of Voice Quality parameters such as Jitter seems to be slower on the phone than other parameters such as cepstral parameters. Supposedly this is due to bottelnecks when accessing memory, as Jitter is computed on the raw waveform data (samples). In future work, the performance could be improved by optimising memory access for the ARM environment, eventually making it possible to extract feature sets such as the Interspeech ComParE feature set in real-time.

IV. Conclusion

We demonstrate the feasibility of mobile recognition of speakers' emotions and characteristics on a mobile platform such as a modern, high-end smart phone in real-time through a standalone Android application.

In future efforts we want to benchmark the resource requirements of the application in terms of battery demand and investigate power saving strategies for using applications based on our sensAI technology in mobile 24/7 monitoring use-cases.

References

[1] K.-K. Lee, Y.-H. Cho, and K.-S. Park, “Robust feature extraction for mobile-based speech emotion recognition system,” in *Intelligent Computing in Signal Processing and Pattern Recognition*. Springer, 2006, pp. 470–477.

[2] W.-J. Yoon, Y.-H. Cho, and K.-S. Park, “A study of speech emotion recognition and its application to mobile services,” in *Ubiquitous Intelligence and Computing*. Springer, 2007, pp. 758–766.

[3] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, “A database of age and gender annotated telephone speech.” in *LREC*, 2010.

[4] A. Tawari and M. Trivedi, “Speech emotion analysis in noisy real-world environment,” in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 4605–4608.

[5] C. McCool, S. Marcel, A. Hadid, M. Pietikainen, P. Matejka, J. Cernocky, N. Poh, J. Kittler, A. Larcher, C. Levy *et al.*, “Bi-modal person recognition on a mobile phone: using mobile phone data,” in *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*. IEEE, 2012, pp. 635–640.

[6] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard *et al.*, “The tardis framework: intelligent virtual agents for social coaching in job interviews,” in *Advances in Computer Entertainment*. Springer, 2013, pp. 476–491.

[7] K.-h. Chang, D. Fisher, J. Canny, and B. Hartmann, “How’s my mood and stress?: an efficient speech analysis library for unobtrusive monitoring on mobile phones,” in *Proceedings of the 6th International Conference on Body Area Networks*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2011, pp. 71–77.

[8] O. Amft, M. Stäger, P. Lukowicz, and G. Tröster, “Analysis of Chewing Sounds for Dietary Monitoring,” in *UbiComp 2005: Proceedings of the 7th International Conference on Ubiquitous Computing*, Tokyo, Japan, 2005, pp. 56–72.

[9] Z.-H. Tan and B. Lindberg, *Automatic speech recognition on mobile devices and over communication networks*. Springer Science & Business Media, 2008.

[10] G. Di Fabbrizio, T. Okken, and J. G. Wilpon, “A speech mashup framework for multimodal mobile services,” in *Proceedings of the 2009 international conference on Multimodal interfaces*. ACM, 2009, pp. 71–78.

[11] A. Kumar, A. Tewari, S. Horrigan, M. Kam, F. Metze, and J. Canny, “Rethinking speech recognition on mobile devices,” in *Proc. ACM Conf. on IUI 2011*. Citeseer, 2011.

[12] W. Han, Z. Zhang, J. Deng, M. Wöllmer, F. Weninger, and B. Schuller, “Towards Distributed Recognition of Emotion in Speech,” in *Proceedings 5th International Symposium on Communications, Control, and Signal Processing, ISCCSP 2012*. Rome, Italy: IEEE, May 2012, pp. 1–4.

[13] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, “Distributing Recognition in Computational Paralinguistics,” *IEEE Transactions on Affective Computing*, 2014, 14 pages, to appear.

[14] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, “Emotionsense: a mobile phones based adaptive platform for experimental social psychology research,” in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010, pp. 281–290.

[15] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” in *Proc. of MM*. ACM, pp. 835–838.

[16] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, “Real-life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies,” in *Proceedings 38th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2013*. Vancouver, Canada: IEEE, 2013, pp. 483–487.

[17] F. Eyben and B. Schuller, “openSMILE: The Munich Open-Source Large-Scale Multimedia Feature Extractor,” *ACM SIGMM Records*, vol. 6, no. 4, December 2014.

[18] M. Schröder, L. Devillers, K. Karpouzis, J.-C. Martin, C. Pelachaud, C. Peter, H. Pirker, B. Schuller, J. Tao, and I. Wilson, “What should a generic emotion markup language be able to represent?” in *Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 440–451.

[19] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, and others., “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing,” *IEEE Transactions on Affective Computing (TAC)*, 2015, to appear.

[20] B. Schuller, S. Steidl, A. Batliner, and F. Jurcicek, “The INTERSPEECH 2009 Emotion Challenge,” in *Proc. of INTERSPEECH 2009*, Brighton, UK, Sep. 2009, pp. 312–315.

[21] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, “The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism,” in *Proceedings INTERSPEECH 2013*. Lyon, France: ISCA, 2013, pp. 148–152.

[22] B. Schuller, S. Steidl, A. Batliner, J. Epps, F. Eyben, F. Ringeval, E. Marchi, and Y. Zhang, “The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load,” in *Proceedings INTERSPEECH 2014*. Singapore, Singapore: ISCA, 2014.

[23] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger, “The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson’s & Eating Condition,” in *Proceedings INTERSPEECH 2015*. Dresden, Germany: ISCA, September 2015, 5 pages, to appear.

[24] B. Schuller, “The Computational Paralinguistics Challenge,” *IEEE Signal Processing Magazine*, vol. 29, no. 4, pp. 97–101, July 2012.

[25] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, “Cooperative Learning and its Application to Emotion Recognition from Speech,” *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 115–126, 2015.

978-1-4799-9953-8/15/$31.00 ©2015 IEEE