

# Contents

<b>1</b>	<b>Introduction</b>	1
1.1	Acoustic Analysis of Speech and Music	2
1.2	Deficiencies of the State-of-the-Art.	2
1.3	Aims of This Thesis.	3
1.3.1	Real-time Analysis Framework	3
1.3.2	Baseline Feature Sets	4
1.3.3	Real-World Robustness.	4
1.3.4	Large-Scale Evaluation	5
1.4	Overview	5
	References	6
<b>2</b>	<b>Acoustic Features and Modelling</b>	9
2.1	Basics of Signal Processing.	9
2.1.1	Signal Representation	10
2.1.2	Frequency Domain	11
2.1.3	Short-Time Analysis.	13
2.1.4	Pre-processing	18
2.2	Acoustic Low-Level Descriptors	19
2.2.1	Time Domain Descriptors	20
2.2.2	Energy	21
2.2.3	Spectrum	22
2.2.4	Spectral Descriptors	35
2.2.5	Autocorrelation	44
2.2.6	Cepstrum	46
2.2.7	Linear Prediction	48
2.2.8	Formants.	53
2.2.9	Perceptual Linear Prediction	55
2.2.10	Cepstral Features	60
2.2.11	Pitch.	62
2.2.12	$F_0$ Harmonics	72
2.2.13	Voice Quality	73

2.2.14	Tonal Features . . . . .	78
2.2.15	Non-linear Vocal Tract Model Features. . . . .	80
2.3	Derived Features and Post-processing of Low-Level Descriptors . . . . .	82
2.3.1	Differences . . . . .	83
2.3.2	Delta Regression Coefficients. . . . .	83
2.3.3	Higher Order Delta Regression Coefficients and Differences . . . . .	84
2.3.4	Temporal Smoothing . . . . .	84
2.4	Supra-Segmental Features . . . . .	85
2.4.1	Stacking of Low-Level Descriptors. . . . .	85
2.4.2	Statistical Functionals . . . . .	86
2.4.3	Modulation Functionals. . . . .	103
2.5	Modelling . . . . .	106
2.5.1	Static Modelling with Support Vector Machines. . . . .	107
2.5.2	Dynamic Modelling . . . . .	109
	References . . . . .	115
<b>3</b>	<b>Standard Baseline Feature Sets. . . . .</b>	<b>123</b>
3.1	INTERSPEECH 2009 Emotion Challenge Set . . . . .	124
3.2	INTERSPEECH 2010 Paralinguistics Challenge Set . . . . .	124
3.3	INTERSPEECH 2011 Speaker State Challenge Set . . . . .	126
3.4	INTERSPEECH 2012 Speaker Trait Challenge Set . . . . .	127
3.5	INTERSPEECH 2013 ComParE Set. . . . .	128
3.6	INTERSPEECH 2014 ComParE Set. . . . .	130
3.7	Audio-Visual Emotion Challenge Sets . . . . .	130
3.8	Geneva Minimalistic Acoustic Parameter Set. . . . .	131
3.9	Music Genre Sets. . . . .	133
3.10	Summary . . . . .	133
	References . . . . .	135
<b>4</b>	<b>Real-time Incremental Processing . . . . .</b>	<b>139</b>
4.1	Segmentation Issues . . . . .	140
4.1.1	On-Line Segmentation . . . . .	141
4.1.2	Incremental Segmentation . . . . .	142
4.2	Feature Issues . . . . .	143
4.3	Architecture of the openSMILE Framework . . . . .	144
4.3.1	Incremental Processing . . . . .	147
4.3.2	Smile Messages . . . . .	151
4.4	Fully Continuous Speech Emotion Recognition . . . . .	151
4.4.1	Related Work . . . . .	152
4.4.2	Proposed Continuous Modelling Approach . . . . .	153
4.4.3	Acoustic Features. . . . .	156
	References . . . . .	157

<b>5 Real-Life Robustness</b>	163
5.1 Voice Activity Detection	164
5.1.1 Related VAD Approaches	164
5.1.2 Proposed VAD Based on LSTM-RNNs	166
5.1.3 Benchmarking of the Proposed Approach	167
5.2 Feature Normalisation	171
5.2.1 Normalisation of Low-Level Descriptors	172
5.2.2 Normalisation of Supra-Segmental Features	173
5.2.3 Incremental Normalisation	174
5.3 Noise Robustness	175
5.3.1 Synthesis of Noisy and Reverberated Data	177
5.3.2 Acoustic Feature Analysis and Selection	179
References	180
<b>6 Evaluation</b>	185
6.1 Speech and Music Databases	185
6.1.1 Airplane Behaviour Corpus (ABC)	186
6.1.2 FAU-AIBO Database (AIBO)	187
6.1.3 TUM Audiovisual Interest Corpus (AVIC)	188
6.1.4 Danish Emotional Speech Database (DES)	189
6.1.5 Berlin Emotional Speech Database (EMO-DB)	189
6.1.6 eNTERFACE'05 Database	190
6.1.7 Geneva Multimodal Emotion Portrayals (GEMEP)	190
6.1.8 Belfast Sensitive Artificial Listener Database (SAL)	191
6.1.9 SEMAINE Database	192
6.1.10 Geneva Singing Voice Emotion (GeSiE) Database	198
6.1.11 Speech Under Simulated and Actual Stress (SUSAS)	199
6.1.12 Vera-Am-Mittag (VAM)	200
6.1.13 Ballroom Dance-Style Database (BRD)	200
6.1.14 Genre Discrimination Database (GeDiDB)	201
6.2 Noise Robust Affective Speech Classification	202
6.2.1 Analysis of Acoustic Features	202
6.2.2 Classification Performance	203
6.3 Evaluation of the Baseline Feature Sets	207
6.3.1 Mapping of Emotions	207
6.3.2 Evaluation Method	208
6.3.3 Results	211
6.4 Continuous Dimensional Affect Recognition	225
6.4.1 Experimental Setup	226
6.4.2 Results	227
References	234

<b>7 Discussion and Outlook</b> . . . . .	237
7.1 Summary . . . . .	237
7.2 Achievements . . . . .	239
7.3 Future Work and Concluding Remarks . . . . .	241
References . . . . .	242
<b>Appendix A</b> . . . . .	247
<b>Appendix B: Mel-Frequency Filterbank Parameters.</b> . . . . .	295
<b>Curriculum Vitae—Florian Eyben</b> . . . . .	297

Real-time Speech and Music Classification by Large  
Audio Feature Space Extraction

Eyben, F.

2016, XXXVIII, 298 p. 41 illus., 39 illus. in color.,

Hardcover

ISBN: 978-3-319-27298-6