Speech Analysis in the Big Data Era

Björn W. Schuller[1,2,3,4,5,6]

[1] University of Passau, Chair of Complex and Intelligent Systems, Passau, Germany
schuller@ieee.org
http://www.schuller.it
[2] Deparment of Computing, Imperial College London, London, UK
[3] audEERING UG, Gilching, Germany
[4] Joanneum Research, Graz, Austria
[5] CISA, University of Geneva, Geneva, Switzerland
[6] Harbin Institute of Technology, Harbin, People's Republic of China

Abstract. In spoken language analysis tasks, one is often faced with comparably small available corpora of only one up to a few hours of speech material mostly annotated with a single phenomenon such as a particular speaker state at a time. In stark contrast to this, engines such as for the recognition of speakers' emotions, sentiment, personality, or pathologies, are often expected to run independent of the speaker, the spoken content, and the acoustic conditions. This lack of large and richly annotated material likely explains to a large degree the headroom left for improvement in accuracy by todays engines. Yet, in the big data era, and with the increasing availability of crowd-sourcing services, and recent advances in weakly supervised learning, new opportunities arise to ease this fact. In this light, this contribution first shows the de-facto standard in terms of data-availability in a broad range of speaker analysis tasks. It then introduces highly efficient 'cooperative' learning strategies basing on the combination of active and semi-supervised alongside transfer learning to best exploit available data in combination with data synthesis. Further, approaches to estimate meaningful confidence measures in this domain are suggested, as they form (part of) the basis of the weakly supervised learning algorithms. In addition, first successful approaches towards holistic speech analysis are presented using deep recurrent rich multi-target learning with partially missing label information. Finally, steps towards needed distribution of processing for big data handling are demonstrated.

1 Introduction

Speech recognition has more and more found its way into our every day lives – be it when searching on small hand-held devices, controlling home-entertainment or entering, e. g., an address into a navigation system. This is yet to come for many other speech analysis tasks – in particular the 'paralinguistic' ones. There, the most

frequently encountered usage in day-to-day live is being the identification of the speaker per se, such as in some telephone banking settings. Next come, most likely – usually unnoticed, gender and age-group, e. g., in dialogue systems or simply to adapt the speech recogniser. A few applications, e. g., in video games such "Truth or Lies" promise to recognise decveptive speech or emotion. However, the plethora of other opportunities such as recognition of a speaker's personality, physical and mental load, health condition, eating condition, degree of nativeness, intoxication, or sleepiness have hardly found their way into applications noticed by the general public. While certainly of high usefulness if running properly, this raises the question on the cause which is likely the still too low reliability. While this is of course a matter of diverse factors such as the right pre-processing including de-noising and de-reverberation, optimal feature representation, optimal classification or regression and optimisation of models, the main bottleneck can likely be attributed to the sparseness of learning data for such systems. In comparison, a speech recogniser is partially being trained on more data than a human is exposed to throughout lifetime. For computational paralinguistic tasks, data often remains at the level of one up to a few hours and a handful to some hundred speakers. This data material is mostly annotated with a single phenomenon such as a particular speaker state at a time. In stark contrast to this, engines such as for the recognition of speakers' emotions, sentiment, personality, or pathologies, are often expected to run independent of the speaker, the spoken content, and the acoustic conditions. While one may argue that still, a human might not need as much data to learn certain paralinguistic characteristics as are needed to learn a whole language, clearly, more data are desired than are given at present – also as one may wish to aim at super-human abilities in some tasks. Three factors are mainly responsible for this sparseness of speech data and suited labels: the data are often 1) sparse per se, such as in the case of a sparsely occurring speaker state or trait, 2) considerably more ambiguous and thus challenging to annotate than, e. g., orthographic transcription of speech usually is, and 3) of highly private nature such as highly emotional or intoxicated data or such of speech disorders. Yet, in the big data era, it is becoming less and less the actual speech data that is lacking, as diverse resources such as the internet, broadcast, voice communication, and increased usage of speech-services including self-monitoring provide access to 'big' amounts. Instead, it is rather the labels that are missing. Luckily, with the increasing availability of crowd-sourcing services, and recent advances in weakly supervised, contextual, and reinforced learning, new opportunities arise to ease this fact.

In this light, this contribution first shows the de-facto standard in terms of data-availability in a broader range of speaker analysis tasks (Section 2). It then presents highly efficient 'cooperative' learning strategies basing on the combination of active and semi-supervised alongside transfer learning to best exploit available data (Section 3). Further, approaches to estimate meaningful confidence measures in this domain are suggested, as they form (part of) the basis of the weakly supervised learning algorithms (Section 4). In addition, first successful approaches towards holistic speech analysis are presented using deep recurrent rich multi-target learning with partially missing label information (Section 5).

Then, steps towards needed distribution of processing for big data handling are demonstrated (Section 6). Finally, some remaining aspects are discussed and conclusions are drawn (Seciton 8). Overall, a system architecture and methodology is thus discussed that holds the promise to lead to a major breakthrough in performance and generalization ability of tomorrow's speech analysis systems.

2 Data: The Availability-Shock

Few speech analysis tasks are lucky enough to have a day of labelled speech material available for training and testing of models. Taking the Interspeech 2009 – 2015 series of Computational Paralinguistics Challenges as a reference [1], one can see that, in fact, mostly around one or 'some' hours is all one is left with as a starting point to train a model for a new speech analysis task such as recognising Altzheimer, Autism, or Parkinson's Condition of a speaker. Obviously, one can hardly expect to train models independent of the speaker, language, cultural background, and co-influencing factors from such little data. Actually, some attempts at cross-corpus studies show the very weak generalisation observed for most systems trained in such a way (e. g., [2]).

3 On Efficiency: Learning Cooperatively

This reality of little labelled speech data, but availability of large(r) amounts of unlabelled such has led to a number of recent approaches in this field to most efficiently exploit both of these with little human labour involved.

3.1 Transfer Learning

Often, one has labelled data from a 'similar' domain or task available, such as recognising emotion of adult speakers, but little to no (labelled) data for the current situation of interest – let's say recognising emotion of children. In such a case, one can train a model that best learns how to 'transfer' the knowledge to the new domain, even if no labels are available at all in the new target domain [3]. An interesting further example has shown that this way, one can even train a model for the recognition of emotion in speech on music and then transfer this knowledge – in [4] this was reached by use of a sparse autoencoder that learns a compact representation of one of the domains (out of speech and music) to 'transfer' features to the respective other one. In [5] a more efficient approach was shown by training several autoencoders and to learn the differences with an additional neural network. Further, usage of related data for the initialisation of models such as in deep learning has been shown useful in general speech processing, e. g., in [6], but is less exploited in paralinguistics as of now.

3.2 (Dynamic) Active Learning

Better models can usually non-the-less be reached if one does label at least some data in the new domain or for the new task. To keep human efforts to a minimum, the computer can first decide which data points are of interest, such as by identifying 'sparse' instances such as emotional data (leaving it to the human to tell which emotion it is) versus 'non-emotional' data (which usually appears in much higher frequency and is thus 'less interesting' after some such data points have already been seen) [7,8]. Accordingly, rather than recognising different emotions, where data for each class may be to sparse, a coarser model is first chosen which is simply neutral versus non-neutral speech. In this sense, one can initialise an active learning system basically by collecting only emotionally neutral speech and then execute a novelty detection or alike for unlabelled data. As neutral emotional speech is available in large amounts or can even be synthesised [9], one can easily train such a 'one class' model by loads of data. Then, when newly seen speech is deviant in some form, a human can be asked for labelling aid. Other aspects can include the likely change of model parameters, i. e., the learning algorithm decides if the data would change its parameters significantly at all before asking for human aid on 'what it is'. Such approaches were also extended for actively learning regression tasks rather than discrete classes [10]. An interesting more recent option for fast labelling is crowd sourcing [11], as it offers to quickly reach a large amount of labellers (in fact often even in real-time which may become necessary when dealing with 'big' and growing amounts of data). However, as often laymen rather than experts in phonetics, linguistics, psychology, medicine or other related disciplines may be of relevance to the speech analysis task of interest form the majority of the crowd, one often needs a factor higher a number of labellers and has to cope with noisy labels. 'Learning' the labellers and dynamically deciding on how many labellers and 'whom to ask when' allows to source the crowd more efficiently [12].

3.3 Semi-supervised Learning

More efficiently, the computer can label data itself once it was trained supervised on some first data [13]. Obviously, this comes at a risk of labelling data erroneously and then re-training the system on partially noisy labels. Accordingly, one usually needs to make a decision based on some form of 'confidence measure' (cf. below) on whether to add a computer-labelled data instance to the learning material for (re-)training or not. In addition, one can use multiple 'views' in 'co-training' to decide on the labels of the data [14,15]. In [16] it was shown for a range of speech analysis tasks that this way, it is indeed possible to have a speech analysis system self-improve by giving it new (unlabelled) speech data observations.

3.4 Cooperative Learning

Putting the above two (i. e., active and semi-supervised learning) together leads to 'cooperative learning' [17]. The principle can best be described as follows:

For a new data instance, have the computer first decide if it can label it itself – if not, make a decision if it is worth or not to ask for human aid. [17] shows that this can be more efficient than any of the above to forms.

4 On Decision-Making: Learning Confidence Measures

As both, active, and semi-supervised learning mostly base on some confidence measure, it seems crucial to find ways of reliably estimating such. In stark contrast to speech recognition [18], there does not yet exist much literature on this topic for the field of paralinguistic speech analysis that would go beyond using a learning's algorithm inherent confidence such as the distance to the separating hyperplane of the winning class as compared to the next best class. As it is, however, this exact learning algorithm that makes the decision on a class and often does so wrongly, it seems more reliable to enable additional ways of measuring the confidence one has in a recognition result. Two ways have been shown recently in the field of Computational Paralinguistics partially exploiting the characteristics of this field.

4.1 Agreement-Based Confidence Measures

The first approach aims at estimating the agreement humans would likely have in judging the paralinguistic phenomenon of interest [19]. Thus, for a subjective task requiring several labellers such as emotion or likability of a speaker, one does not train the emotion class or degree of likability as target, but rather the percentage of human raters that agreed upon the label. Then, one can automatically estimate this percentage also for new speech data which serves as a measure on how difficult it is likely to assess a unique 'correct' label/opinion. Obviously, this can be interpreted as an indirect measure of confidence.

4.2 Learning Errors

Alternatively, one can train additional recognition engines alongside the paralinguistic engine 'in charge' using whether or not it made errors as learning target. If several such engines are trained on different data, their estimates can be used as confidence measure. In fact this measure's reliability can even be improved by semi-supervised learning [20].

5 On Seeing the Larger Picture: Learning Multiple Targets

As all our personal speaker traits as well as our multi-faceted state have an impact on the same voice production mechanism, it seems wise to attempt to see the 'larger picture'. Up to now, most work in the field of Computational Paralinguistics is pre-concerned with one phenomenon at a time such as recognition

of exclusively emotion or exclusively health state or exclusively personality. Obviously, however, the voice sounds different not only because one is angry, but also as one has a flu and depending on whether one is of open or less open personality. Most attempts to learn multiple targets in parallel such that the knowledge of each other target positively influences the overall recognition accuracy have so far been focused on learning several emotion primitives commonly, cf., e. g., [21]. However, more recent attempts aim at learning a richer variety of states and traits as multiple targets [22]. This introduces the challenge to find data that are labelled in such manifold states and traits – something hardly met these days. One can easily imagine how this raises the demand in the above described efficient ways of quickly labelling data by the aid of the crowd in intelligent ways. In addition, it requires learning algorithms not only able to learn with multiple targets, but likely also with partially missing such given that one will not always be able to obtain a broad range of attributes for any voice sample 'found' or newly observed by the computer.

6 On Big Data: Distribution

Referring to 'big data' usually goes along with the amount of data being so large that 'conventional' approaches of processing cannot be applied [23]. This may require partitioning of the data and distribution of efforts [24]. While a vast body of literature exists in the field of 'core' Machine Learning on how to best distribute processing, it will remain to tailor these approaches to the needs of speech analysis. Distributed processing has been targeted considerably for speech recognition, but hardly for paralinguistic tasks where only very first experiences are reported, e. g., on optimal compression of feature vectors [25].

7 Conclusion

The next major leap forward for the field of Computational Paralinguistics and the broader field of Speech Analysis can likely be expected to be made by overcoming the ever-present sparseness of learning data by making efficient use of the big amounts of available speech by adding rich amounts of labels to these likely with help from the crowd. Such resources will have a partially noisy gold standard thus requiring potentially larger amounts of labelled speech than if labelled by experts, but it will be easier to reach large amounts of data. In particular, these may be labelled by a multitude of information rather than targeting a single phenomenon such as emotion or sleepiness of a speaker at a time. The labelling effort will likely become manageable by pre-processing by a machine that makes first decisions on the interest of the data, but that also learns how many and which raters to ask in which situation, i. e., that is not only learning about the phenomena of interest but also about the crowd that helps it to learn about these. One can probably best depict this by the metaphor of a child that not only learns about its world, but also whom to best ask about which parts of it and sometimes to better inquire several opinions. If one does not need to know

"what is inside" the speech data, but simply makes further use of it, e. g., in a spoken dialogue system without attaching a human-interpretable label, unsupervised learning, i. e., clustering may be another suited variant [26,27] allowing for exploitation of big speech data.

Concerning technical necessities, efficient big data speech analysis will require reliable confidence measure estimation to decide which data to label by the computer and which by humans. Further, distributed processing may become necessary if data becomes 'too' large. Processing of big data handling comes with further new challenges such as sharing of the data and trained models, and ethical aspects such as privacy, transparency, and responsibility for what has been learnt by the machine once decisions are made [28,29].

On the other end of the 'big' scale, some speech analysis tasks will remain sparse in terms of data, e. g., for some pathological speech analysis tasks. Here, zero resource [30] or sparse resource approaches are an alternative to circumvent the data sparseness. Such approaches are known from speech recognition and keyword spotting and spoken term detection [31]. The usual application scenario there is to recognise words in 'new' spoken languages where only very sparse resources exist. For the recognition of paralinguistic tasks, an opportunity arises once at least something is known about the phenomenon of interest so to be able to implement rules such as "IF the speech is faster and the pitch is higher THAN the speaker is more aroused" etc.

As a final statement, one can easily imagine that these conclusions may hold in similar ways to a broader range of audio analysis tasks and in fact many other fields – the era of big data and increasingly autonomous machines exploring it has just begun.

Acknowledgments. The research leading to these results has received funding from the European Union's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu), No. 645094 (SEWA), No. 644632 (MixedEmotions), and No. 645378 (ARIA VALUSPA).

References

1. Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J.R., Nöth, E., Zhang, Y., Weninger, F.: The INTERSPEECH 2015 computational paralinguistics challenge: degree of nativeness, Parkinson's & eating condition. In: Proc. INTERSPEECH, Dresden, Germany, p. 5. ISCA (2015)
2. Devillers, L., Vaudable, C., Chastagnol, C.: Real-life emotion-related states detection in call centers: a cross-corpora study. In: Proc. INTERSPEECH, Makuhari, Japan, pp. 2350–2353. ISCA (2010)
3. Deng, J., Zhang, Z., Eyben, F., Schuller, B.: Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition. IEEE Signal Processing Letters **21**(9), 1068–1072 (2014)
4. Coutinho, E., Deng, J., Schuller, B.: Transfer learning emotion manifestation across music and speech. In: Proc. IJCNN, Beijing, China, pp. 3592–3598. IEEE (2014)
5. Deng, J., Zhang, Z., Schuller, B.: Linked source and target domain subspace feature transfer learning - exemplified by speech emotion recognition. In: Proc. ICPR, Stockholm, Sweden, pp. 761–766. IAPR (2014)

6. Swietojanski, P., Ghoshal, A., Renals, S.: Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In: Proc. Spoken Language Technology Workshop (SLT), Miama, FL, pp. 246–251. IEEE (2012)
7. Bondu, A., Lemaire, V., Poulain, B.: Active learning strategies: a case study for detection of emotions in speech. In: Perner, P. (ed.) ICDM 2007. LNCS (LNAI), vol. 4597, pp. 228–241. Springer, Heidelberg (2007)
8. Zhang, Z., Deng, J., Marchi, E., Schuller, B.: Active learning by label uncertainty for acoustic emotion recognition. In: Proc. INTERSPEECH, Lyon, France. ISCA, pp. 2841–2845 (2013)
9. Lotfian, R., Busso, C.: Emotion recognition using synthetic speech as neutral reference. In: Proc. ICASSP, Brisbane, Australia, pp. 4759–4763. IEEE (2015)
10. Han, W., Li, H., Ruan, H., Ma, L., Sun, J., Schuller, B.: Active learning for dimensional speech emotion recognition. In: Proc. INTERSPEECH, Lyon, France, pp. 2856–2859. ISCA (2013)
11. Callison-Burch, C., Dredze, M.: Creating speech and language data with Amazon's mechanical turk. In: Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, CA, pp. 1–12. ACL (2010)
12. Zhang, Y., Coutinho, E., Zhang, Z., Quan, C., Schuller, B.: Agreement-based dynamic active learning with least and medium certainty query strategy. In: Proc. Advances in Active Learning: Bridging Theory and Practice Workshop held in conjunction with ICML, Lille, France, p. 5. IMLS (2015)
13. Yamada, M., Sugiyama, M., Matsui, T.: Semi-supervised speaker identification under covariate shift. Signal Processing **90**(8), 2353–2361 (2010)
14. Liu, J., Chen, C., Bu, J., You, M., Tao, J.: Speech emotion recognition using an enhanced co-training algorithm. In: Proc. ICME, Beijing, P.R. China, pp. 999–1002. IEEE (2007)
15. Jeon, J.H., Liu, Y.: Semi-supervised learning for automatic prosodic event detection using co-training algorithm. In: Proc. Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 2, pp. 540–548. ACL, Stroudsburg (2009)
16. Zhang, Z., Deng, J., Schuller, B.: Co-training succeeds in computational paralinguistics. In: Proc. ICASSP, Vancouver, Canada, pp. 8505–8509. IEEE (2013)
17. Zhang, Z., Coutinho, E., Deng, J., Schuller, B.: Cooperative Learning and its Application to Emotion Recognition from Speech. IEEE/ACM Transactions on Audio, Speech and Language Processing **23**(1), 115–126 (2015)
18. Jiang, H.: Confidence measures for speech recognition: A survey. Speech communication **45**(4), 455–470 (2005)
19. Deng, J., Han, W., Schuller, B.: Confidence measures for speech emotion recognition: a start. In: Fingscheidt, T., Kellermann, W. (eds.) Proc. Speech Communication. 10. ITG Symposium, Braunschweig, Germany, pp. 1–4. IEEE (2012)
20. Deng, J., Schuller, B.: Confidence measures in speech emotion recognition based on semi-supervised learning. In: Proc. INTERSPEECH, Portland, OR. ISCA (2012)
21. Eyben, F., Wöllmer, M., Schuller, B.: A Multi-Task Approach to Continuous Five-Dimensional Affect Sensing in Natural Speech. ACM Transactions on Interactive Intelligent Systems **2**(1), 29 (2012)
22. Schuller, B., Zhang, Y., Eyben, F., Weninger, F.: Intelligent systems' Holistic evolving analysis of real-life universal speaker characteristics. In: Proc. 5th Int. Workshop on Emotion Social Signals, Sentiment & Linked Open Data (ES^{3}LOD 2014), satellite of LREC, Reykjavik, Iceland, pp. 14–20. ELRA (2014)

23. Madden, S.: From databases to big data. IEEE Internet Computing **3**, 4–6 (2012)
24. Chen, M., Mao, S., Liu, Y.: Big data: A survey. Mobile Networks and Applications **19**(2), 171–209 (2014)
25. Zhang, Z., Coutinho, E., Deng, J., Schuller, B.: Distributing Recognition in Computational Paralinguistics. IEEE Transactions on Affective Computing **5**(4), 406–417 (2014)
26. Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., Cowie, R.: Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In: Proc. INTERSPEECH, Brisbane, Australia, pp. 597–600. ISCA (2008)
27. Zhang, Y., Glass, J.R.: Towards multi-speaker unsupervised speech pattern discovery. In: Proc. ICASSP, Dallas, TX, pp. 4366–4369. IEEE (2010)
28. Richards, N.M., King, J.H.: Big data ethics. Wake Forest L. Rev. **49**, 393 (2014)
29. Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. IEEE Transactions on Knowledge and Data Engineering **26**(1), 97–107 (2014)
30. Harwath, D.F., Hazen, T.J., Glass, J.R.: Zero resource spoken audio corpus analysis. In: Proc. ICASSP, Vancouver, BC, pp. 8555–8559. IEEE (2013)
31. Jansen, A., Dupoux, E., Goldwater, S., Johnson, M., Khudanpur, S., Church, K., Feldman, N., Hermansky, H., Metze, F., Rose, R., et al.: A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition, pp. 8111–8115 (2013)