Affective speaker state analysis in the presence of reverberation

Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn. 2011. "Affective speaker state analysis in the presence of reverberation." *International Journal of Speech Technology* 14 (2): 77–87. https://doi.org/10.1007/s10772-011-9090-8.

Affective speaker state analysis in the presence of reverberation

Björn Schuller

Abstract Little is known about the impact of reverberation on performance of running speaker state classification systems. This study thus aims to approach the topic by measuring effects on a state-of-the-art engine with consideration of six public room impulse responses for convolution of the affective speech signals of three standard datasets comprising of emotion and interest. Speech data thereby is given by this year's INTERSPEECH Paralinguistic Challenge corpus TUM AVIC and the frequently used Berlin and eNTERFACE sets. The room impulse responses comprise rooms in private apartments, chapels, a factory hall, and a van. Speaker independent performance after speaker adaptation is investigated. To cope with reverberation, matched condition learning and acoustic space adaptation are considered as efficient means. By that a report is provided on suitability of feature types given the type of impulse response. In the result almost all occurring corruption arising from reverberation can be restored, yet the general impact varies with the type of room or acoustic environment.

Keywords Speaker classification · Affective computing · Reverberation · Model adaptation

1 Introduction

Assessing a speaker's affective state is not only by and large considered useful in a multiplicity of human-machine and -robot communication and media retrieval scenarios and applications (Zeng et al. 2009; Schuller et al. 2010b), it may also be a crucial knowledge factor in other speech technology: Affective speech is known to influence recognition of, e.g. speech recognition (Athanaselis et al. 2005; Steidl et al. 2010) and keyword spotting (Wöllmer et al. 2009), speaker recognition (Raja and Dandapat 2010), or general spoken dialogue (Schröder et al. 2008; Pittermann et al. 2010).

The technology to classify affective speaker states has generally matured to a degree, where—comparable to the development in the related field of automatic speech recognition—more real-world problems can be faced after going from pre-selected prototypical acted speech recorded in the lab to non-filtered natural affect in spontaneous speech recorded in increasingly realistic conditions and acoustic environments (Schuller et al. 2010c).

In such general field conditions, capturing of a speech signal $s[n]$ by a microphone m will then result in the capture signal

$$x_m(t) = h_m(t) * s(t) + \nu_m(t) \tag{1}$$

where h_m is the impulse response of the acoustic channel from the source to microphone and $\nu_m[n]$ is observation noise. Computationally assessing affective speaker states in disturbed signal conditions has so far been mostly investigated for speech with additively superposed noise $\nu_m[n]$ (Schuller et al. 2006a; Grimm et al. 2007; You et al. 2006, 2007; Tawari and Trivedi 2010) and for phone transmission by Yoon et al. (2007) or to design features more 'robust' (Kim et al. 2005; Lee et al. 2006; Lugger et al. 2006) considering diverse influence factors.

However, practically no experience with systematically varied real world room responses $h_m(n)$ exists when it

B. Schuller (✉)
Institute for Human-Machine Communication, Technische Universität München, Theresienstrasse 90, 80333 Munich, Germany
e-mail: schuller@tum.de

comes to the recognition of emotion in speech (Schuller et al. 2007). This is regrettable, as it is well known that reverberation effects, e.g., intelligibility of speech (Payton et al. 1994) and is thus highly likely to impact 'intelligibility' of affect. In a running system used 'in the wild' this will need to be dealt with as room impulse responses will usually be differing. In on-line systems this becomes in particular important when the distance from the speaker's mouth to the microphone is larger than close-talk distance, as given in typical hands-free applications. Most challengingly, the room conditions will often vary dynamically, once speech of moving or turning subjects is dealt with. As it is further known that non-matched learning and classification conditions may heavily downgrade performance (Schuller et al. 2010c), the aim is to provide insight on likely occurring effects when testing affective speaker state classification in non-matched reverberation condition per se and per feature type. Considering the large existing body of literature on coping with and de-reverberation of speech—for an excellent overview the reader is referred to Naylor and Gaubitch (2010)—and effects of reverberation in the discipline of automatic speech recognition, adapting the learnt acoustic model and feature space to the present room acoustics as 'first simple' counter measures is additionally investigated.

To approach the topic gently, only corruption of speech in a 'static' manner is considered, i.e., by convolution of affective speech of a complete test partition with one impulse response at a time. To this end six publicly available responses were selected that cover a reasonable application related variety of diverse indoor and one automotive condition. To further base findings not exclusively on one affective speech corpus, it was decided for three well suited such. These were first selected by choosing a good variety of natural, elicited, and acted affective speaker states and preferring such recorded in studio noise conditions to exclude double effects as much as possible.

Acoustic analysis bases on a variety of typical Low-Level-Descriptors (LLD) to which functionals are applied to obtain a total of 1.4 k acoustic features. Once optimization of this large space by using Correlation-based Feature Selection (CFS), and once type-wise evaluation to analyze impact of reverberation on features in a well-interpretable manner are investigated. The classification is based on the frequently used Support Vector Machines (SVM) in a speaker independent Leave-one-speaker-out (LOSO) manner. In LOSO, all speakers but one are used for training—the one left out is then used for testing. This is repeated until all speakers have once been used for testing and means over speakers are reported. This method is very popular in the field, as data for training and testing is usually sparse, and one wishes to use as much data for training as possible. At the same time, significance of findings benefits from reporting on all speakers. By LOSO one can reach this aim but ensure strict speaker independence at the same time.

In the remainder of this article, first artificial speech reverberation will be introduced in Sect. 2, the room impulse responses decided for in Sect. 3 followed by the emotional speech databases in Sect. 4, and the constructed acoustic feature space in Sect. 5. The experimental protocol for convolution and subject independent evaluation looking at impact on diverse feature types and efficiency of matched condition learning known to improve recognition of reverberated speech (Haderlein et al. 2005) and acoustic space adaptation is presented in Sect. 6 before concluding in Sect. 7.

2 Reverberation

As shown in (1), the room impulse response (RIR)—usually several thousand taps long—affects the speech signal by convolution in the time domain leading to reverberation of the speech signal. As opposed to a single echo, reverberation is characterized by being composed of a large number of single echoes caused by reflections at walls or other items and decaying over time owed to the sound being absorbed. In the frequency domain, this naturally corresponds to the multiplication of the respective speech signal's and RIR's transforms. Being an impulse response, the RIR is often measured by provision of an acoustic impulse such as the blast of an air balloon or a gun shot and recording the room's according response. This response $h_m(n)$ naturally highly depends on the position and, if a directional microphone is used, direction of the capture microphone m, i.e., facing different directions already highly impacts the response. As a consequence already turning of the head may result in speech corrupted by potentially highly differing RIR depending, e.g., on the architecture and furniture of a room.

In principle, one can attempt to measure or identify the room impulse response also in the target environment of an application, e.g., most simply by clapping the hands in a speech pause. This is in praxis likely limited to few application scenarios as for static response characteristics. However, as outlined, already with changing the direction of the microphone severe differences may occur. In a mobile application the response would need to be constantly updated, which is, e.g., in a moving situation, almost infeasible. In addition, even if having identified the acoustic channel properties, direct inversion is not necessarily feasible as the RIR can be very long, usually has non-minimum phase and may contain spectral nulls resulting in strong peaks in the spectrum causing narrow band noise amplification after inversion (Neely and Allen 1979).

Thus, investigation of effects of reverberation impact seems mandatory and is usually observed by systematic artificial reverberation of speech data (Haderlein et al. 2005). For the oncoming experiments this is carried out by convolution of the RIR with the speech from diverse databases. As clipping may occur in artificial reverberation, this is strictly avoided by according normalizing scaling.

3 Six room impulse responses

It was decided for the following six room impulse responses (RIR) exclusively recorded in real spaces which are free for research usage, well documented, and publicly available[1] to cover for a good variety of typical indoor environments and reaching from small to large reverberation time. They are introduced by their short identifier used in the ongoing for better readability and their original name to foster easy reproducibility of experiments:

- *Factory Hall* ("Factory Hall") was recorded in a huge factory hall in Amsterdam/The Netherlands. The recorder captured this RIR with two Schoeps cmc5 MK5 microphones (omni) wide spaced A–B, a Ren Heijnis mic preamp, a Fostex PD4 DAT recorder, and a 6 mm caliber starter gun to ensure an optimal S/N ratio.
- *Van* ("Mercedes-van") was recorded in the interior of an empty small van—a Mercedes one. The recorder captured it with a Genelec S30 speaker faced backwards and the microphone facing front in the back with the otherwise same equipment as for the previously introduced *Factory Hall* RIR.
- *Living Room* ("Amsterdam Living Room") was recorded in a living room—also in Amsterdam/The Netherlands—which is a tight space with wooden floor, fully furnished. It was captured with equivalent equipment as the above introduced *Van* RIR.
- *Hallway* ("@carolas-Livingroom-facing Hallway") was taken from the RIR series "@Carolas's apartment"—an empty apartment due to moving in Hamburg/Germany captured using OKM in-ear-microphones at an ear height of about 1.79 m from ground with a Sony Net-MD NZ-1 by balloon blasting in 3–5 m distance from microphones. This response was recorded facing the narrow hallway of the apartment from the living room side.
- *Bathroom* ("@carolas-Bathroom") is the second room impulse response from this set recorded in the small bathroom tiled to the ceiling and having a bath sink, bath tub and toilet inside. Recording is equivalent to the *Hallway* RIR.
- *Chapel 50ft* ("Chapel-NOS-Rear-Facing-in-50ft") was recorded in a cavernous church under construction. This church now seats 1 500 and is at least 40 feet tall. At the time of recording it was just a shell with hardly anything in it including doors and windows, and no acoustic treatment. The micropohne pair used resembles a Blumlein rear facing NOS pair at 50 feet from the impulse source.

All chosen RIR were either recorded in 44.1 kHz or 48.0 kHz. They have thus been down-sampled for processing with the affective speech databases.

All six impulse responses are depicted in Fig. 1 in the time (non-normalized) and the frequency domain. As can be seen, they differ considerably in overall length and time to decay by 60 dB (T_{60}). Acoustic characteristics of these are further provided in detail in Table 1 and additionally in Table 2, for third octace bands in the range of 125 to 4 000 Hz.

The characteristics of the RIR $h_m(t)$ are provided following ISO 3382, where the Reverberation Times T are determined from the decay curve and Clarity C, Definition D, and Center Time T_s are (logarithmic) ratios between a fraction and the entire or remaining RIR energy (Campanini and Farina 2009):

- Early-to-late arriving sound energy ratio *Clarity* C_{t_e}, where

$$C_{50} = 10 \cdot \lg \frac{\int_0^{t_e} h_m^2(t)dt}{\int_{t_e}^{\infty} h_m^2(t)dt}, \quad t_e = 50 \text{ ms}. \tag{2}$$

In the digital domain with the sampling frequency f_s, the discrete RIR $h_m[n]$ is accordingly described by:

$$C_{50} = 10 \cdot \lg \frac{\sum_0^{n_e} h_m^2[n]}{\sum_{n_e}^{\infty} h_m^2[n]}, \quad n_e = f_s \cdot 50 \text{ ms}. \tag{3}$$

- Early-to-total sound energy ratio *Definition* D_{t_e}, where

$$D_{50} = \frac{\int_0^{t_e} h_m^2(t)dt}{\int_0^{\infty} h_m^2(t)dt}, \quad t_e = 50 \text{ ms}. \tag{4}$$

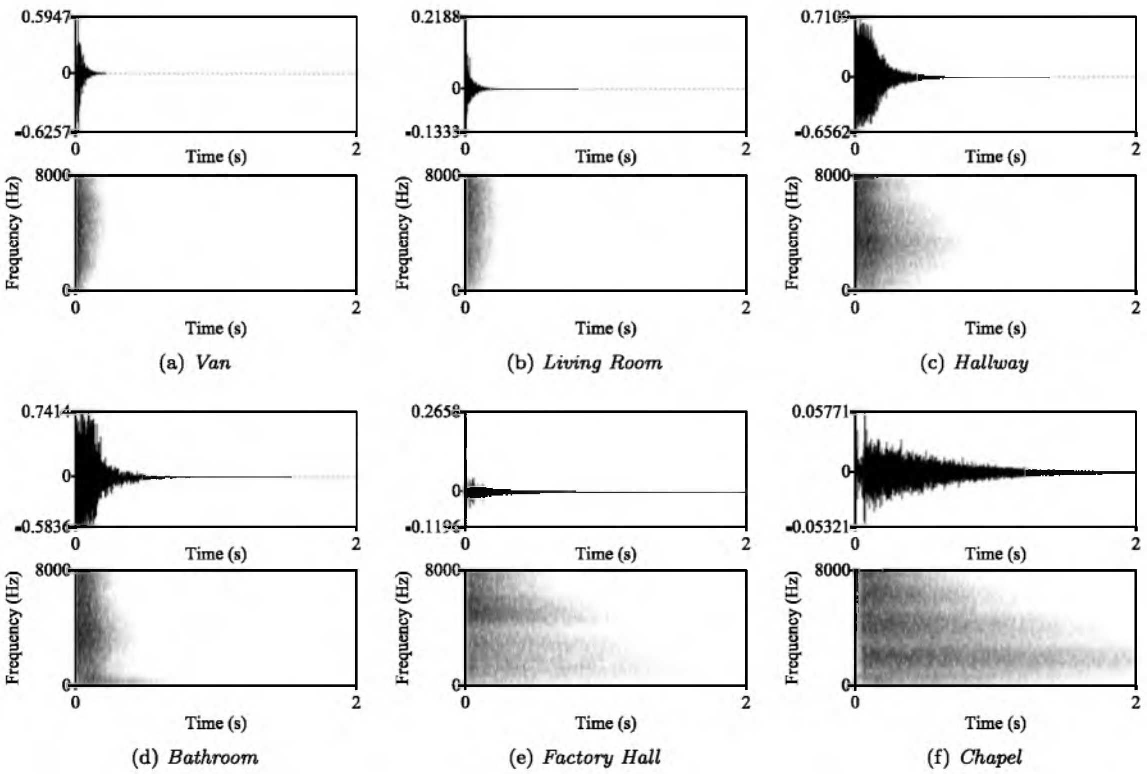
Accordingly, in the digital domain we have:

$$D_{50} = 10 \cdot \lg \frac{\sum_0^{n_e} h_m^2[n]}{\sum_0^{\infty} h_m^2[n]}, \quad n_e = f_s \cdot 50 \text{ ms}. \tag{5}$$

As for the Clarity C, the Definition D also expresses a balance between early and late arriving energy. The idea is to measure clarity and definition as perceived by a human. While Definition is less frequently used, it respects not only the late, but the total signal energy.

Clarity and Definition aim to provide a measure how easily individual sounds can be distinguished from within a general audible stream. For Clarity, this degree highly depends on the type of sound, here speech. By that these measures provide an objective measure of the amount of 'blend': Having two similar sounds arrive at the ear in close temporal proximity (50–80 ms), these are likely to be integrated by the human ear as one sound. By that any reflected sound energy arriving within this period of integration effectively increases the perceived intensity of the direct sound.

[1] http://noisevault.com/



(a) *Van*

(b) *Living Room*

(c) *Hallway*

(d) *Bathroom*

(e) *Factory Hall*

(f) *Chapel*

Fig. 1 Room impulse responses after mono 16 kHz conversion and according spectrograms

Table 1 Per room impulse response: A-weighted C_{50}, D_{50}, T_s, T_{10}, T_{20}, and T_{30}

Characteristics	C_{50} [dB]	D_{50} [%]	T_s [ms]	T_{10} [s]	T_{20} [s]	T_{30} [s]
Van	12.02	94.09	19.15	0.21	0.26	0.26
Living Room	16.25	97.68	5.26	0.22	0.31	0.32
Hallway	−1.87	39.40	76.48	0.89	0.94	1.03
Bathroom	−0.86	45.05	66.12	0.84	0.63	0.85
Factory Hall	9.34	89.58	21.05	1.45	2.37	3.27
Chapel	2.81	65.62	115.13	4.08	3.80	3.88

Note that the chosen time $t_e = 50$ ms resembles the typical value when dealing with speech, as opposed to $t_e = 80$ ms generally preferred for music.

- *Center Time* T_s, where

$$T_s = \frac{\int_0^{\infty} t \cdot h_m^2(t)dt}{\int_0^{\infty} h_m^2(t)dt} \tag{6}$$

or, respectively,

$$T_s = \frac{\sum_0^{\infty} n \cdot h_m^2[n]}{\sum_0^{\infty} h_m^2[n]}, \quad n = f_s \cdot t. \tag{7}$$

It is often given as an alternative to C and D as it avoids discrete division of the RIR into early and late periods. This time of the center of gravity of the squared impulse response is measured in milliseconds. A high Center Time value is an indicator of poor Clarity.

- T_{10} or *Early Decay Time* (EDT) measured over the first 10 dB of the decay, i.e., evaluated between $[0, \ldots, -10]$ dB. It gives a more subjective evaluation of the reverberation time: The initial portion of the sound decay curve process is responsible for subjective impression of reverberation—the later is typically masked by new sounds.
- T_{20} is the reverberation time evaluated between $[-5, \ldots, -25]$ dB of the decay.
- T_{30} is accordingly evaluated between $[-5, \ldots, -35]$ dB.

The tables show the variety of the RIR not only in terms of higher reverberation times for the wider rooms as one would expect, but also clear frequency dependence of these times in the range most important for speech.

Table 2 Per room impulse response: T_{10} and T_{30} in third octave bands

Frequency [Hz]	125	250	500	1 k	2 k	4 k
T_{10} [s]						
Van	0.23	0.11	0.14	0.14	0.21	0.24
Living Room	0.22	0.23	0.25	0.24	0.28	0.17
Hallway	1.02	1.08	0.96	0.87	0.92	0.91
Bathroom	1.11	1.05	0.83	0.83	0.72	0.72
Factory Hall	2.06	1.99	0.75	1.94	1.54	0.02
Chapel	3.77	4.18	3.94	4.64	4.14	3.70
T_{30} [s]						
Van	0.29	0.20	0.19	0.21	0.24	0.26
Living Room	0.47	0.32	0.31	0.33	0.32	0.33
Hallway	0.96	1.21	0.99	1.06	1.07	1.02
Bathroom	1.42	1.18	0.86	0.63	0.50	0.42
Factory Hall	18.19	8.54	4.56	4.19	3.15	1.97
Chapel	5.98	3.87	3.48	4.25	3.94	2.90

4 Three affective speech databases

For the oncoming experiments it was next decided for three ‘nuances’ of affective speaker state: fully spontaneous, elicited, and finally fully acted. The three according sets will be introduced in this order, and are further chosen by their high popularity and high number of existing results reported in the literature for comparisons. Number of instances per affective speaker state are provided in brackets after the type of affective speaker state.

4.1 TUM AVIC

In order to investigate spontaneous speech with non-restricted spoken content and natural expression of speaker state, it was decided to first include the TU Munich Audiovisual Interest Corpus (TUM AVIC) (Schuller et al. 2009) as recently featured in the INTERSPEECH 2010 Paralinguistic Challenge (Schuller et al. 2010a) in the experiments. It is an audiovisual affective corpus containing recordings during which a product presenter leads one of 21 subjects (ten female) through an English commercial presentation. The “level of interest” is annotated for every “sub-speaker-turn” (for details on chunking the reader is referred to Schuller et al. 2009) and reaches from disinterest (subject is tired of listening and talking about the topic, is totally passive, and does not follow) together with indifference (subject is passive, does not give much feedback to the experimenter’s explanations, and asks unmotivated questions, if any) and neutrality (subject follows and participates in the discourse; it cannot be recognized, if she/he is interested or indifferent in the topic) (316 instances for the levels of interest so far) to interest (subject wants to discuss the topic, closely follows the explanations, and asks questions) (510 instances for this level of interest) to curiosity (strong wish of the subject to talk and learn more about the topic) (170 instances). Four annotators listened to the turns and rated them in terms of these three categories. The overall rating of the turn was computed from the majority label of the four annotators. If no majority label exists, the turn is discarded and not included in the database, leaving 996 turns in the database. All final utterances were recorded in an office environment with a headset and stored in 16 kHz sample rate and 16 bit. For the oncoming evaluations these 996 phrases are used as, e.g., employed in Schuller et al. (2006b, 2009).

4.2 eNTERFACE

The eNTERFACE corpus is a public, further audiovisual emotion database (Martin et al. 2006). 42 subjects (eight female) from 14 nations are included. Contained are office environment recordings of pre-defined spoken content in English. Each subject was instructed to listen to six successive short stories, each of them intended to elicit a particular emotion. They then had to react to each of the situations by uttering previously read phrases that fit the short story. Five phrases are available per emotion as “*I have nothing to give you! Please don’t hurt me!*” in the case of fear. Two experts judged whether the reaction expressed the intended emotion in an unambiguous way. Only if this was the case, a sample, i.e., sentence, was added to the database. Therefore, each sentence in the database has one assigned emotion label, which indicates the emotion expressed by the speaker in this sentence. All final 1 170 utterances, which are also used herein, were recorded in a small room furnished with electronic equipments. These contain induced anger (200), disgust (189), fear (189), happiness (205), sadness (195), and surprise (192) as emotions. The audio sample rate was 48 kHz, in an uncompressed stereo 16 bit format. Research results on acoustics-based automatic recognition of these speaker states are reported, e.g., in Datcu and Rothkrantz (2008), Mansoorizadeh and Charkari (2008), Paleari et al. (2008).

4.3 EMO-DB

Finally, the Berlin Emotional Speech database (EMO-DB) (Burkhardt et al. 2005)—likely the most frequently used in the field—will be shortly introduced. The spoken content is again pre-defined by ten German emotionally neutral sentences like “*Der Lappen liegt auf dem Eisschrank.*” (*The cloth is lying on the fridge.*). Ten (five female) professional actors were asked to express each sentence in seven emotional states. The sentences were labeled according to the state they should be expressed in, i.e., one emotion label was assigned to each sentence. It thus provides a high number of repeated words in diverse emotions. While the whole set comprises around 900 utterances, only 494 phrases are marked as minimum 60% natural and minimum 80% agreement by 20 subjects in a listening experiment. This selection is usually used in the literature reporting results on the corpus, as in Meng et al. (2007), Slavova et al. (2008), Schuller et al. (2008b), and in this article. 84.3% mean accuracy is the result of a perception study by 20 independent test-subjects for this limited 'more prototypical' sub-set that contains anger (127), boredom (79), disgust (38), fear (55), happiness (64), neutrality (53), and sadness (78) as emotions. All final 494 utterances were recorded in an anechoic chamber with high-quality recording equipments and saved in mono wave files with 16 kHz sample rate and 16 bit.

5 Acoustic features

In the experiments below, the recognition of affective speaker state is based on speaker turns (EMO-DB, eNTERFACE), as in the vast majority of approaches as by You et al. (2006) or sub-speaker turns (TUM AVIC) as given by the database. Thus, every such speech chunk is windowed in equidistant segments. A Hamming window is used for contours in the time domain and the speech signal is analyzed by using frames of 20 ms for every 10 ms. For every window, a set of Low Level Descriptors (LLD) is calculated and subsequently smoothed by simple moving average low-pass filtering. These LLD can be temporal characteristics as signal envelope and energy, spectral or cepstral characteristics as information on the formants or Mel frequency cepstral coefficients (MFCC) or descriptors of voice quality as harmonics-to-noise-ratio (HNR) and micro perturbations. For every speech chunk of analysis, functionals are calculated from these time series of LLD, e.g., mean values, standard deviations, quartiles, extremes, etc. Additionally, these functionals are calculated from temporal delta coefficients of each LLD.

Table 3 gives an overview on LLD and functionals used for systematic feature space brute-forcing-based acoustic analysis in the context of the presented work. Altogether, 1 406 features are extracted by calculating the 19 functionals from each of the two times 37 Low Level Descriptors. This set is identical to former works by the author and others (cf. Batliner et al. 2006, 2011).

Table 3 Overview on Low Level Descriptors and functionals for chunk-level speech analysis

LLD (2 × 37)	Functionals (19)
Envelope	Mean
Energy	Standard Deviation
Pitch	Zero–Crossing–Rate
Formant 1–5 amplitude	Quartile 1
Formant 1–5 bandwidth	Quartile 2
Formant 1–5 frequency	Quartile 3
MFCC 1–16	Quartile 1—Minimum
Shimmer	Quartile 2—Quartile 1
Jitter	Quartile 3—Quartile 2
HNR	Maximum—Quartile 3
Δ Envelope	Centroid
Δ Energy	Skewness
Δ Pitch	Kurtosis
Δ Formant 1–5 amplitude	Maximum Value
Δ Formant 1–5 bandwidth	Relative Maximum Position
Δ Formant 1–5 frequency	Minimum Value
Δ MFCC 1–16	Relative Minimum Position
Δ Shimmer	Maximum Minimum Range
Δ Jitter	Position 95% Roll-Off-Ponit
Δ HNR	

6 Experimental protocol and results

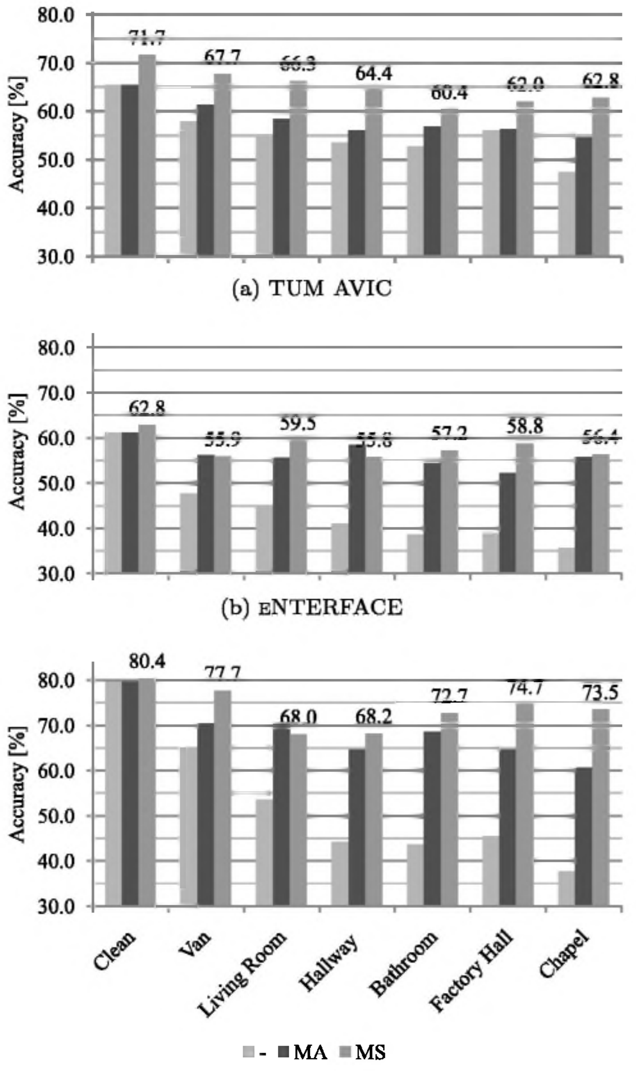
To simulate reverberation, the audio instances of the three affective speech databases as introduced in Sect. 4 are convoluted with the six respective room impulse responses as introduced in Sect. 3. Convolution is carried out in the spectral domain by multiplication of the Fast Fourier transforms of the speech turns and RIR avoiding clipping by according scaling.

As to obtain subject independent performances, it was decided for cyclic Leave-One-Speaker-Out (LOSO) as evaluation strategy to ensure strict speaker independence and at the same time easy reproducibility though exploiting maximum training set sizes given the sparseness of the corpora. However, each speaker is normalized to its mean and variance per feature employing the complete speaker turn context. As classifier it was decided for Support Vector Machines, as they are frequently encountered in the field and proven among best choices (Schuller et al. 2005). A linear kernel, pairwise multi-class discrimination and Sequential Minimal Optimization learning (Witten and Frank 2005) are used.

Two different types of experiments are considered: First, the general impact of reverberation on performance will be observed in non-matched condition. In addition, matching conditions in terms of acoustic model and feature space will be investigated. Second, closer insight on which feature type is more or less affected by reverberation in a non-matched condition will be given.

6.1 Acoustic model and space adaptation

In Fig. 2 a–c one can see the general impact on performance in a non-matched learning condition (left-most bars, each).



(a) TUM AVIC

(b) eNTERFACE

(c) EMO-DB

Fig. 2 Per feature type: speaker independent results by classification accuracy for clean and reverberated speech. Two counter-measures are additionally observed: matched acoustics (MA, *middle bar*, each) and additionally matched space (MS, *rightmost bar*, each) by matched conditions learning and testing

This means that the acoustic affective speaker state model is trained on non-reverberated speech, while the testing is carried out on reverberated such. In the figure, as well as in the further oncoming ones, accuracy is shown per database and testing condition in speaker independent LOSO testing. In comparison to the non-reverberated testing, a clear downgrade in performance is observable independent of the database and for all different types of RIR.

Next to this one finds the effect of matched condition learning to 'repair' the effect of reverberation (bars in the middle in each of Figs. 2a–c). Obviously, this does help significantly in any situation. Yet, the down grade by reverberation cannot be fully restored.

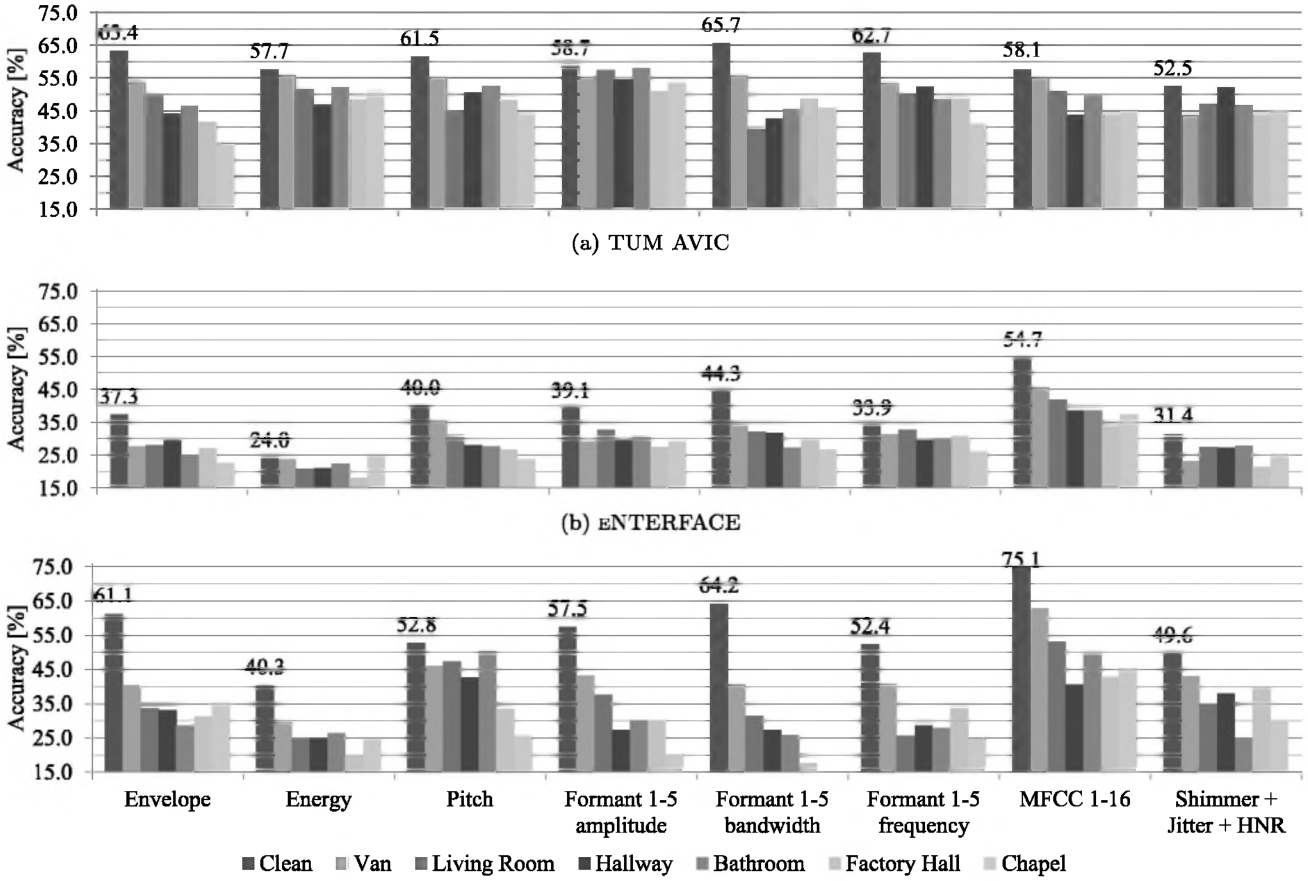
Features are next additionally selected by correlation based feature selection (CFS) (Hall 1998) in matched condition. By that, the acoustic model is trained matched to the reverberation situation and at the same time the optimal feature space for this type of reverberation is provided. This has been shown highly efficient in the case of additive noise (Schuller et al. 2008a). In the next section it will be clarified whether and which features are indeed impacted differently by different types of reverberation. CFS is chosen in order to optimize the space as a whole rather than agglomerating individual top ranked candidates. This optimization of the acoustic space is carried out each independent of the testing instances. The effect of the common adaptation of acoustic model and feature space in matched condition is again seen in Figs. 2a–c (right most bars, each). In three cases this is counter productive over exclusively adapting the acoustic model. However, in the vast majority of cases—18 in total—this results in another considerable boost of accuracy. One can see that it also helps improve in the non-reverberation test case. This fact is known from many other works (Ververidis and Kotropoulos 2006; Schuller et al. 2006c) as it reduces complexity for the classifier by de-correlation of the space and reduction of redundancy thus requiring less free parameters of the classifier to be trained.

Table 4 summarizes these observations: Averaged over the three databases and six RIR, the relative decrease from non-reverberated speech to reverberated such is observed at -29.8% relative, after matching the acoustic model it is highly significantly reduced to -12.7%, and after additionally matching the feature space to the type of RIR, a further highly significant reduction to only -9.8% relative downgrade is reached. Thus, the relative improvement reached by the suggested matching strategies resembles 20% on average.

Comparing impact of reverberation across databases, the automotive environment (*Van*) resulted in least degradation of accuracy, followed by the three home environments (*Living Room*, *Hallway*, then *Bathroom*—as one would expect). As a group, the larger non-home indoor environments (*Factory Hall*, *Chapel*) cause the most severe degradation. This is well predicted by the reverberation times: The higher the reverberation time, the higher the degradation for the observed RIR. The *Factory Hall* RIR is in fact the only RIR that varies in rank of degradation among these three databases and by that slightly disrupts this trend. Of course, given only six RIR, this finding has to be taken with a grain of salt.

6.2 Type-wise feature analysis

While we had seen in the last subsection that matching of the feature space leads to an improvement of recognition accuracy in the case of reverberated speech, an obviously



(a) TUM AVIC

(b) eNTERFACE

(c) EMO-DB

Fig. 3 Speaker independent results for clean and reverberated speech. Two counter-measures are observed: matched acoustics (MA) and additionally matched space (MS) by matched conditions learning and testing

interesting question is which features are impacted most by reverberation. To this end, these are now analyzed individually by looking at one type after the other. Note that it is refrained from reducing features per type to the same number of final features as investigated in Batliner et al. (2011), as in this study there is less interest in which feature type individually performs best, but rather in differences per type owed to different type of reverberation.

Figure 3a–c details this effect per feature type and per RIR for the three databases. Training is carried out exclusively on non-reverberated speech for this analysis. The general trends seem consistent over the different corpora, as MFCC being the best individual feature type—certainly also owed to this being the largest group of features—while the trends are more 'flat' for the eNTERFACE set which is the hardest task and leaves least headroom over chance level.

Table 5 summarizes the results per feature group and type by averaging over all considered types of reverberation and database. In addition to absolute and relative accuracy differences in comparison to non-reverberated speech, the standard deviation is provided. As can be seen, envelope is considerably more susceptible to reverberation than energy. For formants, bandwidth is most effected while at the same time the most relevant formant feature type for affect analysis. Next come—almost on par—amplitude and frequency. Generally least impacted is the voice quality group consisting of shimmer, jitter, and harmonics-to-noise ratio.

7 Conclusions

In this article first insights on performance downgrade by artificial reverberation with real-world room impulse responses on speech for testing of affective speaker state classification were reported. Such considerations will be of increasing importance when emotion or related speaker state and trait classification is used in distant-talking application, as, e.g., surveillance or human-robot communication.

The presented results can be seen as a first step towards effects of and dealing with different types of reverberation of speech in this application context. They document that counter-measures will be needed given an average relative performance loss by 29.8 percentage points in accuracy. They do, however, also show that once the type of reverberation is known, even rather simple counter measures as the presented acoustic model and space adaptation can help to reduce this loss to 9.8 percentage points.

Table 4 Mean downgrade in classification accuracy over the six different types of reverberation in mismatched condition learning on non-reverberated ('clean') speech considered when testing on reverberated speech. Two counter-measures are additionally observed: matched acoustics (MA) and additionally matched space (MS) by matched conditions learning and testing

Accuracy [%]	–	MA	MS
TUM AVIC			
Clean	65.5	65.5	71.7
mean (w/o clean)	54.0	57.3	63.9
Δ absolute	−11.5	−8.2	−7.8
Δ relative	−17.5	−12.5	−10.9
eNTERFACE			
Clean	61.1	61.1	62.8
mean (w/o clean)	41.2	55.5	57.3
Δ absolute	−19.9	−5.6	−5.5
Δ relative	−32.6	−9.2	−8.8
EMO-DB			
Clean	79.6	79.6	80.4
mean (w/o clean)	48.3	66.6	72.5
Δ absolute	−31.3	−13.0	−7.9
Δ relative	−39.3	−16.3	−9.8
mean across sets			
Δ absolute	−20.9	−8.9	−7.1
Δ relative	−29.8	−12.7	−9.8

For the impact of reverberation on feature types, the highest such was observed for formant bandwidth with an average relative downgrade of 41.5% and at the same time highest standard deviation over the different room acoustics at 6.0%. Least impacted—which seems intuitive—is voice quality and energy, at 'only' 20.0% and 21.1% relative average downgrade and at the same time least standard deviation at 3.9% and 2.9% over the different room acoustics.

Obvious needed next steps comprise analysis of dynamic reverberation and 'borrowing' of more sophisticated counter strategies for de-reverberation, e.g., based on processing of the LPC prediction residual or combination of blind channel estimation and channel inversion (Naylor and Gaubitch 2005). Also, multi-condition training was not tested in these experiments but may lead to an easily obtained improvement in unknown reverberation condition (Haderlein et al. 2005).

Further feature types can also be investigated including such obtained by time-frequency transformations (Kandali et al. 2009) or standard features as RASTA-PLP, which are, however, less common in the field of speaker state classification.

Acknowledgements The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

The author is presently a visiting scientist in the Imperial College London's Department of Computing in London/UK.

The research presented in this article highly benefited from the collaboration of the student assistant Yuan Ma.

The author would further like to thank Florian Völk for technical discussions on room impulse response acoustical properties and Fokke van Saane (Factory Hall, Amsterdam Living Room, Mercedes Van) and the recorders of the further employed room impulse responses.

Table 5 Per feature type (with number of features): mean accuracy over the three databases TUM AVIC, eNTERFACE, and EMO-DB for non-reverberated ('clean') speech and additionally over the six different types of reverberation ('w/o clean') in learning on non-reverberated speech. In addition, differences absolute (abs.) and relative (rel.) differences (Δ) and the standard deviation (std. dev.) are provided

Accuracy [%]	LLD Type	# Features	mean clean	mean w/o clean	Δ abs.	Δ rel.	std. dev. w/o clean
Prsodic	Envelope	38	53.9	34.4	−19.5	−36.1	4.6
	Energy	38	40.7	32.1	−8.6	−21.1	2.9
	Pitch	38	51.4	39.0	−12.5	−24.3	5.8
Spectral/Cepstral	Formant 1–5 amplitude	190	51.8	37.6	−14.2	−27.4	4.9
	Formant 1–5 bandwidth	190	58.1	34.0	−24.1	−41.5	6.0
	Formant 1–5 frequency	190	49.7	36.3	−13.3	−26.9	4.5
	MFCC 1–16	608	62.6	44.5	−18.2	−29.0	5.8
Voice Quality	Shimmer + Jitter + HNR	38	44.5	35.6	−8.9	−20.0	3.9

References

Athanaselis, T., Bakamidis, S., Dologlu, I., Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). ASR for emotional speech: clarifying the issues and enhancing performance. *Neural Networks, 18*, 437–444.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., & Aharonson, V. (2006). Combining efforts for improving automatic classification of emotional user states. In *Proc. IS-LTC 2006* (pp. 240–245), Ljubliana.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., & Amir, N. (2011). Whodunnit—searching for the most important feature types signalling emotional user states in speech. *Computer Speech and Language, 25*, 4–28.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. In *Proc. Interspeech* (pp. 1517–1520), Lisbon.

Campanini, S., & Farina, A. (2009). A new audacity feature: room objective acoustical parameters calculation module. In *Proc. Linux audio conference*.

Datcu, D., & Rothkrantz, L. J. M. (2008). Semantic audio-visual data fusion for automatic emotion recognition. In *Proc. Euromedia 2008, Eurosis*.

Grimm, M., Kroschel, K., Harris, H., Nass, C., Schuller, B., Rigoll, G., & Moosmayr, T. (2007). On the necessity and feasibility of detecting a driver's emotional state while driving. In A. Paiva, R. Prada, & R. W. Picard (Eds.), *Affective computing and intelligent interaction* (pp. 126–138). Berlin: Springer.

Haderlein, T., Nöth, E., Herbordt, W., Kellermann, W., & Niemann, H. (2005). Using artificially reverberated training data in distant-talking ASR. In *LNCS: Vol. 3658. Text, speech and dialogue* (pp. 226–233). Berlin: Springer.

Hall, M. A. (1998). *Correlation-based feature selection for machine learning*. PhD thesis, Hamilton, Waikato University, Department of Computer Science.

Kandali, A. B., Routray, A., & Basu, T. K. (2009). Vocal emotion recognition in five native languages of assam using new wavelet features. *International Journal of Speech Technology, 12*, 1–13.

Kim, E. H., Hyun, K. H., & Kwak, Y. K. (2005). Robust emotion recognition feature, frequency range of meaningful signal. In *Proc. IEEE international workshop on robots and human interactive communication (RO-MAN)* (pp. 667–671), Nashville, USA.

Lee, K. K., Cho, Y. H., & Park, K. S. (2006). Robust feature extraction for mobile-based speech emotion recognition system. In *Lecture notes in control and information sciences. Intelligent computing in signal processing and pattern recognition* (pp. 470–477). Berlin: Springer.

Lugger, M., Yang, B., & Wokurek, W. (2006). Robust estimation of voice quality parameters under real world disturbances. In *Proc. ICASSP* (pp. 1097–1100), Toulouse.

Mansoorizadeh, M., & Charkari, N. M. (2008). Bimodal person-dependent emotion recognition comparison of feature level and decision level information fusion. In *Proc. 1st international conference on pervasive technologies related to assistive environments* (pp. 1–4). New York: ACM.

Martin, O., Kotsia, I., Macq, B., & Pitas, I. (2006). The enterface'05 audio-visual emotion database. In *Proc. IEEE workshop on multimedia database management*.

Meng, H., Pittermann, J., Pittermann, A., & Minker, W. (2007). Combined speech-emotion recognition for spoken human-computer interfaces. In *Proc. international conference on signal processing and communications* (pp. 1179–1182), Dubai, United Emirates. New York: IEEE Press.

Naylor, PA, & Gaubitch, N. D. (2005). Speech dereverberation. In *Proc. 2005 international workshop on acoustic echo and noise control, EURASIP*.

Naylor, P., & Gaubitch, N. D. (2010). *Speech dereverberation*. London: Springer.

Neely, S. T. & Allen, J. B. (1979). Invertibility of a room impulse response. *Journal of the Acoustical Society of America, 66*, 165–169.

Paleari, M., Benmokhtar, R., & Huet, B. (2008). Evidence theory-based multimodal emotion recognition. In *Proc. 15th international multimedia modeling conference on advances in multimedia modeling* (pp. 435–446). Berlin: Springer.

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America, 95*, 1581–1592.

Pittermann, J., Pittermann, A., & Minker, W. (2010). Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology, 13*, 49–60.

Raja, G. S., & Dandapat, S. (2010). Speaker recognition under stressed condition. *International Journal of Speech Technology, 13*, 141–161.

Schröder, M., Cowie, R., Heylen, D., Pantic, M., Pelachaud, C., & Schuller, B. (2008). Towards responsive sensitive artificial listeners. In *Proc. 4th international workshop on human-computer conversation*, Bellagio.

Schuller, B., Jiménez Villar, R., Rigoll, G., & Lang, M. (2005). Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In *Proc. ICASSP* (Vol. I, pp. 325–328), Philadelphia.

Schuller, B., Arsić, D., Wallhoff, F., & Rigoll, G. (2006a). Emotion recognition in the noise applying large acoustic feature sets. In *Proc. speech prosody 2006*, Dresden.

Schuller, B., Köhler, N., Müller, R., & Rigoll, G. (2006b). Recognition of interest in human conversational speech. In *Proc. interspeech* (pp. 793–796), Pittsburgh.

Schuller, B., Reiter, S., & Rigoll, G. (2006c). Evolutionary feature generation in speech emotion recognition. In *Proc. international conference on multimedia and Expo ICME 2006* (pp. 5–8), Toronto, Canada.

Schuller, B., Seppi, D., Batliner, A., Meier, A., & Steidl, S. (2007). Towards more reality in the recognition of emotional speech. In *Proc. ICASSP* (pp. 941–944), Honolulu.

Schuller, B., Wimmer, M., Arsic, D., Moosmayr, T., & Rigoll, G. (2008a). Detection of security related affect and behaviour in passenger transport. In *Proc. interspeech* (pp. 265–268), Brisbane.

Schuller, B., Wimmer, M., Mösenlechner, L., Kern, C., Arsic, D., & Rigoll, G. (2008b). Brute-forcing hierarchical functionals for paralinguistics: a waste of feature space. In *Proc. ICASSP* (pp. 4501–4504), Las Vegas.

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., & Konosu, H. (2009). Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing Journal, 27*, 1760–1774. Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2010a). The INTERSPEECH 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010* (pp. 2794–2797), Makuhari, Japan.

Schuller, B., Steidl, S., Batliner, A., & Seppi, D. (2010b). Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Communication*. Special issue on "Sensing emotion and affect—facing realism in speech processing".

Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G. (2010c). Cross-corpus acoustic emotion recognition: variances and strategies. *IEEE Transactions on Affective Computing, 1.*

Slavova, V., Verhelst, W., & Sahli, H. (2008). A cognitive science reasoning in recognition of emotions in audio-visual speech. *International Journal Information Technologies and Knowledge*, *2*, 324–334.

Steidl, S., Batliner, A., Seppi, D., & Schuller, B. (2010). On the impact of children's emotional speech on acoustic and language models. *EURASIP Journal on Audio, Speech, and Music Processing, 2010*, 783954.

Tawari, A., & Trivedi, M. (2010). Speech emotion analysis in noisy real world environment. In *Proc. ICPR* (pp. 4605–4608), Istanbul, Turkey.

Ververidis, D., & Kotropoulos, C. (2006). Fast sequential floating forward selection applied to emotional speech features estimated on des and susas data collection. In *Proc. European signal processing conf. (EUSIPCO 2006)*, Florence.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques* (2nd edn.). San Francisco: Morgan Kaufmann.

Wöllmer, M., Eyben, F., Keshet, J., Graves, A., Schuller, B., & Rigoll, G. (2009). Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks. In *Proc. of ICASSP* (pp. 3949–3952), Taipei, Taiwan.

Yoon, W. J., Cho, Y. H., & Park, K. S. (2007). A study of speech emotion recognition and its application to mobile services. In *Lecture notes in computer science. Ubiquitous intelligence and computing* (pp. 758–766). Berlin: Springer.

You, M., Chen, C., Bu, J., Liu, J., & Tao, J. (2006). Emotion recognition from noisy speech. In *Proc. ICME* (pp. 1653–1656), Toronto.

You, M., Chen, C., Bu, J., Liu, J., & Tao, J. (2007). Manifolds based emotion recognition in speech. *Computational Linguistics and Chinese Language Processing*, *12*, 49–64.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58.