Detection of Negative Emotions in Speech Signals Using Bags-of-Audio-Words

Florian B. Pokorny[1,2,3,4], Franz Graf[1,5], Franz Pernkopf[3,5], Björn W. Schuller[1,6,7]
[1]Institute for Information and Communication Technologies, Joanneum Research Forschungsgesellschaft mbH, Graz, Austria
[2]Machine Intelligence & Signal Processing group, MMK, Technische Universität München, Germany
[3]Brain, Ears & Eyes – Pattern Recognition Initiative (BEE-PRI), BioTechMed-Graz, Austria
[4]Institute of Physiology, Medical University of Graz, Austria
[5]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
[6]Chair of Complex & Intelligent Systems, University of Passau, Germany
[7]Machine Learning Group, Department of Computing, Imperial College London, UK
Email: florian.pokorny@medunigraz.at

***Abstract*—Boosted by a wide potential application spectrum, emotional speech recognition, i. e., the automatic computer-aided identification of human emotional states based on speech signals, currently describes a popular field of research. However, a variety of studies especially concentrating on the recognition of negative emotions often neglected the specific requirements of real-world scenarios, for example, robustness, real-time capability, and realistic speech corpora.**

Motivated by these facts, a robust, low-complex classification system for the detection of negative emotions in speech signals was implemented on the basis of a spontaneous, strongly emotionally colored speech corpus. Therefore, an innovative approach in the field of emotion recognition was applied as the core of the system – the bag-of-words approach that is originally known from text and image document retrieval applications. Thorough performance evaluations were carried out and a promising recognition accuracy of 65.6 % for the 2-class paradigm negative versus non-negative emotional states attests to the potential of bags-of-words in speech emotion recognition in the wild.

I. Introduction

The field of speech emotion recognition deals with the automatic, computer-aided identification of human emotional or affective states on the basis of speech signals. Even though emotions do not necessarily influence the semantic content of an utterance, they provide feedback information playing an important role in human communication [1].

Already in the mid-1980s, the arising idea of making computers to detect emotions from the human voice resulted in first investigations (e. g., [2], [3]). One decade later, the improvements in computer technology allowed the implementation of more complex algorithms. Additionally, the market requirements for automatic services boosted further research. In recent years, the use of telecommunication services, multimedia devices and, consequently, the number of possible new applications rapidly increased. The multitude of studies describing a variety of new approaches as well as focusing on the search for powerful combinations of existing approaches (e. g., [4]), attests the actuality of the area of emotional speech recognition these days [1].

Similar to the field of automatic speech recognition (cf. [5]), architectural stages of feature extraction, feature processing, and classification play an important role in emotional speech recognition. Moreover, for both fields the need of a suitable speech corpus with appropriate annotations for training and testing the system is of immense value. As emotions cannot be measured by objective means, especially emotion recognition systems have to deal with ambiguities already in the development phase, caused by different annotators' interpretations of emotions. Hence, in the first place, emotion recognition systems can never achieve a better performance than the annotators did with respect to the underlying training data. In the second place, the annotators can hardly reflect the exact emotional state of a speaker [6].

The huge number of continuously arising approaches, models, algorithms and considerations dealing with an automatic detection of human emotional states based on speech, testifies the great popularity and presence of the field of emotional speech recognition. However, in spite of intensive research since almost 30 years (cf. [1]) that caused a considerable technological progress, the lack of large realistic emotional speech data collections still represents one of the main drawbacks limiting the performance of today's emotion recognition systems.

Apart from the requirement of spontaneity in emotional speech recordings and their preferable accurate annotations used as training data for a classifier, also the acoustic environment of the speaker plays an important role. It ought to be considered that many emotion recognition systems are intended for real-world applications and real-world scenarios are usually characterised by an overlay of multiple voices and background noise. Aspects concerning emotion recognition in context of spontaneous field data and speaker independence were discussed in Schuller et al. [7]. On the one hand, they showed results for two acted data collections originally exhibiting studio quality manipulated by step-wisely adding noise. On the other hand, the spontaneous FAU Aibo Emotion Corpus (cf. [8]) was investigated by comparing recognition

accuracies obtained when using audio from a standard headset microphone, from a room microphone, and artificially reverberated audio.

Beside acoustic real-world factors influencing the input quality of automatic recognition systems, also system complexity and, consequently, calculation time, may not be disregarded. Especially these days the market demand on high-capacity, but at the same time energy-efficient, mobile, electronic everyday tools forces the use of sophisticated algorithms, for example, when thinking of smart-phone applications. Furthermore, many applications must be capable of operating in real-time [9].

Over the last years lots of studies have especially focused on the investigation of negative emotions and their recognition, as well as on the recognition of stressed speech in everyday situations, which is justifiable by many application possibilities these days. For instance, in automated call centers, the detection of the telephoner's level of annoyance, anger, or frustration allows individual adapted answers and, thereby, enhances his/her satisfaction. A corresponding system that automatically classifies telephone quality speech into the speaker's emotional states neutral, low anger, and high anger, was described in Lee et al. [10]. They used speech recordings in which emotions were acted by amateurs and, thereafter, categorised by the actors themselves.

A wide operational area for all kinds of recognition tools that handle negative emotions is the detection of conflicts in spontaneous conversations. In this regard, in Kim et al. [11] an annotation scheme for the rating of the degree of conflict as well as an investigation of various prosodic conversational features based on recordings of broadcast political debates were presented. In another study, Kim et al. [12] revealed a method not for automatically recognising the level of conflict, but for detecting conflict escalations, i. e., if the level of conflict increases or not.

Solid systems for the detection of negative emotions in spontaneous real-word scenarios could contribute considerable advancements in the field of security applications. An acoustic monitoring of public places with a higher incidence of violence, such as subway or train stations as well as fan zones inside and around stadiums, in addition to a video surveillance system could facilitate a more efficient alerting of police and/or emergency forces. Also when thinking of bank robberies, an automatic alarming system triggered by acoustically detected emotions in the clients' and the employees' voices could have potential. Certainly, the main requirements for real-world emotion recognisers in applications like these are represented by a real-time processing and an insensitivity against stationary background noise and time-variant everyday sound events.

On the basis of existing approaches (e. g., [4]), the ideas for requested and innovative fields of application, and the non-neglectable conditions for a use in everyday life, the aim of this project was to develop a robust, but low-complex/calculation-efficient classification system for the detection of negative emotions in speech signals. On this purpose, the bag-of-words (BoW; pl. BoWs) method was implemented as the core of the system. BoWs represent a valuable instrument in information retrieval [13]. Beside their appliance for preparing feature data for classifying text documents or image documents (bag-of-visual-words approach) [14], the method was adapted to process audio features for the proposed system. BoWs in audio-related investigations were applied, for example, for multimedia event classification [14], video copy detection [15], content-based retrieval of digital music [16], detection of violence in movies [17], or the diagnosis and monitoring of depression [18]. To the best of our knowledge, BoWs have been rarely used in context of emotional speech recognition so far. This study provides a first comprehensive evaluation of the BoW approach as the essential component of an emotion recogniser on basis of a spontaneous, strongly emotionally colored, freely accessible speech corpus.

The remainder of this paper is organised as follows: Section II – describing the built recogniser in detail – is divided into three subsections that focus on the extraction of acoustic features, feature/BoW processing, and classification. In Section III the used speech corpus is introduced and recognition results are presented. The recogniser's performance in context of existing investigations in the field of emotional speech recognition is discussed in Section IV. Section V concludes the study and provides an outlook on possible improvements of the introduced system and, on implications for potential future work in emotional speech recognition applications in real-world scenarios.

II. Method

The data processing architecture of the proposed system comprises three stages. In the first stage, a speech signal is analyzed frame-by-frame by extracting acoustic features. In the second stage, the features are quantised and transformed into the BoW representation – in context of audio processing henceforth referred to as bag-of-audio-words (BoAW; pl. BoAWs) representation. Finally, BoAWs constitute the input data for the system's third stage, the classification stage.

A. Feature Extraction

To ensure reproducibility, feature extraction was carried out by means of the open-source toolkit openSMILE [19]. We further decided to extract the features of the official baseline feature set of the INTERSPEECH 2009 Emotion Challenge (cf. [20]) on a frame basis of $0.2\,\mathrm{s}$. As presented in Table I, in particular, 12 functionals were calculated for the trajectories of 16 low-level descriptors (LLDs) and their delta coefficients, leading to a feature vector of length 384 for each frame.

B. Feature/Bag-of-Audio-Words Processing

In general, feature processing serves the purpose of transforming raw feature vectors into a certain kind of representation that provides benefits for the crucial classification step and/or for a potential data transmission (cf. [21]).

As the later generation of BoAWs required each feature vector's representation as a single, discrete symbol, the first feature processing step passed through consecutively by each

TABLE I
LOW-LEVEL DESCRIPTORS AND FUNCTIONALS OF THE USED FEATURE SET (CF. [20]). LLD = LOW-LEVEL DESCRIPTOR, ZCR = ZERO-CROSSING RATE, RMS = ROOT MEAN SQUARE, F0 = FUNDAMENTAL FREQUENCY, HNR = HARMONICS-TO-NOISE RATIO, MFCC = MEL-FREQUENCY CEPSTRAL COEFFICIENT.

LLDs	Functionals
ZCR, ΔZCR	mean, standard deviation
RMS energy, ΔRMS energy	kurtosis, skewness, linear
F0, ΔF0	regression (offset, slope,
HNR, ΔHNR	MSE), extremes (values,
MFCCs 1–12, ΔMFCCs 1–12	relative position, range)

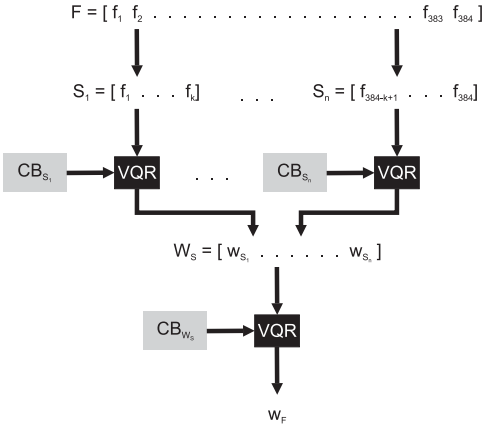

Fig. 1. Split vector quantization (SVQ) procedure applied to an input feature vector F. CB = codebook, VQR = vector quantiser.

extracted feature vector was its compression by applying vector quantization (VQ, cf. [22]). For the created system an hierarchical subspace quantization scheme was implemented due to better recognition results compared to those when testing conventional VQ. The applied method, known as split vector quantization (SVQ), is based on the use of independent codebooks for multiple feature subspaces [21]. Figure 1 illustrates the SVQ procedure applied to an input feature vector F handed over from the system's feature extraction stage.

In a first step, the input vector consisting of 384 feature values is divided into an adjustable[1] number of n sub-vectors of length k ($k = 384/n$). Next, VQ is applied to each sub-vector by using individual codebooks of same length. This step results in a vector W_S which contains the symbols/words output from the particular sub-vector quantization stages. So far, the number of features was reduced from 384 to n and their values' variability was limited to a discrete value from 0 to the sub-vector codebooks' length. As a final step, the word vector W_S is mapped to an ultimate feature word w_F by again conducting VQ using another individual codebook of independent length. Hence, the SVQ algorithm attains the input feature vector's representation as one discrete word. The required number of bits for storing such a front-end word is

[1]The length of the original input feature vector (384) must be divisible by n without remainder.

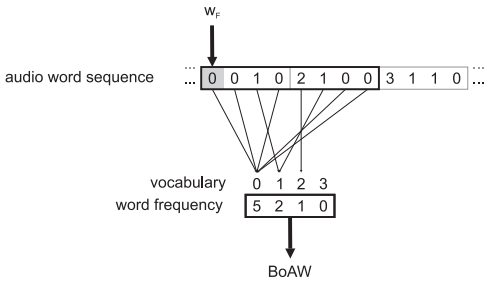

Fig. 2. Concept of generating a bag-of-audio-words (BoAW) from an input sequence of discrete 2-bit symbols/audio words.

determined by the codebook length of the last quantization step.

The split vector quantiser's (SVQR's) codebooks were generated by applying the k-means clustering algorithm (cf. [23]) to the training partition of the speech database that was selected for the system's development, training and evaluation. This database will be introduced in Section III.

Subsequent to the transformation of each 384-dimensional feature vector into one discrete (audio) word using the SVQ algorithm, the second feature processing step, i. e., the generation of BoAWs was carried out. Figure 2 demonstrates the basic concept of transforming a sequence of discrete 2-bit audio words into the BoAW representation.

In a first step, the audio word sequence is usually buffered into overlapping frames of an adjustable length, i. e., the BoAW buffer length. Subsequently, each frame is mapped to a vector of fixed length specified by the number of codewords used for the preceding VQ procedure and constituting both the VQR's and the BoAW processor's vocabulary [16]. Each output vector representing one BoAW simply encodes the frequency of occurrence of each of the vocabulary's audio words in a predefined, usually ascending order within one frame [16].

To attain a more convenient value range within each bag that led to better classification results in preliminary evaluation tests during the implementation phase, the decadic logarithm was applied to each bin value after adding the value 1 for avoiding the log of zero.

Generally, the projection of a word sequence to BoWs involves a loss of timing information by assuming that the order of words within each bag is irrelevant [13]. In other words, the generation of BoWs is based on counting the occurrences of words from a vocabulary within sets of words not taking into account the words' exact positions within the sets [5].

C. Classification

By reasons of its computational simplicity we used the naive Bayes approach for classification. Furthermore, we could not achieve better recognition results by more advanced techniques such as support vector machines or neural networks.The naive Bayes classifier is known for providing a good performance for lots of real-world data sets [23].

III. Evaluation

A. Material

After a thorough investigation of different existing speech corpora, the choice of an appropriate speech data collection for training and testing the implemented system fell on the Vera am Mittag (VAM) German Audio-Visual Emotional Speech Database (VAM corpus/database) that was collected at the Communications Engineering Lab of the Karlsruhe Institute of Technology (cf. [24]) in 2008. The VAM corpus comprises 12 hours of audio-visual material extracted from recordings of the German television talk show Vera am Mittag (Vera at noon). It contains spontaneous utterances from discussions between the talk show guests. Due to the particular topics mainly concerning personal and very emotional issues, for example, fatherhood questions or affairs, a variety of affective states is provided. The fact that the talk show guests were not aware of the recordings' later use for purposes of emotion analysis constitutes an essential advantage with regard to the requirement of authenticity in emotional speech corpora. Furthermore, it was assured that the discussions were unscripted and the guests did not perform as paid lay actors [24].

The VAM database is divided into the three sub-databases VAM-Video, VAM-Audio and VAM-Faces. For training and testing the built emotion recogniser exclusively the VAM-Audio database was applied. VAM-Audio consists of 947 utterances by 47 speakers with an average duration of 3 s. The average number of utterances per speaker is 21.7 [24].

Advantageously, the VAM database additionally provides emotion annotations. One part of the database comprising the utterances of 19 speakers was evaluated by 17 annotators. The other part comprising the utterances of 28 speakers was evaluated by 6 annotators. For evaluation the method of self assessment manikins according to Lang [25] was used. Thereby, the emotional content of each utterance was three-dimensionally interpreted in terms of the emotion primitives valence, activation and dominance [24].

Consistent with the original intent of the recognition system to detect negative emotions, that is to distinguish between negative and non-negative emotions, the corpus' utterances were assigned to either a defined negative emotion class 1 or a defined non-negative emotion class 0 depending on the annotators' ratings for the entity valence.

As valence for most of the utterances was rated as slightly negative the decision boundary between class 0 and class 1 was placed at -0.2 instead of 0. Thereby, a roughly balanced number of utterances for each class could be achieved. Figure 3 illustrates the applied division of emotion classes within the three-dimensional emotion space spanned by the primitives valence, activation and dominance within the normalised interval of [-1,1] . The gray volume indicates the defined negative emotion space, the unfilled volume the non-negative emotion space.

Furthermore, the division into a training and a test partition was carried out. To assure speaker independence, for each class the test partition was composed from utterances of the last third of the corpus' speakers. The BoAW buffer length was individually set to the lengths of each utterance. Consequently, one BoAW was calculated for each utterance representing one training or test instance. Table II shows the instance distribution of both classes 0 and 1 and the training and the test partitions as well as the respective overall number of instances used for the system's performance evaluation.

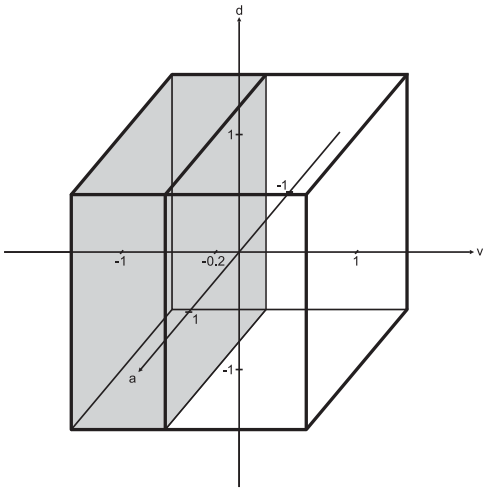

Fig. 3. Division of the three-dimensional emotion space spanned by the normalised emotion primitives valence, activation and dominance into two classes. The non-negative class 0 is represented by the unfilled block, the negative class 1 by the gray block. a = activation, d = dominance, v = valence.

TABLE II
Total and specific number of instances for the classes 0 and 1 applied for evaluating the system's recognition performance.

#	**0**	**1**	Σ
train	318	420	738
test	133	76	209
Σ	451	496	**947**

Admittedly, beside many essential advantages of the VAM database, such as its free availability for research purposes, a considerable number of segmented, emotionally colored, spontaneous, unscripted utterances extracted from real discussions, as well as their annotations by several listeners following internationally approved labeling methods, the lack of everyday background noise, the fact that more than one speaker never talks at once, and the speech materials restriction to German language, slightly reduces the data collections value for the training of robust, global, real-world applications.

B. Results

Table III reveals the system's recognition results for varying basic settings influencing the SVQ process, i. e., for a varying number of feature sub-vectors and varying sizes of both the sub-vector codebook and the SVQR's ultimate codebook.

The accuracy values in Table III are rounded to one decimal point. Considering both the unweighted accuracy (UA) and the class weighted accuracy (WA), the best recognition results could be achieved by dividing the original feature vector

TABLE III
RECOGNITION RESULTS FOR VARYING BASIC SETTINGS. SV = SUB-VECTOR, SVCS = SUB-VECTOR CODEBOOK SIZE, CS = CODEBOOK SIZE, UA = UNWEIGHTED ACCURACY, WA = WEIGHTED ACCURACY.

Configuration			Recognition	
# SV	SVCS [bit]	CS [bit]	UA [%]	WA [%]
2	2	4	61.6	65.1
2	4	2	60.3	62.7
3	**2**	**5**	**64.2**	**65.6**
3	2	6	63.7	65.6
3	4	2	64.6	64.6
3	4	3	63.8	63.2
4	4	3	62.1	61.7
6	3	3	63.2	63.2
8	4	3	65.3	62.7
12	3	4	63.4	64.1

TABLE IV
RECOGNITION RESULTS FOR A VARYING CODEBOOK SIZE WHEN REPLACING THE SYSTEM'S SPLIT VECTOR QUANTISER BY A SINGLE VECTOR QUANTISER. CS = CODEBOOK SIZE, UA = UNWEIGHTED ACCURACY, WA = WEIGHTED ACCURACY.

Configuration	Recognition	
CS [bit]	UA [%]	WA [%]
2	59.3	60.8
3	59.2	61.7
4	**61.6**	**63.6**
5	57.9	60.8
6	54.0	57.9
7	54.1	57.4
8	53.3	57.4
9	52.8	56.0

into 3 sub-vectors and by using a sub-vector codebook size and a SVQR's ultimate codebook size of 2 bits and 5 bits, respectively. This configuration led to an UA of 64.2 % and a WA of 65.6 %. By comparison, using support vector machines with linear kernel of complexity $C = 1.0$ (trained based on sequential minimal optimization) instead of the naive Bayes classifier an UA of 65.2 % and a WA of 62.2 % were achieved.

Whereas in other studies, e. g., [21], the SVQ algorithm served the purpose of compressing each raw feature vector to a sufficient number of attributes, the proposed system requires the mapping of each raw feature vector to one discrete symbol by reason of the subsequent generation of BoAWs. Thus, the SVQR consisting of two hierarchical VQ stages could also be replaced by a single, conventional VQ stage. However, when applying SVQ instead of conventional VQ a better recognition performance could be achieved. Table IV presents the respective naive Bayes classification results for the case of applying conventional VQ for varying codebook sizes.

For just using the unprocessed feature vectors extracted for each utterance with its individual length as feature extraction frame length, i. e., raw instances each consisting of 384 feature values, an UA of 54.3 % and a WA of 61.2 % were achieved using the naive Bayes classifier. Support vector machines – again with linear kernel of complexity $C = 1.0$ – performed even worse ($UA = 48.0\%$ and $WA = 46.0\%$).

IV. DISCUSSION

A comparison of the proposed system to other systems described in various studies is very difficult due to the huge number of degrees of freedom when both implementing a recognition system and evaluating its performance. Depending on the applied feature set, the feature processing steps, and the classification technique, as well as the speech corpus and the method for generating training and test partitions from the corpus reduces the meaningful comparability between different systems. Moreover, a system's intended application should always be considered. For example, a robust system for providing recognition results in real-time might be implemented in a different way from a system with no requirements concerning calculation complexity.

Beside its actuality, the study by Han et al. [21] was chosen to serve as a reference for the proposed system, because (i) the same set of features was extracted by means of openSMILE, (ii) also SVQ was applied by using the k-means algorithm for creating the codebooks, and (iii) the system evaluation was similarly carried out utterance-by-utterance on the basis of a German, non-acted speech corpus that was divided into a negative and a non-negative emotion class. However, in contrast to the proposed system, in Han et al. [21] SVQ was applied in order to reduce the number of bits for representing each front-end feature vector. For this purpose, only one VQ stage was passed through by the particular feature sub-vectors. As no further feature processing steps are performed subsequent to the SVQ procedure the classifier's input vector in the system by Han et al. [21] corresponds to the vector W_S (see Figure 1) in the proposed system that constitutes the output of the SVQR's first VQ stage. In Han et al. [21] support vector machines with a linear kernel, a complexity of 0.05 and a pairwise multi-class discrimination on the basis of sequential minimal optimization were employed for classification. The intent of the study by Han et al. [21] was to find a trade-off between feature compression via SVQ and recognition accuracy by testing different numbers of feature sub-vectors and codebook sizes. For evaluation the FAU Aibo Emotion Corpus (cf. [8]) was used in Han et al. [21]. The instances from a total of 18216 speech chunks labelled as either emotionally negative or non-negative were divided into a training and a test partition in consideration of speaker independence. In Han et al. [21] an UA of 67.4 % and a WA of 69.1 % could be achieved [21].

Even though the proposed system differs from the system in Han et al. [21] with respect to processing steps subsequent to the SVQ procedure, it might be inferred that the BoAW approach shows potential in its appliance for emotion recognition. Anyway, when comparing the performance of the proposed system with the system in Han et al. [21] it must be considered that – beside the systems' implementations – also the corpora used for evaluation are different. Different

corpora involve different annotations that restrict the classification system's performance to the performance of the human annotators. The corpus used in Han et al. [21] comprises about 20 times as many utterances as VAM-Audio.

V. Conclusion and Outlook

In consideration of the initial aims of this study, it can be concluded that the final system implementation fulfills the requirements. On the one hand, mainly robust, approved and non-complex algorithms were applied to allow for a reliable and calculation-efficient system performance. On the other hand, an innovative approach in the field of emotional speech recognition, i. e., the BoAW approach, was embedded as the core of the system. For preparing the system to a potential appliance in any kind of real-world scenario, the classifier was trained on the basis of a non-acted, strongly emotionally colored speech corpus of sufficient size. The system's performance was thoroughly evaluated. The final implementation turned out to achieve recognition results that can keep up with results reported for similar systems in comparable classification scenarios. Hence, the BoAW approach seems to work well in emotion recognition tasks. However, its real potency in contrast to other methods might appear under real-world conditions due to its straightforward concept. Moreover, the VQ procedure required for the generation of BoAWs guarantees that new acoustic input data are always related to the system's most similar prototype data. Thereby, the system always provides the best matching results, regardless of potential acoustic disturbances.

As the presented system serves as a proposed initial setup for a variety of potential real-world applications, its practical real-time installation in respective scenarios for test and evaluation purposes would be of great interest. However, even though the VAM-Audio corpus fulfills essential requirements that allow for modeling non-acted, rather spontaneous emotions, for the proposed system's use in real-world scenarios a different training corpus might be needed due to the fact that an acoustic real-world environment is characterised by speech simultaneously produced by more than one speaker as well as by stationary and/or transient background noise. The adaption of the VAM database by combining multiple utterances and/or by superposing utterances with background noise could be a first future step.

Acknowledgment

The research leading to these results has received funding from the European Union's Horizon 2020 Framework Programme under grant agreement No. 645094 (SEWA). The authors are grateful to Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan for creating the Vera am Mittag German Audio-Visual Emotional Speech Database and making it accessible to the community.

References

[1] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[2] R. Van Bezooijen, *Characteristics and Recognizability of Vocal Expressions of Emotion*. Foris Publications Holland, 1984.

[3] F. J. Tolkmitt and K. R. Scherer, "Effect of experimentally induced stress on vocal parameters," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 12, no. 3, pp. 302–313, 1986.

[4] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller, "Distributing Recognition in Computational Paralinguistics," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 406–417, October–December 2014.

[5] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Second Edition*, 2nd ed. Pearson Education, 2009.

[6] J. Pittermann, A. Pittermann, and W. Minker, *Handling Emotions in Human-Computer Dialogues*. Springer, 2009.

[7] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 4. IEEE, 2007, pp. 941–944.

[8] S. Steidl, *Automatic Classification of Emotion Related User States in Spontaneous Children's Speech*. Logos, 2009.

[9] F. Weninger, M. Wöllmer, and B. Schuller, "Emotion recognition in naturalistic speech and language - A survey," in *Emotion Recognition: A Pattern Analysis Approach*, 1st ed., A. Konar and A. Chakraborty, Eds. Wiley, December 2015, ch. 10, pp. 237–267.

[10] F.-M. Lee, L.-H. Li, and R.-Y. Huang, "Recognizing low/high anger in speech for call centers," in *Proceedings of the International Conference on Signal Processing, Robotics and Automation*. World Scientific and Engineering Academy and Society, 2008, pp. 171–176.

[11] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2012, pp. 5089–5092.

[12] S. Kim, S. H. Yella, and F. Valente, "Automatic detection of conflict escalation in spoken conversation," in *Proceedings of Interspeech*, 2012.

[13] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, 2002.

[14] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Proceedings of Interspeech*, 2012.

[15] Y. Liu, W.-L. Zhao, C.-W. Ngo, C.-S. Xu, and H.-Q. Lu, "Coherent bag-of audio words model for efficient large-scale video copy detection," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2010, pp. 89–96.

[16] M. Riley, E. Heinen, and J. Ghosh, "A text retrieval approach to content-based audio retrieval," in *Proceedings of the International Symposium on Music Information Retrieval*, 2008, pp. 295–300.

[17] E. Acar and S. Albayrak, "DAI Lab at MediaEval 2012 affect task: The detection of violent scenes using affective features," in *MediaEval*, 2012.

[18] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal assistive technologies for depression diagnosis and monitoring," *Journal on MultiModal User Interfaces*, vol. 7, no. 3, pp. 217–228, 2013.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[20] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of Interspeech*, 2009, pp. 312–315.

[21] W. Han, Z. Zhang, J. Deng, M. Wöllmer, F. Weninger, and B. Schuller, "Towards distributed recognition of emotion from speech," in *Proceedings of the International Symposium on Communications Control and Signal Processing*. IEEE, 2012, pp. 1–4.

[22] R. M. Gray, "Vector quantization," *ASSP Magazine*, vol. 1, no. 2, pp. 4–29, 1984.

[23] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, 4th ed. Academic Press, 2009.

[24] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proceedings of the International Conference on Multimedia and Expo*. IEEE, 2008, pp. 865–868.

[25] P. J. Lang, "Behavioral treatment and bio-behavioral assessment: Computer applications," in *Technology in Mental Health Care Delivery Systems*, J. B. Sidowski, J. H. Johnson, and T. A. Williams, Eds. Norwood, NJ: Ablex, 1980, pp. 119–137.