Chapter 1
Introduction

In our daily lives we are interacting with a constantly growing number of computerised devices, such as in-car computers, electronic vending machines, automated telephone call-centres, smartphones, and wearable computing devices. In order to make the interaction with these devices as simple and intuitive as possible, more natural interaction methods are essential (Pentland 2005; Sinha et al. 2010). Currently, most man-machine interfaces rely on haptic input, i. e., keyboard, mouse, and recently touchscreens (cf. Grudin 2007; Schenk and Rigoll 2010). For every new device the user has to learn how to operate it by reading the manual, for example. In the future, computing and communication devices will be even more ubiquitous and high-functionality (Sinha et al. 2010), making it more and more important to have intuitive interfaces, which accept visual and spoken input, and with which we can communicate naturally and intuitively just like with any other person—without reading a manual first (Sinha et al. 2010).

An important ingredient for more natural man-machine interaction is to enable machines to analyse and understand human intentions and social behaviour as well as humans do. This is a very complex and ambitious goal that we are, yet, far from reaching (Pentland 2005). To reach such an ambitious goal, machines have to be able to analyse social (Schuller et al. 2008; Vinciarelli et al. 2009), natural affective (Zeng et al. 2009; Schuller et al. 2009; Batliner et al. 2011), paralinguistic (Schuller and Batliner 2013), and contextual signals (Vinciarelli et al. 2009), for example. Further, machines have to be able to deal with ambiguities (ter Maat and Heylen 2009), and robustly interpret all input and deduce the correct intention of the user from them, i. e., the machines should behave in a way that would be considered socially intelligent (Schroeder and McKeown 2010).

F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, Springer Theses, DOI 10.1007/978-3-319-27299-3_1

1.1 Acoustic Analysis of Speech and Music

This thesis focuses on the analysis of human signals, limited to the acoustic channel, in particular speech and music, thereby connecting fields of automatic speech recognition (cf. e. g., Furui 1996) and Music Information Retrieval (MIR) (cf. Bello et al. 2000; Fingerhut 2004). The human voice is presumably the most important mean of direct human to human communication: Speech carries linguistic content, i. e., sentences and words, and paralinguistic content (Schuller and Batliner 2013), such as mood, affect, speaker states such as intoxication and sleepiness, and speaker traits such as age, gender, and personality (Schuller and Batliner 2013). As was shown by the author of this thesis in other studies, speech possibly also carries information about eye contact in dyadic conversations (Eyben et al. 2013b), or heart-rate and skin conductivity in some cases (Schuller et al. 2013). Further, the average subjective likeability of a voice was estimated automatically in Eyben et al. (2013a). Singing and—more generally—music allows for an even better expression of mood and emotions, in particular, through rhythm and melody.

From a technical point of view (cf. Schuller 2013), speech and music are acoustic signals, represented in the physical world by micro variations of pressure, mostly air pressure, in the range from approx. 50–8,000 Hz. When speech or music are recorded with a microphone, the air pressure modulation is converted to an electrical signal. In digital sound processing, as this thesis deals with exclusively, the analogue electrical signal is digitised by analogue digital conversion and sampling. Digital signal processing techniques (Oppenheim and Schafer 1975) are then applied to the audio samples to reduce the amount of information to the aspects of interest, and discard unnecessary or redundant information. This process is known as feature extraction (cf. e. g., Ruske 1993).

1.2 Deficiencies of the State-of-the-Art

While the automatic, computer-based recognition of the linguistic content from speech has matured over three decades of research (cf. Rabiner 1989; Furui 1996; Dahl et al. 2012), the analysis of speech paralinguistics (Schuller and Batliner 2013) and high level music attributes such as mood (e. g., Schmidt and Kim 2010) is a comparably young field (Schuller 2013).

Automatic recognition of emotions, for example, in acted speech and in acoustic laboratory conditions, i. e., on the same database, where training and test data have the same acoustic conditions, was shown to be at human performance levels (e. g., Schuller 2006). However, recognition of affect and paralinguistic information "in the wild" (Dhall et al. 2013), i. e., in noisy and reverberated acoustic conditions, on spontaneous speech, and in a continuum of non-prototypical, real-life emotions, or in music, for example, remains a challenge (cf. e. g., You et al. 2006; Schuller et al. 2006, 2007; Schmidt and Kim 2010). This is most evident, if the good performance

of, e. g., emotion classification within a corpus of prototypical emotions is contrasted to the performance obtained in cross corpus classification (e. g., Schuller et al. 2010; Eyben et al. 2010a), i. e., when training a model on one database, and evaluating on another which contains similar emotions but was likely recorded with a different protocol and in a different environment. Robustness under such conditions is very important for, e. g., use in in-car environments (Eyben et al. 2010b).

Experiments conducted so far, were mostly done using small acoustic feature sets (cf. e. g., Ververidis and Kotropoulos 2006), extracted with various toolkits, such as Praat (Boersma 2001) or Mirtoolbox (Lartillot and Toiviainen 2007), limiting the comparability and reproducibility of experiments across research sites.

1.3 Aims of This Thesis

In order to advance the field of paralinguistic speech and music analysis research, this thesis addresses the challenge of analysis of speech and music beyond the textual and musical note content. This means a higher level analysis based on the acoustic mid- and long-term properties, in order to identify paralinguistic attributes such as affect, or high level music attributes such as genre or mood. A large space of acoustic parameters is discussed and evaluated in this context, and—to encourage comparability and reproducibility of results—standard acoustic parameter sets are proposed and evaluated for voice based affect recognition and music style identification. Further, a strong emphasis is put on the applicability of the proposed methods to real-world use-cases. This includes discussion of real-time processing issues related to the signal processing, introduction of an incremental processing software-framework, and the evaluation of novel methods to increase noise robustness.

In particular, the following four major aims have been defined for this thesis:

1. Development of an open-source framework and toolkit for incremental speech and music analysis with a strong focus on real-time, incremental feature extraction algorithms and a broad coverage of acoustic descriptors,
2. Definition of novel, large-scale standard acoustic feature sets for incremental speech and music analysis,
3. Design of speech and music analysis methods which can robustly handle real-life conditions,
4. Evaluation of the framework, the methods, and the standard acoustic feature sets on a wide range of music classification and speech affect recognition tasks.

These aims are elaborated in more details in the following sections.

1.3.1 Real-time Analysis Framework

Compared to video/visual feature extraction, the computational complexity of audio feature extraction is significantly lower on average. Still, not all acoustic features can

be extracted in real-time, and some modifications to the overall chain of processing must be made in order to efficiently process inputs in real-time and—especially—obtain incremental classification results with a latency as low as possible. Also, no standard feature extraction toolkit which unites a large set of speech and music descriptors existed prior to this thesis. Thus, the first goal defined for this thesis in the scope of the SEMAINE EU-FP7 project[1] was to create a flexible and modular framework and toolkit for real-time, incremental audio feature extraction. The framework should be usable for both large-scale, efficient off-line batch processing for research, as well as for on-line, incremental processing in interactive speech dialogue systems, for example. Further, a large set of acoustic descriptors should be included and algorithms for real-time, incremental extraction of these features should be implemented, described, and made publicly available, in order to allow researchers to use a common implementation for many tasks, instead of a multitude of different libraries.

1.3.2 Baseline Feature Sets

Next, from the large set of acoustic descriptors implemented in the framework, high-dimensional feature sets based on feature brute-forcing should be proposed and published in order to set baseline standards for international evaluation campaigns and comparison of audio analysis and classification algorithms across the world. The feature sets should be shared freely with the research community, and should be extractable with an open-source toolkit (see previous aim, Sect. 1.3.1), in order to encourage reproducible research basing on standard acoustic features and standard feature extraction algorithms.

1.3.3 Real-World Robustness

As current methods mostly deal with laboratory conditions and show results on clean speech corpora, real-life conditions are mainly neglected. This thesis aims to propose (and evaluate) solutions to improve the analysis performance under real-world conditions, such as high levels of background noise, reverberation, or speaker variability. Further, another important issue to be addressed in this respect was the segmentation of continuous input audio streams, especially for real-time interactive systems. In contrast to textual speech recognition, in acoustic analysis of speech larger units of speech (Zeng et al. 2009), e. g., sentences or phrases (Batliner et al. 2010) are commonly used, and the performance of the whole approach is highly dependent on the segmentation chosen (Batliner et al. 2010). In most research the units are pre-segmented by hand and therefore all results obtained so far have the precondition of

[1] http://www.semaine-project.eu/.

perfect segmentation. In real-time, real-world settings, perfect segmentation is not possible due to limited context and often limited resources. Moreover, long segments such as sentences increase the latency of the analysis algorithm unnecessarily. Thus, for this thesis an alternate input segmentation method should be developed which is suitable for incremental processing with low latency.

1.3.4 Large-Scale Evaluation

The final, fourth goal, which links to the three previous goals, was to evaluate the proposed baseline feature sets, real-life robustness methods, and the quality of the feature extraction algorithms implemented for this thesis systematically on a vast number of speech and music analysis tasks. The goal thereby was to confirm the validity and suitability of the proposed feature sets, highlight the differences of the sets with respect to different tasks and data-sets, and investigate the influence of classifier parameters and variations of the training and evaluation procedure.

1.4 Overview

This thesis is grouped into five main Chaps. (2–6): Chap. 2 describes all the methods applied in this thesis for the automatic classification of speech and music signals. These include in particular the signal processing and audio feature extraction algorithms (Sect. 2.2) and the supra-segmental feature summarisation and large-scale feature brute-forcing (Sect. 2.4), as well as the static and dynamic modelling (machine learning) methods (Sect. 2.5).

The proposed standard baseline acoustic feature sets are introduced in Chap. 3, The description of the incremental processing framework developed for this thesis is given in Chap. 4 and the methods developed to increase real-world robustness of speech and music analysis are described in Chap. 5.

Chapter 6 describes the extensive, large-scale evaluations of the standard acoustic parameter sets conducted on several data-sets for the automatic recognition of speech affect categories, music styles, and fully time and value continuous affect recognition. Moreover, the effectiveness of the real-world robustness methods is evaluated on realistic data.

The findings of this thesis are summarised and concluded in Chap. 7. Further, it is critically evaluated how well the goals described in Sect. 1.3 were achieved and which topics are still open and deserve more attention in follow up work.

References

A. Batliner, D. Seppi, S. Steidl, B. Schuller, Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach. *Advances in Human Computer Interaction, Special Issue on Emotion-Aware Natural Interaction*, 2010, p. 15 (Article ID 782802). (on-line)

A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, N. Amir, The Automatic Recognition of Emotions in Speech, in *Emotion-Oriented Systems: The Humaine Handbook, Cognitive Technologies*, ed. by P. Petta, C. Pelachaud, R. Cowie (Springer, Berlin, 2011), pp. 71–99

J.P. Bello, G. Monti, M. Sandler. Techniques for automatic music transcription. In *Proceedings of the International Symposium on Music Information Retrieval (MUSIC-IR 2000)*. ISMIR, 2000, p. 8

P. Boersma, Praat, a system for doing phonetics by computer. Glot Intern. **5**(9/10), 341–345 (2001)

G. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Trans. Audio, Speech Lang. Process. **20**(1), 30–42 (2012)

A. Dhall, R. Goecke, J. Joshi, M. Wagner, T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM International conference on multimodal interaction (ICMI) 2013*, ACM. Sydney, Australia, pp. 509–516, December 2013

F. Eyben, A. Batliner, B. Schuller, D. Seppi, S. Steidl, Cross-Corpus Classification of Realistic Emotions—Some Pilot Experiments. In L. Devillers, B. Schuller, R. Cowie, E. Douglas-Cowie, A. Batliner, editors, *Proceedings of the 3rd International Workshop on EMOTION: Corpora for Research on Emotion and Affect, satellite of LREC 2010*, Valletta, Malta, pp. 77–82 May 2010a. European Language Resources Association (ELRA)

F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, N. NguyenspsThien, Emotion on the Road—Necessity, Acceptance, and Feasibility of Affective Computing in the Car. *Advances in Human Computer Interaction (AHCI), Special Issue on Emotion-Aware Natural Interaction*, 2010b. p. 17. doi:10.1155/2010/263593. Article ID 263593

F. Eyben, F. Weninger, E. Marchi, B. Schuller, Likability of human voices: A feature analysis and a neural network regression approach to automatic likability estimation. In *Proceedings of the 14th International Workshop on Image and Audio Analysis for Multimedia Interactive Services (WIAMIS) 2013*, IEEE. Paris, France, pp. 1–4 July, 2013a

F. Eyben, F. Weninger, L. Paletta, B. Schuller, The acoustics of eye contact—Detecting visual attention from conversational audio cues. In *Proceedings of the 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction (GazeIn '13), held in conjunction with the 15th International Conference on Multimodal Interaction (ICMI) 2013*, ACM. Sydney, Australia, pp. 7–12, December 2013b

M. Fingerhut, *Music information retrieval, or how to search for (and maybe find) music and do away with incipits*. In Proceedings of the IAML-IASA Congress, Oslo, Norway, August 2004

S. Furui, *Digital Speech Processing: Synthesis, and Recognition*, 2nd edn., Signal Processing and Communications (Marcel Denker Inc., New York, 1996)

J. Grudin, in *Human-Computer Interaction Handbook*, 2nd edn., A moving target: The evolution of human-computer interaction, ed. by A. Sears, J .A. Jacko (CRC Press, Boca Raton, 2007), pp. 1–24. ISBN 0-8058-5870-9

O. Lartillot, P. Toiviainen, MIR in Matlab (II): a toolbox for musical feature extraction from audio. In *Proceedings of the ISMIR 2007*, ISMIR. Vienna, Austria (2007)

A.V. Oppenheim, R.W. Schafer, *Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, 1975)

A. Pentland, Socially aware computation and communication. IEEE Comput. **38**(3), 33–40 (2005)

L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)

G. Ruske, *Automatische Spracherkennung. Methoden der Klassifikation und Merkmalsextraktion*, 2nd edn. (Oldenbourg, Germany, 1993)

J. Schenk, G. Rigoll. *Mensch-Maschine-Kommunikation: Grundlagen von sprach- und bild-basierten Benutzerschnittstellen*. Springer, p. 348 (2010). ISBN 978-3-6420-5456-3

E.M. Schmidt , Y.E. Kim, Prediction of time-varying musical mood distributions from audio. In *Proceedings of the ISMIR 2010*, ISMIR. Utrecht, The Netherlands (2010)

M. Schroeder, G. McKeown, Considering social and emotional artificial intelligence. In *Proc.eedings of the AISB 2010 Symposium Towards a Comprehensive Intelligence Test*, SSAISB. Leicester, UK (2010)

B. Schuller, *Automatische Emotionserkennung aus sprachlicher und manueller Interaktion*. Doctoral thesis, Technische Universität München, Munich, Germany, June 2006

B. Schuller, D. Arsić, F. Wallhoff, G. Rigoll, Emotion Recognition in the Noise Applying Large Acoustic Feature Sets. In *Proceedings of the 3rd International Conference on Speech Prosody (SP) 2006*, ISCA. Dresden, Germany, pp. 276–289, May 2006

B. Schuller, F. Eyben, G. Rigoll, Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech, in *Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-based Systems (PIT 2008)*, vol. 5078, Lecture Notes in Computer Science, ed. by E. André (Springer, Kloster Irsee, Germany, 2008), pp. 99–110

B. Schuller, F. Friedmann, F. Eyben, Automatic Recognition of Physiological Parameters in the Human Voice: Heart Rate and Skin Conductance. In *Proceedings of the ICASSP 2013*, IEEE. Vancouver, Canada, pp. 7219–7223, May 2013

B. Schuller, G. Rigoll, M. Grimm, K. Kroschel, T. Moosmayr, G. Ruske, Effects of In-Car Noise-Conditions on the Recognition of Emotion within Speech. In *Proceedings of the 33. Jahrestagung für Akustik (DAGA) 2007*, DEGA. Stuttgart, Germany, pp. 305–306, March 2007

B. Schuller, S. Steidl, A. Batliner, F. Jurcicek, The INTERSPEECH 2009 Emotion Challenge. In *Proceedings of the INTERSPEECH 2009*, Brighton, UK, pp. 312–315, September 2009

B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, G. Rigoll, Cross-corpus acoustic emotion recognition: Variances and strategies. IEEE Trans. Affect. Comput. (TAC) **1**(2), 119–131 (2010). doi:10.1109/T-AFFC.2010.8

B. Schuller, Signals and communication technology, *Intelligent Audio Analysis* (Springer, Berlin, 2013)

B. Schuller, A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing* (Wiley, Hoboken, 2013), p. 344. ISBN 978-1119971368

G. Sinha, R. Shahi, M. Shankar, Human computer interaction. In *Proceedings of the 3rd International Conference on Emerging Trends in Engineering and Technology (ICETET)*, IEEE. pp. 1–4 (2010)

M. ter Maat, D. Heylen, Using context to disambiguate communicative signals, in *Multimodal Signals: Cognitive and Algorithmic Issues*, Lecture Notes in Computer Science, ed. by A. Esposito, A. Hussain, M. Marinaro, R. Martone (Springer, Berlin, 2009), pp. 67–74. doi:10.1007/978-3-642-00525-1_6. ISBN 978-3-642-00524-4

D. Ververidis, C. Kotropoulos, Emotional speech recognition: resources, features, and methods. Speech Commun. **48**(9), 1162–1181 (2006)

A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: Survey of an emerging domain. Image Vis. Comput. **27**(12), 1743–1759 (2009). doi:10.1016/j.imavis.2008.11.007

M. You, C. Chen, J. Bu, J. Liu, J. Tao. Emotion recognition from noisy speech. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2006)*, IEEE. Toronto, Canada, pp. 1653–1656, July 2006. doi:10.1109/ICME.2006.262865

Z. Zeng, M. Pantic, G.I. Rosiman, T.S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)