# Chapter 5
# Real-Life Robustness

With rapidly growing interest in and market value of social signal and media analysis (Zeng et al. 2009; Eyben et al. 2012b, 2013c; Vinciarelli et al. 2009), interactive speech systems (Pittermann et al. 2010; Schröder et al. 2012), and multi-modal user profiling (Schuller et al. 2009b; Lee et al. 2010; Schuller et al. 2012b) as well as stress measurement (Lu et al. 2012), the technologies and algorithms for automatic affect recognition from speech get more and more commercial attention. While good results are reported in research papers in laboratory settings (cf. Schuller et al. 2009a) or with systems tailored towards specific databases, real-life applications still remain challenging (Schuller et al. 2011b; Mower et al. 2011; Schuller et al. 2012a) due to various factors.

These factors can be roughly summarised by three categories (Eyben et al. 2013a): The large variability of affective expression across different speakers, languages, and cultures; contextual dependencies of the meaning and significance of affective expressions; and varying and degraded acoustic conditions caused by reverberation, background noise, and acoustic properties of the recording devices used.

The variability across subjects can only be effectively addressed by analysing data from each of the target groups and deriving rules (such as by Banse and Scherer 1996) or using annotated data from the target groups (in-domain data) for training of data-driven systems. The contextual dependency has to be addressed on a higher level, when the results of a speech/music analyser are interpreted and actions of the system are planned (e.g., as in the SEMAINE system Schröder et al. 2012). In this thesis, the focus is on the robustness due to varying and unpredictable acoustic conditions—in particular additive and convolutive noise, which is an important issue for virtually all imaginable use case scenarios.

Besides robustness to changing acoustic conditions, this chapter discusses two other related robustness issues: on-line discrimination of speech and non-speech segments in a continuous audio stream (Sect. 5.1), and the normalisation of acoustic features with the goal of eliminating speaker and corpus variability (Sect. 5.2). Finally a systematic data synthesis approach to increase robustness of classification algorithms to background noise is presented in Sect. 5.3.





F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, Springer Theses, DOI 10.1007/978-3-319-27299-3_5

## 5.1 Voice Activity Detection

Voice Activity Detection (VAD), also referred to as Speech Activity Detection (SAD), which discriminates speech and non-speech sounds, is an important first step in many speech-based systems. It is important in Automatic Speech Recognition (ASR) applications, for example, to avoid word insertions and false positives due to noise and background speech; it is also used in audio coding to improve compression and save bandwidth (Syed and Wu 2007), for example. For the speech analysis tasks investigated in this thesis, segmentation of a continuous input stream into speech segments incrementally in real-time is crucial (cf. Sect. 4.1). Moreover, it is important that this segmentation is reliable and highly accurate, even in highly noisy environments. In this section, thus, related, existing VAD methods are summarised and a novel data-driven, context aware VAD approach developed for this thesis is described in detail and evaluated on a large data set. This approach has been published and presented at the 2013 IEEE ICASSP conference by the author of this thesis (cf. Eyben et al. 2013b).

### *5.1.1 Related VAD Approaches*

Early approaches to VAD were based on energy thresholds or pitch and Zero-Crossing Rate (ZCR) rules (e.g., Woo et al. 2000; Marzinzik and Kollmeier 2002; Hahn and Park 1992). An energy threshold VAD roughly is based on the following simple algorithm: The Root Mean Square (RMS) or logarithmic energy is computed on a frame level, and a voiced frame is found if the energy crosses a pre-defined speech energy threshold for that given frame. Only if the energy falls below the silence energy threshold (can be equal to or less than the speech energy threshold) on one of the following frames, that frame is labelled as unvoiced. All following frames below the speech energy threshold are also unvoiced until the next voiced frame is detected with an energy above the speech energy threshold.[1] Post-smoothing of the frame-wise decisions can be applied to reach a more robust final VAD decision. Commonly, for most on-line use-cases, a hysteresis based smoothing is implemented as described in Sect. 4.1.[2] Such simple approaches perform well in settings where there is little to no background noise and no non-speech sounds. They fail, however, with high levels of non-speech sounds.

A better discrimination of speech and non-speech sounds can be achieved with glottal and/or spectral features such as Linear Predictive Coding (LPC) coefficients (Rabiner and Sambur 1977) and cepstral coefficients (Haigh and Mason 1993). More recent approaches consider advanced parameters like autoregressive (AR) model parameters (Mousazadeh and Cohen 2011) and Line Spectral Frequencies (LSFs)

[1] In openSMILE the energy based VAD can be implemented with a `cEnergy` component and a `cTurnDetector` component.

[2] This type of smoothing is done by the `cTurnDetector` component in openSMILE.

to actively discriminate between speech-like and non-speech sounds based on static and dynamic statistics. The most promising approaches in strongly corrupted conditions seem to be data-driven methods, where a classifier is trained to predict the classes speech versus non-speech from acoustic features (cf. e.g., Misra 2012). Misra (2012) compares Gaussian Mixture Models (GMMs) with a discriminative classifier and proposes novel features instead of the standard Mel-Frequency Cepstral Coefficient (MFCC)/Perceptual Linear Prediction (PLP) frontends. Realistically noisy, manually labelled YouTube videos are used for evaluation. Many of related data-driven approaches also rely on Gaussian mixture modelling and adaptation (as is typical for ASR systems) to adapt the VAD models to new speakers (e.g., Matsuda et al. 2012) and changing background noise conditions (e.g., Deng et al. 2011; Suh and Kim 2012; Fujimoto et al. 2012). Omar (2012) adapts GMMs to both channel and noise conditions. Thambiratnam et al. (2012) propose to couple the VAD with the acoustic models in the speech recogniser, which is a first step towards inclusion of context in the VAD method. Still, the performance of such approaches goes down when background noise with spectral characteristics similar to speech is audible. Also, the non-stationarity of speech as well as the noise sounds is a big problem (Syed and Wu 2007). Very recent studies suggest that the use of long time-span features clearly improves the robustness in realistic, non-stationary noisy settings because the decision for each frame can be performed in the context of the previous frames: Ng et al. (2012) compare a standard GMM system using 14 PLP cepstral coefficients with a Multi-Layer Perceptron (MLP) based system incorporating long-span acoustic features which are computed over 0.5 s windows. MLP based speech/non-speech posteriors are then decoded over time with two ergodic Hidden Markov Models (HMMs). A similar approach has been proposed by Thomas et al. (2012) who use PLP based and similar, more advanced temporal features in combination with GMMs.

Yet, all of these systems do not use adaptive context learning as provided by Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs) (Sect. 2.5.2.2) in the approach presented below (Sect. 5.1.2). This featured approach is thoroughly benchmarked in Sect. 5.1.3, where a comparison to other state-of-the-art VAD algorithms designed for the use in noisy conditions (cf. Sohn and Kim 1999; Ramirez et al. 2005; Mousazadeh and Cohen 2011) is made (published by the author of this thesis in Eyben et al. 2013b). The reference algorithms all belong to the category of *statistical methods*, where a Likelihood Ratio (LR) test is applied to the hypotheses of speech presence (denoted as $H_1$) and speech absence ($H_0$) on each frame of the observed noisy signal $x_t = s_t + n_t$, where $s_t$ and $n_t$ denote the clean speech signal and the noise signal, respectively. The VAD described by Sohn and Kim (1999) (*SOHN* in the ongoing) is based on a statistical model in the time-frequency domain for the derivation of the LR test. The algorithm introduced by Ramirez et al. (2005) (*RAM*05 in the ongoing), similar to the one in (Ramirez et al. 2004), is based on the concept that more consecutive speech frames concur into the definition of the LR function—thus also taking into account a fixed amount of neighbouring frames, i.e., context information. The algorithm presented by Mousazadeh and Cohen (2011) (*ARG* in the ongoing) is based on the idea of modelling the speech signal by means

of an autoregressive-generalised autoregressive conditional heteroskedasticity (AR-GARCH) model directly in the time domain. Compared to the proposed LSTM-RNN approach this method is computationally very complex because it operates on the sample level. It is thus not very suitable for on-line applications, unless implemented efficiently on a Digital Signal Processor (DSP) platform.

### *5.1.2 Proposed VAD Based on LSTM-RNNs*

For this thesis, a novel data-driven method for voice activity detection based on (unidirectional) LSTM-RNNs (cf. Sect. 2.5.2.2 and Hochreiter and Schmidhuber 1997) has been developed by the author of this thesis and evaluated in noisy scenarios (Eyben et al. 2013b). The motivation behind the use of LSTM-RNNs is their ability to model long range dependencies between the input features and the voice activity labels and dynamically learn those dependencies from the training data. As discussed above, other common data-driven VAD approaches, such as those based on GMMs or MLPs do not consider temporal relations in the model. Delta features, modulation or long-span features (Ng et al. 2012) are used to overcome those limitations. Standard Recurrent Neural Networks (RNNs) (as applied by Gemello et al. 2005 to VAD, for example), are able to model a limited amount of temporal dependency—which, however, vanishes exponentially over time (vanishing gradient problem, cf. Hochreiter et al. 2001). LSTM-RNNs do not have this limitation due to their memory cells with constant error carousels (see Sect. 2.5.2.2 for details).

The networks implemented here have one input layer, which size-wise matches the low-level acoustic feature vectors, one or two recurrent hidden layers with Long Short-Term Memory (LSTM) blocks, and one output layer with a single linear summation unit. The networks are trained to produce a continuous voice activity score in the range $[-1; +1]$ for every input frame. Thereby $+1$ indicates voice/speech and $-1$ indicates silence or non-speech/noise sounds. Two network topologies have been investigated (cf. Eyben et al. 2013b), one shallow and one deep:

- *Net*1: 1 recurrent hidden layer (50 LSTM blocks)
- *Net*2: 3 recurrent hidden layers (50 LSTM blocks; 10 standard sigmoid neurons; 20 LSTM blocks)

On the input side of the networks RelAtive Spectral TrAnsform Perceptual Linear Prediction (RASTA-PLP) (Hermansky 1990) cepstral parameters 1–18 (Sect. 2.2.9.4) are applied together with their first order delta regression coefficients ($W = 2$, cf. Sect. 2.3.2). The size of overlapping audio frames is 25 ms with a frame step (period) of 10 ms. Frames are multiplied with a Hamming window. It is important to highlight that this 36 dimensional feature vector does not contain direct energy information (e.g., the 0th cepstral coefficient or RMS energy or similar). This decision was made on purpose to make the behaviour of the networks invariant to the input level. Features have been extracted with the openSMILE toolkit (Eyben et al. 2010) which was developed by the author of this thesis and $z$-normalisation has been applied to all

features (mean 0, variance 1). The parameters for the $z$-normalisation are computed from the training set only in order to simulate realistic on-line evaluation conditions (for more details on feature normalisation see Sect. 5.2). The LSTM-RNNs were trained and evaluated with the *rnnlib* by Graves et al. (2007). More details on the training parameters are found in (Eyben et al. 2013b).

LSTM-RNNs are perfectly suitable for real-time on-line applications because they (a) provide frame-level decisions (low latency), and (b) the computational complexity for evaluating the networks is asymptotically linear with respect to the number of input frames, i.e., for every frame a constant number of operations must be performed. Many of these operations can be run in parallel, which is ideal for implementation on embedded hardware such as DSPs or Field Programmable Gate Arrays (FPGAs), or on multi-core processors or graphics processors.

### *5.1.3 Benchmarking of the Proposed Approach*

For a thorough evaluation of the proposed VAD approach, a large amount of labelled data in realistic noisy conditions is required. To obtain such data for training and validating the networks, new data were synthesised by building random utterance sequences overlaid with additive and convolutive noise as first published by the author of this thesis in (Eyben et al. 2013b). Clean speech data from the Buckeye (Pitt et al. 2007) and the TIMIT corpus (Garofolo et al. 1993) are used. The Buckeye corpus consists of 26 h of spontaneous speech from 40 subjects (20 male, 20 female) recorded in informal interview settings. Only the subjects' speech is used and segments corresponding to utterances between silence parts of at least 0.5 s length are extracted according to the automatic alignment shipped with the Buckeye corpus. The corpus is split subject-independently into a training, validation, and test partition, stratified by age and gender. The segmentation and subdivision is exactly equal to the one used by Weninger et al. (2011a). TIMIT already provides a split into a training and test partition. The original TIMIT training set is further split speaker-independently into a training and validation set. Speech for the synthesised VAD test set is taken from the original TIMIT and Buckeye test sets. Four types of noises are considered: *babble*, *city*, white and pink *noise*, and instrumental *music*. The babble noise recordings are taken from the *freesound.org* website. Samples from the categories pub-noise, restaurant chatter, and crowd noise are concatenated together. The music recordings resemble instrumental and classical music pieces from the *last.fm* website. The city recordings consist of recordings conducted at the Technische Universität München (TUM) in Munich, Germany on smartphones while volunteers where cycling and walking through the city. White and pink noise samples were generated with pseudo random number generators (white noise) followed by a lowpass filter for the pink noise.

The noise samples used for synthesising the VAD training, validation, and test samples are fully disjoint (i.e., different original pieces of music, different babble samples, etc.). Noise patches for the test and validation partitions span 30 min for

**Table 5.1** Length of the samples available for each noise type for corrupting the training set (Eyben et al. 2013b)

| Noise-type: | Babble | City | Noise | Music |
|---|---|---|---|---|
| Length (h:mm) | 1:34 | 1:56 | 2:00 | 2:56 |

each type, the remaining noise audio is used for the training set. The lengths of these samples vary from 94 min (babble) to 176 min (music). The lengths of these samples detailed per noise type are found in Table 5.1.

Each synthetic utterance in the VAD training set is built by concatenating $N \in \{1, \ldots, 5\}$ original clean speech utterances, which are randomly selected either from TIMIT or Buckeye. A pause (silence) before the first utterance, pauses between all utterances, and a pause after the last utterance are inserted with a randomly (uniform distribution) chosen length between 0.5 and 5 s. Each of the original utterances is level normalised to have a peak amplitude of 0 dB and then the normalised utterances are concatenated to a meta-utterance which is multiplied with a gain factor $g_{s,lin} = 10^{\frac{g_s}{20.0}}$ where $g_s \in [+3\,\text{dB}; -20\,\text{dB}]$ is randomly sampled from a uniform pseudo random number generator. For 80 % of the synthetic utterances, a random noise sample, which matches the length of the meta-utterance ($N$ original utterances and pauses), is selected from the training noise pool and scaled to have a peak amplitude of 0 dB. In order to mix speech and noise at a controlled Signal-to-Noise Ratio (SNR), a gain factor $g_{n,lin}$ as described by Eq. (5.1) is applied multiplicatively to the noise patch:

$$g_{n,lin} = 10^{(log(g_{s,lin}) - \frac{SNR}{20.0})}. \tag{5.1}$$

It is to note that, the SNR is based on peak signal amplitudes and no A-weighting was performed, as the goal was to measure the amount of signal distortion as it would affect linear audio signal processing and not human hearing. The SNR is randomly chosen for each mixed instance from $[-6\,\text{dB}; +25\,\text{dB}]$. The remaining 20 % of all synthetic utterances are not overlaid with noise, i.e., they remain clean speech utterances. 1,948 meta-utterances are created with speech from Buckeye. This corresponds to 15 h of total audio, where 6:43 h are non-speech and 8:17 h are speech. From TIMIT, 3,493 meta-utterances are generated, resembling 19:45 h of total audio, where 12:54 h are non-speech and 6:51 h are speech. In total there are 34:54 h of audio material in the VAD training set used for this study.

The validation set is built in a similar way, however, one single meta-utterance with a total length of 22.5 min is generated from Buckeye and TIMIT speech. The gain of each of the original utterances is varied randomly over the same range as is used for the training set, and silence segments (pauses) are added with random lengths using the same parameters. This same meta-utterance is overlaid with four continuous 22.5 min segments of babble, music, city, and white+pink noise (all normalised to 0 dB peak amplitude). A fixed gain $g_{n,lin}$ is chosen for this noise segment as $g_{n,lin} = 0.5(g^{\mu}_{s,lin} + g^{min}_{s,lin})$, where $g^{\mu}_{s,lin}$ and $g^{min}_{s,lin}$ are the mean and minimum gain factors,

respectively, of all the speech utterances in the meta-utterance. This yields a total 90 min of speech audio for validation from TIMIT and Buckeye. In total, the VAD validation set has 3 h of audio, where 1:22 h are speech and 1:38 h are non-speech. The validation set is used to evaluate the neural network performance after every training epoch and stop the network training when no improvement of the performance on the validation set is obtained for more than 20 epochs. The training is also stopped when a maximum of 100 epochs has been reached.

For the VAD test set, 15 min long meta-utterances are created each from TIMIT and Buckeye speech. Thus, the total length of each test meta-utterance is 30 min. The clean version of the 30 min test audio contains 12 min of speech and 18 min of silence. A single fixed gain of −6 dB for the clean speech is applied, and noise with a peak SNR (noise gain relative to speech gain) of 0 dB is added to the speech.

In order to test the VAD in challenging real-life conditions, a second test set consisting of the full-length English audio tracks of four Hollywood film DVDs is used (Eyben et al. 2013b). The films are chosen from the official development set of the 2012 MediaEval campaign's violence detection task (Demarty et al. 2012). Speech and non-speech segments in the films were manually annotated by a single rater. The list of films and statistics on the lengths of speech/non-speech segments are given in Table 5.2.

Results for the synthetic test and validation sets are given in Table 5.3. Two evaluation metrics are used: the area under Receiver Operating Characteristic (ROC) curves (AUC) and the combined error rate (False Positive Rate (FPR) + False Negative Rate (FNR)). Fixed thresholds which correspond to the thresholds at the Equal Error Rate (EER) on the validation set are used for all test set evaluations in order to ensure fair evaluation conditions. For nets *Net*1 and *Net*2 the selected thresholds are −0.268 and −0.071, respectively. The same thresholds are applied to the DVD film test set. For computation of FPR and FNR, the predictions are binarised by applying the threshold (both for the reference and the LSTM output) and the binarised predictions are smoothed with a silence hysteresis of five frames (i.e., non-speech segments shorter than five frames are joined with the adjacent speech segments).

It can be observed that both the *Net*1 and *Net*2 network topologies outperform all baseline algorithms in terms of Area Under (ROC) Curve (AUC) and combined error rate (FNR + FPR). Notably, this is also the case for clean speech. The largest margin of improvement can be reported for music, babble, white, and pink noise. For city noise, the baselines seem relatively robust, which can be attributed to the fact that the average energy of these noise samples is much lower than the peak amplitude (e.g.,

**Table 5.2** DVD film test set. Film length in [hh:mm] and percentage of parts with speech (sp); minimum, average, and maximum duration of continuous speech segments (Eyben et al. 2013b)

| Film title | Duration (h:mm) | % speech | Min/avg/max (s) |
|---|---|---|---|
| I am legend | 1:36 | 39.2 | 0.5/21.4/174.9 |
| Kill bill volume 1 | 1:46 | 33.9 | 0.4/39.3/321.2 |
| Saving Private Ryan | 2:42 | 48.6 | 0.5/25.2/230.4 |
| The bourne identity | 1:53 | 40.7 | 0.6/32.6/185.6 |

**Table 5.3** Area under (ROC) curve (AUC) frame-level results on the synthetic validation and test sets of LSTM-RNN approaches *Net*1 and *Net*2 and the *RAM*05, *ARG*, and *SOHN* reference algorithms as reported in (Eyben et al. 2013b)

| set | AUC | | | | |
|---|---|---|---|---|---|
| | *Net*1 | *Net*2 | *RAM*05 | *ARG* | *SOHN* |
| Validation | 0.814 | **0.838** | 0.713 | 0.685 | 0.709 |
| Test clean | 0.980 | **0.985** | 0.955 | 0.962 | 0.959 |
| Test babble | 0.909 | **0.932** | 0.877 | 0.875 | 0.826 |
| Test city | 0.968 | **0.972** | 0.928 | 0.935 | 0.931 |
| Test music | 0.921 | **0.940** | 0.725 | 0.675 | 0.677 |
| Test noise | 0.941 | **0.949** | 0.878 | 0.773 | 0.878 |
| Test ALL | 0.951 | **0.961** | 0.821 | 0.794 | 0.805 |

Equal error rate (EER) and combined error rate (false negative rate (FNR) + false positive rate (FPR)) are not given here; the reader is referred to (Eyben et al. 2013b). Test set: −6 dB gain applied to the original speech signal, average SNR is 0 dB

loud cars passing by). ROC curves for the proposed and the baseline algorithms are plotted in Fig. 5.1. The apparent 'smoothness' of the curves for the proposed LSTM-RNN approach compared to the curves of the baseline systems is due to the modelling as a regression task, which delivers a 'continuum' of scores in testing (by varying the binarisation threshold). As far as ROCs are concerned, the behaviour of the two network topologies is almost identical. The EER in the validation and test

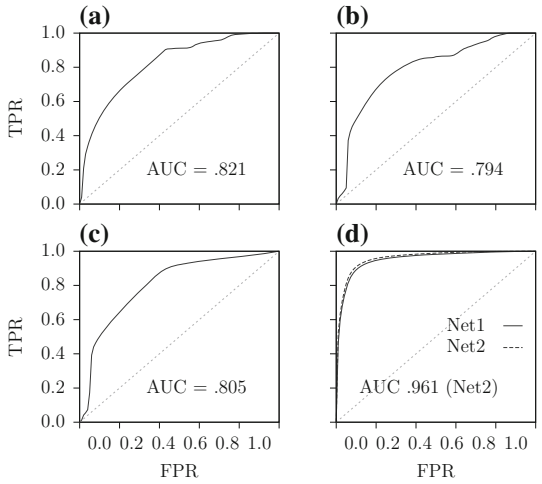


**Fig. 5.1** Receiver operating characteristic (ROC) curves for all VAD algorithms on the synthetic test set: true positive rate (TPR) versus false positive rate (FPR) and area under (ROC) curve (AUC) values for Ramirez's algorithm (Ramirez et al. 2005) (**a**), AR-GARCH (Mousazadeh and Cohen 2011) (**b**), Sohn's approach (Sohn and Kim 1999) (**c**) and the novel LSTM-RNN method (**d**) using network topologies *Net*1 and *Net*2; plots drawn after (Eyben et al. 2013b)

**Table 5.4** Frame-level results for the DVD film test set of nets *Net*1 and *Net*2 and the *SOHN* algorithm

| Film | AUC | | |
|---|---|---|---|
| | *Net*1 | *Net*2 | *SOHN* |
| I am legend | 0.704 | 0.676 | 0.567 |
| Kill bill volume 1 | 0.627 | 0.601 | 0.554 |
| Saving Private Ryan | 0.743 | 0.680 | 0.577 |
| Bourne identity | 0.685 | 0.647 | 0.603 |
| **ALL** | 0.722 | 0.676 | 0.556 |

Area under (ROC) curve (AUC) as reported in (Eyben et al. 2013b); more results given in (Eyben et al. 2013b). Results for *RAM*05 and *ARG* are not included due to their heavy computational load on the large DVD test set, which makes them unsuitable for real-time processing

partitions is around 10 % for both network topologies as opposed to 25 % and above for the baseline algorithms (Eyben et al. 2013b).

The results for the DVD film test set are given in Table 5.4. Compared to the synthetic test set, the performance on the film test set is clearly lower. One main reason for the lower performance on the DVD film set might be that many noise types occur that have not been seen in training, such as gunshots, fighting, or synthetic background ambience, for example. Further, noises which are likely to be confused with speech, such as animal sounds or human like sounds are found. Another possible reason might be the coarse annotation of speech segments; for the sake of efficiency, longer conversations were labelled as continuous speech segments, even though they included small pauses. In the evaluations this results in a higher miss rate than actual. In comparison to the work of Misra (2012) (25.3 % EER on YouTube videos) the EERs of the VAD approach presented in this thesis are very competitive, considering that the system of Misra (2012) was trained on in-domain data, while the approach presented here was trained on synthesised data only.

The LSTM-RNN VAD developed for this thesis outputs frame level voicing predictions. Due to the context modelling of the LSTM-RNN the predictions are already fairly smooth. However, to reliably find continuous segments of speech—which inherently include small pauses between (e.g., at a comma) and within words (e.g., plosives), an additional stage of smoothing suitable for incremental, on-line processing is implemented based on the hysteresis method as described in Sect. 4.1.

## 5.2 Feature Normalisation

Most models, such as Support Vector Machines (SVMs) and Neural Networks (NNs) benefit from a normalised feature space, i.e., when all features are in the same range of values, best between $-1$ and $+1$ or 0 and 1. For NNs the fact is obvious in two ways: (a) the input of each neuron is a weighted sum of the inputs—thus, it is best when all inputs are in a similar range of values, and (b) the gradient of the sigmoid

function is highest around the origin (0)—thus, the backpropagated error will have the most effect on the weights in this case, resulting in a fast converging training and numerically optimal weights in the trained network (Sola and Sevilla 1997). Similarly, all distance based classifiers benefit from well scaled feature dimensions where all values in the same range, in order to obtain numerically well behaved and meaningful distances. For SVMs it was shown by Herbrich and Graepel (2001) that feature space normalisation is an essential pre-processing step needed prior to model training. Further, Graf and Borer (2001) investigate and discuss the effect of various normalisations on SVMs from a theoretical point of view.

Next to being a pre-requisite for training of classifier/regressor models, normalisation also serves the purpose of adaptation to diverse conditions, such as speakers, acoustic conditions, noise settings, etc. Adaptation, however, for general audio analysis is a very wide spread field, which has not yet received enough attention. For on-line processing this issue becomes even more challenging and deserves more attention in future work. In this thesis only a short summary of the challenges of adaptation and normalisation will be presented and the standard approach for (on-line) feature normalisation, that has been employed here, is presented.

Generally, feature normalisation can be applied at two different feature levels: for Low-level Descriptors (LLDs) (Sect. 5.2.1), or for supra-segmental features (Sect. 5.2.2). The following two sections discuss these two possible approaches to feature normalisation, present related work in the area, and finally briefly describe the normalisation method which is applied on the supra-segmental feature level for on-line speech and music analysis in this thesis (Sect. 5.2.3).

### *5.2.1 Normalisation of Low-Level Descriptors*

Most work on acoustic feature normalisation has been performed in the specific field of ASR, where cepstral Mean Variance Normalisation (MVN) applied to MFCC features is a common method (Junqua and Haton 1996; Young et al. 2006): Given a set of values $x(n)$ for $n = 0 \ldots N$, $\mu_x$ and $\sigma_x^2$ are the mean and the variance of these values, respectively. The MVN for every sample $x$ is then expressed as:

$$x' = \frac{1}{\sigma_x^2} (x - \mu_x) \,. \tag{5.2}$$

Thereby the values are normalised to have zero mean and a variance of one. In ASR, typically, MVN is applied to each utterance, i.e., the means and variances are computed individually for each utterance and each feature. This effectively eliminates channel noise and speaker variability in each recording separately. However, especially for short utterances, also phonetic information is discarded. Recently, more advanced feature normalisation approaches have been presented for normalising intra speaker differences in GMM/HMM speech recognition frameworks, e.g., by Sethu et al. (2007).

For paralinguistic tasks most static classification approaches apply MVN to the supra-segmental features (which is discussed in Sect. 5.2.2). While for some functionals, such as mean and variance, there is mathematically no difference in normalising the LLD before applying the functional to compute the supra-segmental feature, or normalising the supra-segmental feature directly, for others, such as linear and quadratic regression coefficients or peak amplitude based functionals, there is an inherent difference. This calls for more work in the future to investigate the difference between the two approaches, and take a deeper look at the role of individual LLDs and functionals. This thesis provides a basis for such analysis by introducing large, comprehensive feature sets. Up to now, only few publications study the effects of LLD normalisation:

For instance, Suzuki et al. (2012) normalise prosodic features to a neutral reference obtained from a speech synthesiser on a frame-level. They show the effectiveness of the approach for two affective dimensions: pleasantness and arousal. However, the method is limited by the fact that the textual content must be known in order to synthesise the neutral reference. This involves costly ASR and text-to-speech algorithms to run in parallel to the speech analyser.

Busso et al. (2009) describe and evaluate a pitch-only based emotion classification model, where emotional speech is compared to a neutral reference. The article itself does not directly describe a normalisation method, but the presented concept of the neutral pitch reference could be applied to normalisation. In that case, for instance, the mean pitch and the variance of the pitch would be computed from all neutral instances and then all instances would be scaled to have zero mean and unit variance. Alternatively this scaling could be done for each speaker separately to effectively eliminate intra-speaker pitch variability.

### 5.2.2 *Normalisation of Supra-Segmental Features*

In most work up to now, including this thesis, a quasi standard for the static classification approaches of normalising the supra-segmental feature vectors has been followed (cf. Schuller et al. 2011b, 2014). Thereby either MVN (also referred to as standardisation or $z$-normalisation, cf. previous section) is applied to the supra-segmental features, or Mean Range Normalisation (MRN), which is defined as follows[3]:

$$x' = \frac{2}{x_{max} - x_{min}} (x - x_{min}) - 1.0, \tag{5.3}$$

where $x_{max}$ and $x_{min}$ are the maximum and minimum values in the set of all $x(n)$ values. The resulting values $x'$ are normalised in a way that $x' \in [-1; +1]$. Compared to MVN this approach is more vulnerable to single outliers, which would significantly impact the range of $x(n)$. Thus, MRN might work well for laboratory

[3] In openSMILE, both MVN and MRN are implemented in the `cVectorMVN` component.

conditions where the data are outlier free, but for realistic conditions, MVN should be preferred.

### 5.2.3 Incremental Normalisation

A crucial question for feature normalisation is on what data set the normalisation parameters (max/min or mean/variance) should be computed. In traditional off-line experiments on a single corpus, typically the corpus is divided into a training and test partition (cf. Schuller and Batliner 2013b). The normalisation parameters are then estimated either from the training partition only and applied to the test partition without adaptation to simulate realistic test conditions where only a single unknown instance from the test partition is presented to the system at each time—e.g., for interactive systems (*on-line* normalisation). This normalisation approach was followed for the baselines of all the INTERSPEECH Challenges. Alternatively, if a batch of, or the whole test set is available at evaluation time, e.g., when doing large-scale, off-line data-mining in voice recordings, then the normalisation parameters for the test data can be computed on and only on the test data directly (*off-line* normalisation). If the test partition is large enough, this theoretically ensures the best possible adaptation (numerically, at least) to the test set conditions.

In order to eliminate inter-speaker differences, the normalisation parameters can be computed individually for data from each speaker. This is referred to as *speaker normalisation*. It, however, assumes that enough test data are available for each speaker to compute meaningful normalisation parameters.

To allow for adaptation to test conditions in an incremental set-up, i.e., where the test data are presented to the system instance by instance, a gradual adaptation to the test set or test speaker (for speaker normalisation) must be performed. This can be implemented, for instance, (a) by exponential adaptation of the parameters, (b) by implementing a fixed buffer of previous instances over which normalisation parameters are computed, or (c) by recomputing the normalisation parameters for every new instance. For method (a), the parameters for MVN, for example, will be updated with the following exponential update rule:

$$\mu_n = \alpha\mu_{n-1} + (1.0 - \alpha)x(n), \tag{5.4}$$

$$\sigma_n^2 = \alpha\sigma_{n-1}^2 + (1.0 - \alpha)(x(n) - \mu_n)^2. \tag{5.5}$$

The influence of the initial parameters decays exponentially over time with this update rule. The amount of decay is controlled by the parameter $\alpha$. Typical values are $\alpha = 0.95$ to $\alpha = 0.999$. For method (c), all data recorded so far have equal influence. The update rules for the $n$th test sample, are ($b$ is an initial weight bias, discussed below):

$$\mu_n = \frac{n-1+b}{n+b}\mu_{n-1} + \frac{1}{n+b}x(n), \tag{5.6}$$

$$\sigma_n^2 = \frac{n-1+b}{n+b}\sigma_{n-1}^2 + \frac{n-1+b}{(n+b)^2}\left(x(n)-\mu_{n-1}\right)^2. \tag{5.7}$$

The above is equivalent to re-computing the mean and variance from all samples collected so far for every new test sample which is arriving, but is computationally much more efficient. In both cases (a) and (c) the initial mean $\mu_{n-1}$ and the initial variance $\sigma_{n-1}^2$ are computed from the training set. For case (c) an initial weight should be chosen for those values by adding a bias $b$ to $n$ in the above update equations. A choice of $b = 100$ results in the initial parameters having a weight equivalent to 100 new test instances.

A pragmatic constraint to consider when doing incremental adaptation of normalisation parameters, is the balance of classes. I.e., if a class-wise balanced training set was used to estimate the initial parameters and train the model, and then in the test case a high number of neutral samples are provided to the system before emotional samples are presented, the normalisation parameters adapt to the neutral samples. That is, the neutral test samples are stretched in the feature space to cover the full range of the normalised feature space. This will result in neutral samples being more and more misclassified as emotional samples, the further the adaptation goes. Thus, during an adaptation phase, the test samples should be balanced regarding the classes. In practical situations this, however, is rarely the case as often many more neutral cases occur. This highlights the importance for future research work to investigate these issues of incremental updates to the normalisation parameters in more detail.

## 5.3 Noise Robustness

As mentioned previously, a core aspect to robustness in speech and music analysis is the robustness of the algorithms against background noise. Background noise is always an issue when a system deals with live recordings in a non-studio environment. Music query services such as Shazam[4] have implemented very robust algorithms (Wang 2003) which ensure robust identification of song titles even from highly noise corrupted snippets, such as recordings of background music in a pub. For emotion analysis, for example, only very little work in this direction has been done, although it has been shown that the classification performance is highly affected by noise (cf. e.g., Eyben et al. 2013a, 2012a; Schuller et al. 2007; You et al. 2006). Although in most cases the methods developed for emotion recognition are transferable to other, related domains, such as speaker state and trait analysis, or more general, Computational Paralinguistics (cf. Schuller and Batliner 2013b), no well known specific work on noise robustness exists in these fields.

[4] http://www.shazam.com/.

In a limited sense the normalisation of acoustic parameters provides noise robustness to some extent. However, the amount of control is low, and normalisation also affects other parameters besides noise, as only a statistical scaling of the feature space is performed (see previous section). Therefore, the author of this thesis distinguishes between the following three explicit approaches to implement noise robustness, which do not include feature normalisation:

1. the *pre-processing approach*, where it is attempted to filter or transform the input audio signal in a way which reduces or removes noise. Many methods and standards exist in the areas of telecommunications and conferencing, as well as ASR. The topic is known as speech enhancement there (Benesty et al. 2005). Popular examples are spectral subtraction, Non-negative Matrix Factorisation (NMF), Wiener Filtering, or simply bandpass filtering (cf. Schuller 2013a). If a multi-channel signal exists, beam forming methods can be used to isolate speech sources spatially (cf. e.g., Parra and Alvino 2001).
2. the *feature approach* involves the design of acoustic features which are tailored very specifically to precisely reflect the target information and which are as little as possible affected by noise (anything but the target information of interest). An example would be the use of spectral peaks as for music identification as suggested by Wang (2003), or implementing an $F_0$ detection algorithm which is noise robust in order to extract noise robust prosodic features (e.g., Talkin 1995 or the Subharmonic Summation (SHS) algorithm used in this thesis—see Sect. 2.2.11 for details).
3. the *modelling approach* is about adapting the classification/regression models to be able to deal with noise. Typically this involves training the model in a way that it can capture the increased variance of noisy data. The best examples are matched conditions training, i.e., where the training data are corrupted with the same type (or a similar type) of noise as the training data, or multi-condition training, where a variety of corruptions are applied to the training set in order to cover a large variety of test conditions. Other methods theoretically would involve splitting the model in one part which models noise and in another part which models the actual speech or music.

Combinations of the three approaches are possible. However, noise removal in the pre-processing steps also discards target information. Especially for affective and paralinguistic speech analysis the common speech enhancement algorithms might remove too much information for highly aroused voices, or pathological voices, due to the assumptions they make on the properties of a normal speech signal. The feature approach, if well designed, does not have these problems. However, a significant amount of engineering work is required to find robust features. Then again, these features are very specific to the setting (databases, tasks, etc.) in which they were evaluated during design and might not work well in other acoustic environments or for other paralinguistic analysis tasks. Both the pre-processing and the feature extraction approaches potentially increase the computational load by a non neglectable amount, which is especially important for real-time, on-line processing. Speech enhancement methods, for example, try to re-construct the clean speech signal, and thus operate

on a sample level which leads to a high number of computing operations to perform. Thus, in this thesis the modelling approach is favoured. No overhead in terms of complex pre-processing and filtering nor complex feature extraction algorithms are required at run-time. Adaptations to the model in order to make it noise robust can all be performed at design/training time. Further, the approach published by the author of this thesis in (Eyben et al. 2013a) is fully data-driven and fully unsupervised. Thus, it contains no task and data specific manual optimisations and can automatically be extended to new domains. In the light of the big data era this seems the most feasible approach for obtaining highly flexible and robust systems which work with real world data across different acoustic environments, speakers, countries, languages, cultures, and tasks.

Previous work in this direction was focussed either on additive noise (e.g., Eyben et al. 2012a) or reverberation in isolation (Schuller 2011a). Techniques from ASR for acoustic pre-processing and signal enhancement, or multi-condition training have typically been applied to boost performances in such noisy conditions (cf. Eyben et al. 2012a; Weninger et al. 2011b). The approach taken here (cf. Eyben et al. 2013a) focuses on finding acoustic features which are least degraded by noise and most correlated with the targets. This is combined with multi-condition training. In order to obtain large amounts of degraded training data for feature analysis and model training, such data are automatically synthesised on a large scale by applying realistic convolutive and additive noise to clean speech recordings.

In the following, the noise robust training method first published by the author of this thesis in (Eyben et al. 2013a) is described. Section 5.3.1 describes how large amounts of new data were synthesised in an automated and unsupervised way to produce degraded training data in four different acoustic conditions. The algorithm to select noise robust acoustic features is described in Sect. 5.3.2. An evaluation of the approach, including a discussion of feature relevance, on two affective speech corpora is given in the next chapter in Sect. 6.2.

### *5.3.1 Synthesis of Noisy and Reverberated Data*

Realistic noise recordings of three types serve as additive noise (cf. Eyben et al. 2013a): Babble noise (*babble*), city street noise (*city*), and instrumental music (*music*). Babble noise recordings are taken from the *freesound.org* website out of the categories pub-noise, restaurant chatter, and crowd noise. Music recordings consist of instrumental and classical music from the *last.fm* website. The city recordings were recorded by the author and his coleagues in Munich, Germany with smartphones while cycling and walking through the city similarly to the procedure described by Schuller et al. (2013c). The noise samples used for synthesising training set samples are fully disjunctive of those used for the test set samples, i.e., no original sample occurs in both sets. The total length of the noise sample pool is 30 min for each noise type in the test set and 94 min for babble, 116 min for city, and 176 min for music in the training set.

To simulate convolutive noise, Room Impulse Responses (RIRs) from the Aachen Impulse Response Database (Jeub et al. 2009) were chosen. To keep the complexity of the evaluations low, and yet simulate realistic conditions, a few meaningful combinations of additive noise types and RIRs were selected:

- babble noise and lecture room (A),
- babble noise and stairway (B),
- city noise and meeting room (C),
- and music noise and chapel (Aula Carolina) (D).

The combinations in that selection range from more favourable reverberation conditions (meeting room) to heavily reverberated (chapel) conditions and at the same time represent a wide range of common non-stationary additive noises. Three different virtual microphone distances (relative) from the virtual sound sources in an azimuth angle of 90° (facing the sound source) in 'near', 'mid', and 'far' distance categories are employed to simulate various intensities of the convolutive noise. The further the virtual microphone is away from the virtual sound source, the larger the amount of convolutive noise (indirect reflections of the room) is in relation to the direct source signal. The actual distance in meters for each category depends on the room type (cf. Table 5.5).

In order to eliminate any influence of the average utterance energy, all clean speech utterances were normalised to a −1 dB peak amplitude. From these normalised clean utterances, speech samples with degraded acoustic conditions were then created. A clean (normalised) utterance is thereby convolved with a RIR, then normalised to −6 dB peak amplitude, and finally overlaid with an additive noise sample, which is scaled in order to achieve a given SNR in the created sample. The test partitions of the clean data are convolved with the 'near', 'mid', and 'far' impulse responses and noise at SNRs from 0 to 12 dB in steps of 3 dB is added resulting in 18 different test sets (including three reverberated sets without noise) for each of the four acoustic conditions (cf. Table 5.5).

**Table 5.5** Four acoustic conditions: additive noise type and room impulse response (RIR) from the Aachen Impulse Response Database along with the source-microphone distances, and signal-to-noise ratio (SNR) ranges for the additive noise (Eyben et al. 2013a)

| | Noise | RIR | Speaker-microphone distance (m) | | | SNR range (dB) |
|---|---|---|---|---|---|---|
| | | | Near | Mid | Far | |
| A | Babble | Lecture room | 2.25 | 5.56 | 10.2 | 0–12 |
| B | Babble | Stairway | 1 | 2 | 3 | 0–12 |
| C | City | Meeting room | 1.45 | 1.9 | 2.8 | 0–12 |
| D | Music | Chapel | 1 | 5 | 20 | 0–12 |

The training partition for each acoustic condition has three times the size of the original training partition because each utterance is included once for the 3 RIR distances. Additive noise at random SNRs (uniformly distributed over the range 0–15 dB and with a 10 % probability of clean utterances) was overlaid. SNRs are calculated after first order high pass filtering from the difference signal of speech and noise to approximate A-weighting and thus to better match human perception. The noise samples are picked at random locations in the training and test noise sample pools, matching the length of the speech samples. These *noise samples* are then convolved with the RIR of the current acoustic condition ('far' distance only) to ensure realistic reverberant conditions also for the noise, and then normalised to −6 dB peak amplitude before additive mixing with the (reverberated) clean speech sample.

### 5.3.2 Acoustic Feature Analysis and Selection

A large set of acoustic features is extracted—the ComParE 2013 set (Sect. 3.5). It contains 6,368 features. From this set, features relevant for the target task are selected by computing the Pearson correlation coefficients (CC) (Pearson 1895) of each feature $x(n)$ over discrete time $n$ with a continuous target label $y(n)$:

$$CC = \frac{\sum_{n=1}^{N} \left(x(n) - x_\mu\right)\left(y(n) - y_\mu\right)}{\sqrt{\sum_{n=1}^{N} \left(x(n) - x_\mu\right)^2} \sqrt{\sum_{n=1}^{N} \left(y(n) - y_\mu\right)^2}}, \tag{5.8}$$

where $x_\mu$ and $y_\mu$ are the arithmetic means of the time series $x(n)$ and $y(n)$ over all timesteps. This method is termed CC-FS in the ongoing. The 400 most relevant features—ranked by CC—for each of the activation, valence, and Level of Interest (LOI) tasks are chosen based on CCs computed from a training set. This set of reduced acoustic features is referred to as the CC-FS feature set.

Feature selection on clean speech data is contrasted with feature selection performed on multi-condition degraded data, in order to identify features which are robust in both clean and the degraded conditions. Results of the feature reduction, as well as classification performances with the reduced set and multi-condition training are shown and discussed in Sect. 6.2.

## References


J. Benesty, S. Makino, J. Chen (eds.), *Speech Enhancement* (Springer, Berlin, 2005). ISBN 978-3-540-24039-6

R. Banse, K.R. Scherer, Acoustic profiles in vocal emotion expression. J. Personal. Soc. Psychol. **70**(3), 614–636 (1996)

C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Audio Speech Lang. Process. **17**, 582–596 (2009)

C.H. Demarty, C. Penet, G. Gravier, M. Soleymani, The MediaEval 2012 affect task: violent scenes detection in hollywood movies, in *Proceedings of the MediaEval 2012 Workshop*. CEUR-WS.org, Pisa, Italy, Oct 2012

S. Deng, J. Han, T. Zheng, G. Zheng, A modified MAP criterion based on hidden Markov model for voice activity detecion, in *Proceedings of ICASSP 2011* (IEEE, Prague, 2011), pp. 5220–5223

F. Eyben, M. Wöllmer, B. Schuller, openSMILE—the munich versatile and fast open-source audio feature extractor, in *Proceedings of ACM Multimedia 2010* (ACM, Florence, 2010), pp. 1459–1462

F. Eyben, B. Schuller, G. Rigoll, Improving generalisation and robustness of acoustic affect recognition, in *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI) 2012*, ed. by L.-P. Morency, D. Bohus, H.K. Aghajan, J. Cassell, A. Nijholt, J. Epps (ACM, Santa Monica, 2012a), pp. 517–522

F. Eyben, F. Weninger, N. Lehment, G. Rigoll, B. Schuller, Violent scenes detection with large, brute-forced acoustic and visual feature sets, in *Proceedings of the MediaEval 2012 Workshop*. CEUR-WS.org, Pisa, Italy, Oct 2012b

F. Eyben, F. Weninger, B. Schuller, Affect recognition in real-life acoustic conditions—a new perspective on feature selection, in *Proceedings of INTERSPEECH 2013* (ISCA, Lyon, 2013a), pp. 2044–2048

F. Eyben, F. Weninger, S. Squartini, B. Schuller, Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies, in *Proceedings of ICASSP 2013* (IEEE, Vancouver, 2013b), pp. 483–487

F. Eyben, F. Weninger, N. Lehment, B. Schuller, G. Rigoll, Affective video retrieval: violence detection in hollywood movies by large-scale segmental feature extraction. PLoS ONE **8**(12), e78506 (2013c). doi:10.1371/journal.pone.0078506

M. Fujimoto, S. Watanabe, T. Nakatani, Frame-wise model re-estimation method based on gaussian pruning with weight normalization for noise robust voice activity detection. Speech Commun. **54**(2), 229–244 (2012). doi:10.1016/j.specom.2011.08.005. ISSN: 0167-6393

J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgrena, V. Zue, TIMIT acoustic-phonetic continuous speech corpus (1993)

R. Gemello, F. Mana, R.D. Mori, Non-linear esimation of voice activity to improve automatic recognition of noisy speech, in *Proceedings of INTERSPEECH 2005* (ISCA, Lisbon, 2005), pp. 2617–2620

A.B. Graf, S. Borer, Normalization in support vector machines, in *Pattern Recognition*, Lecture Notes in Computer Science (Springer, Berlin, 2001), pp. 277–282

A. Graves, S. Fernández, J. Schmidhuber, Multidimensional recurrent neural networks, in *Proceedings of the 2007 International Conference on Artificial Neural Networks (ICANN)*. Lecture Notes in Computer Science, vol. 4668 (Springer, Porto, 2007), pp. 549–558

M. Hahn, C.K. Park, An improved speech detection algorithm for isolated korean utterances, in *Proceedings of ICASSP 1992* (IEEE, San Francisco, 1992), vol. 1, pp. 525–528

J.A. Haigh, J.S. Mason, Robust voice activity detection using cepstral features, in *Proceedings of the IEEE Region 10 Conference on Computer, Communication, Control, and Power Engineering* (IEEE, 1993), vol. 3, pp. 321–324

R. Herbrich, T. Graepel, A PAC-Bayesian margin bound for linear classifiers: why SVMs work, in *Advances in Neural Information Processing Systems* (MIT press, Cambridge, 2001), pp. 224–230

H. Hermansky, Perceptual linear predictive (PLP) analysis for speech. J. Acoust. Soc. Am. (JASA) **87**, 1738–1752 (1990)
S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber, Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, in *A Field Guide to Dynamical Recurrent Neural Networks*, ed. by S.C. Kremer, J.F. Kolen (IEEE Press, New York, 2001)
M. Jeub, M. Schäfer, P. Vary, A binaural room impulse response database for the evaluation of dereverberation algorithms, in *Proceedings of the International Conference on Digital Signal Processing (DSP)* (IEEE, Santorini, 2009), pp. 1–4
J.-C. Junqua, J.-P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications* (Kluwer Academic Publishers, Boston, 1996)
C.-C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P.G. Georgiou, S.S. Narayanan, Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples, in *Proceedings of INTERSPEECH 2010* (ISCA, Makuhari, 2010), pp. 793–796
H. Lu, M. Rabbi, G. Chittaranjan, D. Frauendorfer, M. Schmid Mast, A.T. Campbell, D. Gatica-Perez, T. Choudhury, Stresssense: detecting stress in unconstrained acoustic environments using smartphones, in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (Ubicomp'12)* (ACM, Pittsburgh, 2012), pp. 351–360
M. Marzinzik, B. Kollmeier, Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Trans. Speech Audio Process. **10**, 109–118 (2002)
S. Matsuda, N. Ito, K. Tsujino, H. Kashioka, S. Sagayama, Speaker-dependent voice activity detection robust to background speech noise, in *Proceedings of INTERSPEECH 2012* (ISCA, Portland, 2012)
A. Misra, Speech/nonspeech segmentation in web videos, in *Proceedings of INTERSPEECH 2012* (ISCA, Portland, 2012)
S. Mousazadeh, I. Cohen, AR-GARCH in presence of noise: parameter estimation and its application to voice activity detection. IEEE Trans. Audio Speech Lang. Process. **19**(4), 916–926 (2011)
E. Mower, M.J. Mataric, S.S. Narayanan, A framework for automatic human emotion classification using emotional profiles. IEEE Trans. Audio Speech Lang. Process. **19**(5), 1057–1070 (2011). doi:10.1109/TASL.2010.2076804
T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesel, P. Matjka, Developing a speech activity detection system for the darpa rats program, in *Proceedings of INTERSPEECH 2012* (ISCA, Portland, 2012)
M.K. Omar, Speech activity detection for noisy data using adaptation techniques, in *Proceedings of INTERSPEECH 2012* (ISCA, Portland, 2012)
L. Parra, C. Alvino, Geometric source separation: merging convolutive source separation with geometric beamforming, in *Proceedings of the 2001 IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing XI* (IEEE, 2001), pp. 273–282. doi:10.1109/NNSP.2001.943132
K. Pearson, Notes on regression and inheritance in the case of two parents. Proc. R. Soc. Lond. **58**, 240–242 (1895)
M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, E. Fosler-Lussier, in *Buckeye Corpus of Conversational Speech (2nd release)*. Department of Psychology, Ohio State University (Distributor), Columbus, OH, USA (2007). http://www.buckeyecorpus.osu.edu/
J. Pittermann, A. Pittermann, W. Minker, Emotion recognition and adaptation in spoken dialogue systems. Int. J. Speech Technol. **13**, 49–60 (2010)
L.R. Rabiner, M.R. Sambur, Voice-unvoiced-silence detection using the itakura LPC distance measure, in *Proceedings of ICASSP 1977* (IEEE, Hartford, 1977), vol. 2, pp. 323–326
J. Ramirez, J. Segura, M. Benitez, A. De La Torre, A. Rubio, Efficient voice activity detection algorithms using long-term speech information. Speech Commun. **42**(3), 271–287 (2004)
J. Ramirez, J. Segura, C. Benitez, L. Garcia, A. Rubio, Statistical voice activity detection using a multiple observation likelihood ratio test. IEEE Signal Process. Lett. **12**(10), 689–692 (2005)

M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, M. Wöllmer, Building autonomous sensitive artificial listeners. IEEE Trans. Affect. Comput. **3**(2), 165–183 (2012)

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, 2007), pp. 305–306

B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, Acoustic emotion recognition: a benchmark comparison of performances, in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2009* (IEEE, Merano, 2009a), pp. 552–557

B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, H. Konosu, Being bored? recognising natural interest by extensive audiovisual integration for real-life application. Image Vis. Comput. **27**(12), 1760–1774, Special issue on visual and multimodal analysis of human spontaneous behavior (2009b)

B. Schuller, Affective speaker state analysis in the presence of reverberation. Int. J. Speech Technol. **14**(2), 77–87 (2011a)

B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. **53**(9/10), 1062–1087, Special issue on sensing emotion and affect—facing realism in speech processing (2011b)

B. Schuller, M. Valstar, R. Cowie, M. Pantic, AVEC 2012: the continuous audio/visual emotion challenge—an introduction, in *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI) 2012*, ed. by L.-P. Morency, D. Bohus, H.K. Aghajan, J. Cassell, A. Nijholt, J. Epps (ACM, Santa Monica, 2012a), pp. 361–362

B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet, G. Mohammadi, B. Weiss, The INTERSPEECH 2012 speaker trait challenge, in *Proceedings of INTERSPEECH 2012* (ISCA, Portland, 2012b)

B. Schuller, *Intelligent Audio Analysis*, Signals and Communication Technology (Springer, Berlin, 2013a). ISBN 978-3642368059

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

B. Schuller, F. Pokorny, S. Ladstätter, M. Fellner, F. Graf, L. Paletta, Acoustic geo-sensing: recognising cyclists' route, route direction, and route progress from cell-phone audio, in *Proceedings of ICASSP 2013* (IEEE, Vancouver, 2013c), pp. 453–457

B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, F. Eyben, Medium-term speaker states—a review on intoxication, sleepiness and the first challenge. Comput. Speech Lang. **28**(2), 346–374, Special issue on broadening the view on speaker analysis (2014)

V. Sethu, E. Ambikairajah, J. Epps, Speaker normalisation for speech-based emotion detection, in *Proceedings of the 15th International Conference on Digital Signal Processing (DSP 2007)*, pp. 611–614, Cardiff, UK, July 2007

J. Sohn, N. Kim, A statistical model-based voice activity detection. IEEE Signal Process. Lett. **6**(1), 1–3 (1999)

J. Sola, J. Sevilla, Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Trans. Nucl. Sci. **44**(3), 1464–1468 (1997). doi:10.1109/23.589532. ISSN: 0018-9499

Y. Suh, H. Kim, Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection. IEEE Signal Process. Lett. **19**(8), 507–510 (2012). doi:10.1109/LSP.2012.2204978. ISSN: 1070-9908

M. Suzuki, S. Nakagawa, K. Kita, Prosodic feature normalization for emotion recognition by using synthesized speech, in *Advances in Knowledge-Based and Intelligent Information and Engineering Systems—16th Annual KES Conference*, vol. 243, Frontiers in Artificial Intelligence and Applications, ed. by M. Graña, C. Toro, J. Posada, R.J. Howlett, L.C. Jain (IOS Press, San Sebastian, 2012), pp. 306–313

W.Q. Syed, H.-C. Wu, Speech waveform compression using robust adaptive voice activity detection for nonstationary noise in multimedia communications, in *Proceedings of Global Telecommunications Conference, 2007 (GLOBECOM'07)* (IEEE, Washington DC, 2007), pp. 3096–3101

D. Talkin, A robust algorithm for pitch tracking (RAPT), in *Speech Coding and Synthesis*, ed. by W.B. Kleijn, K.K. Paliwal (Elsevier, New York, 1995), pp. 495–518. ISBN 0444821694

K. Thambiratnam, W. Zhu, F. Seide, Voice activity detection using speech recognizer feedback, in *Proceedings of INTERSPEECH 2012* (ISCA, Portland, 2012)

S. Thomas, S.H. Mallidi, T. Janu, H. Hermansky, N. Mesgarani, X. Zhou, S. Shamma, T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, Acoustic and data-driven features for robust speech activity detection, in *Proceedings of INTERSPEECH 2012* (ISCA, Portland, 2012)

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

A.L. Wang, An industrial-strength audio search algorithm, in *Proceedings of ISMIR* (Baltimore, 2003)

F. Weninger, B. Schuller, M. Wöllmer, G. Rigoll, Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and long short-term memory, in *Proceedings of ICASSP 2011* (IEEE, Prague, 2011a), pp. 5840–5843

F. Weninger, B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognition of non-prototypical emotions in reverberated and noisy speech by non-negative matrix factorization. EURASIP J. Adv. Signal Process. (Article ID 838790), Special issue on emotion and mental state recognition from speech (2011b)

K. Woo, T. Yang, K. Park, C. Lee, Robust voice activity detection algorithm for estimating noise spectrum. Electron. Lett. **36**(2), 180–181 (2000)

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, 2006), pp. 1653–1656. doi:10.1109/ICME.2006.262865

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, *The HTK Book*, Cambridge University Engineering Department, for HTK version 3.4 edition (2006)

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)