Improving Generalisation And Robustness Of Acoustic Affect Recognition

Florian Eyben
Technische Universität München
Institute for Human-Machine Communication
Munich, Germany

Björn Schuller
JOANNEUM RESEARCH Forschungsgesellschaft mbH*
DIGITAL - Institute for Information and Communication Technologies
Graz, Austria

Gerhard Rigoll
Technische Universität München
Institute for Human-Machine Communication
Munich, Germany

ABSTRACT

Emotion recognition in real-life conditions faces several challenging factors, which most studies on emotion recognition do not consider. Such factors include background noise, varying recording levels, and acoustic properties of the environment, for example. This paper presents a systematic evaluation of the influence of background noise of various types and SNRs, as well as recording level variations on the performance of automatic emotion recognition from speech. Both, natural and spontaneous as well as acted/prototypical emotions are considered. Besides the well known influence of additive noise, a significant influence of the recording level on the recognition performance is observed. Multi-condition learning with various noise types and recording levels is proposed as a way to increase robustness of methods based on standard acoustic feature sets and commonly used classifiers. It is compared to matched conditions learning and is found to be almost on par for many settings.

Categories and Subject Descriptors

H.5.5 [**Information Systems Applications**]: Sound and Music Computing

General Terms

Experimentation, Reliability

Keywords

Emotion recognition, noise robustness, recording level, multi-condition training

*The author is further affiliated with Technische Universität München, Munich, Germany

1. INTRODUCTION

Recently, a shift in emotion recognition research from working with controlled, lab-recorded data to more naturalistic data is observed. Challenges at INTERSPEECH from 2009 to 2011 literally challenged participants with highly naturalistic data for tasks ranging from identification of emotions all the way to identification of speaker states such as intoxication or sleepiness [12]. The reported results are still low, however. Even human agreement on these data-sets is far from perfect, indicating the difficulty of the task and the fuzziness of emotions.

When we want to use affect recognition technology in a live system, either a research demonstrator or a commercial product, robustness and high accuracies are a key factor. For achieving this goal, there are two main factors to consider: (a) optimising the classification performance on given data-sets, and (b) making the system robust to factors, such as background noise. For (a), an optimal choice of features and classifiers are the main research topics, which are dealt with extensively in many other studies such as the INTERSPEECH Challenge contributions. However, experience has shown, that often a system highly optimised for one data-set or task does not perform too well in cross data-set evaluations (cf. [14]), or when external factors such as the environment or the recording equipment change. Thus, we investigate issue (b) – the influence of background noise and level variations – in this paper.

Section 2 gives an overview on related work, section 3 describes our experiments and the proposed method of multi-condition training to build more robust models, section 4 describes the three corpora used for evaluations, and section 5 includes a discussion of the obtained results. A conclusion is drawn in section 6.

2. RELATED WORK

Compared to the large amount of work that has been published on optimising feature sets (e. g., [7]), fusing of modalities (e. g., [5]), and choice of classifiers (e. g., [4]) for affect recognition, few people have investigated the influence of additive noise, and none – to our best knowledge – the effect of level variations. Some work exists on the influence of the environment, specifically the reverberation caused by it (e. g., [15] and [8]).

The effects of additive noise were first investigated on a broad scale on three publicly available databases in [9].

No speaker independent evaluation was performed, however, and features specific to the noise type were selected. A similar study has been performed for noise types occurring in an in-car environment in [11]. Another report on in-car emotion recognition dealing with the effects of in-car noises, such as engine and road noise, is presented in [3]. Weninger et. al, present a method of improving recognition performance in noisy and reverberated settings by using Non-Negative Matrix Factorisation (NMF) based acoustic features in [15].

3. EXPERIMENTS

None of the methods referenced in section 2 have investigated the influence of recording level variations within the test set as well as extensive multi-condition training. Many studies from the ASR field deal with the topic of robustness and background noise, however the findings cannot be transferred to emotion recognition in most cases, as emotion recognition uses different types of features and a different approach to modelling. Low-level signal enhancement and noise reduction can be used in both fields, however this shall not be part of this study.

In previous studies typically matched conditions learning is used, which refers to training on data that is corrupted with the same noise type and Signal-to-Noise Ratio (SNR) as the test data. This is an unrealistic setting for a real-world application, as the noise type in many cases is not known when designing the system and training the models. Mismatched conditions training refers to training on the original, clean data and testing on noisy data. It investigates the performance of a generic model in varying noise conditions. A third alternative is multi-condition training, which combines benefits from matched and mismatched conditions training. Multiple copies of the data overlaid with different noise types at various SNRs are used during training. Thus, a generic model is generated, which is expected to work well in a variety of noise types. [15] investigates multi-condition learning and reports good performance on children's emotional speech from the FAU AIBO corpus.

In this paper, in addition to an extensive study on the influcence of additive noise under mismatched, matched, and multi-condition learning, we evaluate the influence of recording level variations for the first time. While the recording level plays a non-significant role in ASR, the correct recordning level is a major factor for the success of current emotion recognition systems. Level variations (scaling of the audio samples) by applying a gain of -20 dB to +10 dB are investigated without noise overlay. We would like to note at this point that this evaluation might seem unnecessary, if the recording level was normalised. However, this requires normalisation of the recording level of speech parts only. This is a non-trivial task if reverberation and esp. background noise are present.

Further, we overlay five noise types without scaling the original audio: babble noise, street noise, office noise, white noise, and music. For babble and street noise we use the Aurora noise samples (cf. [6]). The office noise consists of typical sounds occurring in a busy office environment, such as typing, printer machines, writing, beep sounds, and occasional talk in the background. The noise samples have been sampled from YouTube^{TM} videos which were recorded in office environments. The music noise type has been compiled from a mix of instrumental classic and pop music snippets (no vocals).

Our noise samples have lengths which vary from 1 (street noise) to 47 minutes (office noise). When overlaying an audio chunk with noise a random region of the noise sample is selected, and scaled accordingly to match the desired SNR before adding it to the audio chunk. The SNRs are computed based on Root Mean Square (RMS) amplitudes $A_{sig}^{(rms)}$ and $A_{noise}^{(rms)}$ of signal and noise chunks, respectively, according to the following equation:

$$SNR = 20 \log_{10} \frac{A_{sig}^{(rms)}}{A_{noise}^{(rms)}} \quad (1)$$

Six SNR levels are investigated: -5 dB, 0 dB, 5 dB, 10 dB, 20 dB, 30 dB.

In section 5 we provide and discuss results for matched conditions training, mismatched conditions training, and multi-condition training. Multi-condition training in general refers to the fact that a classifier is trained with target data (e.g. emotional speech sentences), which is repeated in the training material with different conditions applied to it, such as various background noise types or recording level variations. The multi-condition training approach chosen in this paper differs slightly from the one applied in [15]. Weninger et. al create duplicates of the data at all investigated SNRs and for all noise types. In contrast, we synthesise the training data by tripling the amount of original data, applying a random gain between -20 dB and +5 dB to each instance (sentence, phrase), and then adding randomly chosen noise samples from all noise types at random SNRs from the range of 0 dB to 60 dB. 5% of the instances are not overlaid with noise.

All evaluations are performed in a speaker independent manner by leave-one-speaker-out (LOSO) cross validation. We use the INTERSPEECH 2011 Speaker State Challenge acoustic feature set [12] and extract the features with openSMILE [2]. The feature set consists of 4,368 audio features, which are statistical functionals of an exhaustive set of low-level audio descriptors, such as loudness, fundamental frequency, voice quality, Mel Frequency Cepstral Coefficients (MFCC), etc. We use linear kernel Support Vector Machines (SVM) as classifier. The linear kernel allows for fast processing even in a high dimensional space, which is important for live emotion recognition. Results are reported in terms of unweighted average recall rate (UAR), which is the unweighted average of the per-class recall rates.

4. DATABASES

To asses the influence of additive noise and audio gain on different tasks, we present experimental results on three databases: (a) the Berlin Emotional Speech-Database (EmoDB) [1], (b) an acted telephone speech anger corpus, and (c) the TUM Audio-Visual Interest Corpus (AVIC) [10].

The choice of EmoDB is motivated by the fact that often models trained on small, acted data sets which contain prototypical emotions, perform well in cross-validation experiments on the same corpus [13], but show bad performance when evaluated cross-corpus [14] or when used in a live system. The anger database was chosen to asses the effect of audio gain on the discrimination performance between neutral and angry speech. AVIC was chosen to asses the effects of noise and scaling in a natural environment for spontaneous affect.

We briefly describe each of the data sets in the ongoing: EmoDB is a well known set chosen to test the effectiveness of

emotion classification. The studio recorded database covers seven emotions: *anger, boredom, disgust, fear, joy, neutral, and sadness*. The spoken content is pre-defined by ten emotionally neutral German sentences. Ten (five female) professional actors speak these sentences multiple times. While the whole set comprises around 900 utterances, only 494 phrases are marked as minimum 60 % natural and minimum 80 % assignable by 20 subjects in a listening experiment. 84.3 % mean accuracy – for identifying the 7 emotions – is the result of this perception study on this limited 'more prototypical' set. As this set is usually used in the manifold works reporting results on the corpus we restrict ourselves to this selection, as well. The 494 phrases have a combined recording length of approx. 23 minutes.

The TUM AVIC corpus contains recordings of personalised commercial product presentations. A product presenter leads each one of 21 subjects (10 female) through an English commercial presentation. The level of interest is annotated for every sub-speaker turn reaching from *boredom* (subject is bored with listening and talking about the topic, very passive, does not follow the discourse; this state is also referred to as level of interest (loi) 1, i.e. loi1), over *neutral* (subject follows and participates in the discourse, it can not be recognised, if she/he is interested or indifferent in the topic; loi2) to *joyful* interaction (strong wish of the subject to talk and learn more about the topic; loi3). Additionally, the spoken content and non-linguistic vocalisations are labeled in the AVIC set, however, not used in this study.

For our evaluation we consider all 3,002 phrases (approx. 108 minutes recording time), in contrast to only 996 phrases with high inter-labeler agreement as e. g. employed in [10]. However, we select a sub-sampled and re-sampled subset of these 3k phrases to eliminate problems caused by the uneven distribution of instances over the three interest levels. The sub-set is created by first sub-sampling the instances with a maximum number of 250 per class and the up-sampling the minority classes with a bias to an even distribution using the re-sampling algorithm implemented in the WEKA data mining toolkit [16]. 243 instances of loi1, 264 of loi2, and 233 of loi3 are retained.

	#	length (m:s)
Anger	319	57:07
Neutral	341	58:33

Table 1: Number of instances and combined recording length for each class in the Anger database.

The Anger database is a private database, which contains telephone quality speech recordings where callers were asked to speak a single English sentence in a neutral and in an acted way multiple times over a land-line telephone connection. There are 9 speakers in total, 4 female and 5 male. The number of instances for each class and the recording times are given in table 1. Unfortunately, the database is not publicly available for research. The mean sentence length is 10.5 seconds, the minimum length is 4.1 seconds and the maximum length 25.3 seconds.

5. RESULTS

The obtained results are summarised in table 4 for noise corruption and in table 2 for recording level variations. The

UAR [%]	gain/dB					
	-20	**-10**	**-5**	**0**	**+5**	**+10**
EmoDB (*mi*)	77.8	78.1	79.8	80.6	79.6	66.5
EmoDB (*ma*)	81.3	79.9	79.8	80.6	80.7	79.2
EmoDB (*mu*)	78.2	78.3	77.0	78.1	77.5	75.8
Anger (*mi*)	71.1	71.4	70.1	81.7	62.9	56.7
Anger (*ma*)	79.3	82.3	81.5	81.7	82.8	82.0
Anger (*mu*)	80.9	79.5	80.6	69.5	80.0	80.0
AVIC (*mi*)	46.0	45.2	48.8	59.4	53.2	51.7
AVIC (*ma*)	59.2	59.1	57.8	59.4	57.5	60.1
AVIC (*mu*)	58.0	58.5	59.4	56.5	59.2	58.8

Table 2: Results with gain applied to test partitions to simulate varying recording levels. Unweighted Average Recall (UAR) for leave-one-speaker-out cross validation. Training splits with original audio (mismatched condition, *mi*), audio with gain applied (matched condition, *ma*), and noise corrupted and scaled audio with all noise types (multi-condition, *mu*).

results in table 4 are an average of the individual results of the noise types. They show the general performance degradation for each data-set and each of the three training methods (mismatched condition (*mi*), matched condition (*ma*), and multi-condition (*mu*) training). The result for clean training and testing is given in the 0 dB column in table 2 for *mi* and *ma* evaluation (identical).

By looking at the *mi* results, we can see that recognition performance is heavily affected by additive noise at low SNRs. The effect is much greater as was reported in [9], which could be related to the use of non-hierarchical functionals as features in this paper and the fact that [9] did not perform speaker independent evaluation with leave-one-speaker-out cross validation. At -5 dB SNR the UAR for EmoDB and Anger drop to almost chance level (14.3% and 50% respectively). Performance on AVIC degrades substantially too, however it is still well above chance level (33.3%). A possible explanation is that AVIC contains a minimal amount of background noise and level variations in the training data. EmoDB, for example, has been recorded under studio conditions and the sentences have been normalised to RMS amplitude 1.0, therefore no level variations are present in the training set.

For SNRs of 5 dB and higher multi-condition training performs almost as well as matched conditions training, whereby clean training performs significantly worse (on a level of 0.01). Multi-condition training performance could be further improved for lower SNRs, if data with such SNRs was included in the training set (in our setting the lowest is 0 dB).

Applying a gain to the test data does affect the recognition performance (cf. table 2), but not as strongly as additive noise. The effect is strongest for AVIC (-13.4% for -20 dB gain on test partitions compared to unmodified test partitions) and the Anger database (-25% for +10 dB gain on test compared to unmodified data). A 14.1% decrease is observed for +10 dB gain on EmoDB, however only a 1% decrease for +5 dB is observed. We attribute the large decrease to non-linear distortions introduced through clipping when scaling by the 10 dB factor, as the original (0 dB) samples

are already normalised to the maximum peak amplitude. AVIC and Anger databases are recorded at lower levels.

Table 3 shows the confusion matrices for AVIC for LOSO evaluation on the unmodified audio, and -20 and +10 gain applied to the test splits. For the low gain test split (-20 dB) the distinguishability for separating loi3 from loi1 and loi2 decreases, while the distinguishability between loi1 and loi2 does not seem to be affected so much. This can be interpreted in a way that energy/loudness is an important feature for discriminating interested from neutral speech, while pitch and spectral characteristics are more important for neutral vs. boredom. For the high gain test split (+10 dB) the classifier shows a bias towards loi3. This bias is stronger for loi2 instances than for loi1 instances.

For multi-condition training a noteable decrease in performance for 0 dB gain is observable for Anger and AVIC compared to all other gain settings. The only explanation for this behaviour is that the muli-condition training data contains too little actual clean examples at 0 dB gain.

UAR [%]	SNR/dB					
	-5	**0**	**+5**	**+10**	**+20**	**+30**
EmoDB (*mi*)	16.4	17.7	23.5	32.9	54.7	71.0
EmoDB (*ma*)	74.6	77.0	78.1	79.5	81.6	80.2
EmoDB (*mu*)	45.2	65.1	72.7	74.9	78.1	78.7
Anger (*mi*)	56.9	59.7	62.3	65.7	62.9	63.1
Anger (*ma*)	74.6	77.0	78.1	79.5	81.6	80.2
Anger (*mu*)	68.7	74.6	77.8	79.6	81.6	81.5
AVIC (*mi*)	44.5	47.5	49.2	51.2	52.0	51.6
AVIC (*ma*)	56.3	57.4	58.2	59.0	59.1	58.5
AVIC (*mu*)	47.3	53.2	55.7	58.3	57.9	58.9

Table 4: Results for testing on noise corrupted audio with SNRs from -5 dB to 30 dB. Unweighted Average Recall (UAR) for leave-one-speaker-out cross validation. Training on original audio (mismatched condition, *mi*), noise corrupted audio (matched condition, *ma*), and noise corrupted and scaled audio with all noise types (multi-condition, *mu*).

Table 5 shows detailed results for each noise type on the AVIC set. As expected, for matched conditions training we see least variation in the results (min. 54.3 % UAR, max. 63.0 % UAR). Noteably, there is even an increase in performance over the clean case for some settings. This is most evident for white noise at an SNR of 30dB. This suggests that adding a small amount of white noise to the training data might improve generalisation performance. Besides that we cannot see a clear, significant trend as to which noise type affects the performance most over all SNRs. However, for both multi-condition and mismatched-condition training, we observe that noise types which degrade performance a lot at low SNRs seem to have the inverse effect at high SNRs (see white noise vs. babble noise).

6. CONCLUSION

We have evaluated the effects of additive noise on automatic affect recognition performance of naturalistic and acted affect. In contrast to previous work we chose a broad variety of emotional content. Moreover, this paper is the first to investigate the effect of recording level variation on

UAR [%]	SNR/dB					
	-5	**0**	**5**	**10**	**20**	**30**
Clean training (mismatched condition):						
babble	48.3	48.8	50.3	49.6	52.2	49.8
music	49.0	47.9	49.4	53.6	50.5	51.6
office	41.8	46.9	46.7	48.9	53.5	53.0
street	48.8	52.4	50.0	49.6	52.5	51.1
white	34.7	41.5	49.7	54.5	51.3	52.3
Matched condition training:						
babble	55.7	60.7	58.7	57.9	58.3	56.2
music	57.1	57.5	59.0	58.8	61.4	58.2
office	54.3	58.1	60.1	58.7	58.2	58.5
street	54.4	56.1	56.6	61.3	59.3	56.5
white	59.3	54.5	56.4	58.3	58.1	63.0
Multi-condition training:						
babble	54.8	56.9	52.6	57.8	57.9	57.2
music	48.8	52.3	56.8	60.6	59.8	58.9
office	44.2	49.9	58.7	57.7	58.4	58.6
street	50.3	55.3	53.2	56.5	56.9	59.6
white	38.5	51.4	57.1	58.8	56.3	60.3

Table 5: Detailed results for clean, matched condition, and multi-condition training with 5 noise types and 6 SNR levels. Unweighted Average Recall (UAR) for all noise types and SNRs on AVIC. Leave-one-speaker-out cross validation. For comparison: clean result = 59.4 % UAR.

the recognition performance. We propose a method for robust multi-condition training of a single, robust model, which is able to deal with additive noise and varying recording gain. For SNRs of 5 dB or higher the performance of our multi-condition approach is almost on par with matched conditions training. A significant effect of the recording gain on recognition performance was found for the naturalistic data-set (TUM AVIC), and the task of anger detection. This effect is almost completely eliminated by multi-condition training.

Future work will deal with extending the multi-condition learning approach to reverberation effects and more diverse noise types, as well as an in-depth analysis of which acoustic features are most affected by noise and recording level. This paper helps to better understand the influence of some external factors commonly encountered when building a real-world, speaker independent emotion and affect recognition system from given data-sets and successfully proposed an approach to build more robust models by synthesising new training data.

7. 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. 289021 (ASC-Inclusion).

8. REFERENCES

[1] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. A database of german emotional speech. In *Proc. INTERSPEECH 2005, Lissabon, Portugal* (2005), pp. 1517–1520.

[#]	-20 dB			0 dB			+10 dB		
as →	**loi1**	**loi2**	**loi3**	**loi1**	**loi2**	**loi3**	**loi1**	**loi2**	**loi3**
loi1	**140**	65	37	**170**	65	7	**126**	56	60
loi2	95	**111**	57	106	**124**	33	78	**60**	125
loi3	44	100	**88**	27	64	**141**	37	9	**186**
Sum	279	276	182	303	253	181	241	125	371

Table 3: Confusion matrices for AVIC with -20 dB, 0 dB, and +10 dB gain applied to the test split.

[2] Eyben, F., Wöllmer, M., and Schuller, B. openSMILE – the munich versatile and fast open-source audio feature extractor. In *Proc. ACM Multimedia (MM), Florence, Italy* (2010), ACM, pp. 1459–1462.

[3] Jones, C., and Jonsson, I. Performance analysis of acoustic emotion recognition for in-car conversational interfaces. In *Universal Access in Human-Computer Interaction. Ambient Interaction*, vol. 4555 of *Lecture Notes in Computer Science*. Springer, 2007, pp. 411–420.

[4] Kockmann, M., Burget, L., and Cernocky, J. Brno university of technology system for INTERSPEECH 2010 paralinguistic challenge. In *Proc. of INTERSPEECH 2010* (2010), ISCA, pp. 2822–2825.

[5] Mower, E., Lee, S., Mataric, M. J., and Narayanan, S. Joint-processing of audio-visual signals in human perception of conflicting synthetic character emotions. In *Proc. of ICME 2008, Hannover, Germany* (June 2008).

[6] Pearce, D., and Hirsch, H.-G. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ICSLP '00* (Beijing, China, October 2000).

[7] Rahman, T., Marioryad, S., Keshavamurthy, S., Liu, G., Hansen, J., and Busso, C. Detecting sleepiness by fusing classifiers trained with novel acoustic features. In *Proc. of INTERSPEECH 2011, Florence, Italy* (2011), ISCA.

[8] Schuller, B. Affective speaker state analysis in the presence of reverberation. *International Journal of Speech Technology 14*, 2 (2011), 77–87.

[9] Schuller, B., Arsic, D., Wallhoff, F., and Rigoll, G. Emotion recognition in the noise applying large acoustic feature sets. In *Proc. Speech Prosody 2006* (May 2006), ISCA.

[10] Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., and Konosu, H. Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application. *Image and Vision Computing Journal (IVCJ), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior 27*, 12 (November 2009), 1760–1774.

[11] Schuller, B., Rigoll, G., Grimm, M., Kroschel, K., Moosmayr, T., and Ruske, G. Effects of in-car noise-conditions on the recognition of emotion within speech. In *Proc. DAGA 2007, Stuttgart* (March 2007), DEGA, pp. 305–306.

[12] Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. The INTERSPEECH 2011 speaker state challenge. In *Proc. of INTERSPEECH 2011, Florence, Italy* (2011), ISCA, pp. 3201–3204.

[13] Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., and Wendemuth, A. Acoustic emotion recognition: A benchmark comparison of performances. In *Proc. of ASRU* (2009), IEEE, pp. 552–557.

[14] Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., and Rigoll, G. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing (TAC) 1*, 2 (December 2010), 119–131.

[15] Weninger, F., Schuller, B., Batliner, A., Steidl, S., and Seppi, D. Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization. *EURASIP Journal on Advances in Signal Processing 2011* (2011).

[16] Witten, I. H., and Frank, E. *Data mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.