Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2011, Article ID 838790, 16 pages
doi:10.1155/2011/838790

Research Article

Recognition of Nonprototypical Emotions in Reverberated and Noisy Speech by Nonnegative Matrix Factorization

Felix Weninger,[1] Björn Schuller,[1] Anton Batliner,[2] Stefan Steidl,[2] and Dino Seppi[3]

[1] *Lehrstuhl für Mensch-Maschine-Kommunikation, Technische Universität München, 80290 München, Germany*
[2] *Mustererkennung Labor, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany*
[3] *ESAT, Katholieke Universiteit Leuven, 3001 Leuven, Belgium*

Correspondence should be addressed to Felix Weninger, weninger@tum.de

Received 30 July 2010; Revised 15 November 2010; Accepted 18 January 2011

Academic Editor: Julien Epps

We present a comprehensive study on the effect of reverberation and background noise on the recognition of nonprototypical emotions from speech. We carry out our evaluation on a single, well-defined task based on the FAU Aibo Emotion Corpus consisting of spontaneous children's speech, which was used in the INTERSPEECH 2009 Emotion Challenge, the first of its kind. Based on the challenge task, and relying on well-proven methodologies from the speech recognition domain, we derive test scenarios with realistic noise and reverberation conditions, including matched as well as mismatched condition training. As feature extraction based on supervised Nonnegative Matrix Factorization (NMF) has been proposed in automatic speech recognition for enhanced robustness, we introduce and evaluate different kinds of NMF-based features for emotion recognition. We conclude that NMF features can significantly contribute to the robustness of state-of-the-art emotion recognition engines in practical application scenarios where different noise and reverberation conditions have to be faced.

1. Introduction

In this paper, we present a comprehensive study on automatic emotion recognition (AER) from speech in realistic conditions, that is, we address spontaneous, nonprototypical emotions as well as interferences that are typically encountered in practical application scenarios, including reverberation and background noise. While noise-robust automatic speech recognition (ASR) has been an active field of research for years, with a considerable amount of well-elaborated techniques available [1], few studies so far dealt with the challenge of noise-robust AER, such as [2, 3]. Besides, at present the tools and particularly evaluation methodologies for noise-robust AER are rather basic: often, they are constrained to elementary feature enhancement and selection techniques [4, 5], are characterized by the simplification of additive stationary noise [6, 7], or are limited to matched condition training [8–11].

In contrast, this paper is a first attempt to evaluate the impact of nonstationary noise and different microphone conditions on the same realistic task as used in the INTERSPEECH 2009 Emotion Challenge [12]. For a thorough and complete evaluation, we implement typical methodologies from the ASR domain, such as commonly performed with the Aurora task of recognizing spelt digit sequences in noise [13]. On the other hand, the task is realistic because emotions were nonacted and nonprompted and do not belong to a prototypical, preselected set of emotions such as joy, fear, or sadness; instead, all data are used, including mixed and unclear cases (open microphone setting). We built our evaluation procedures for this study on the two-class problem defined for the Challenge, which is related to the recognition of negative emotion in speech. A system that performs robustly on this task in real-life conditions is useful for a variety of applications incorporating speech interfaces for human-machine communication, including human-robot interaction, dialog systems, voice command applications, and computer games. In particular, the Challenge task is based on the FAU Aibo Emotion Corpus which consists of recordings of children talking to the dog-like Aibo robot.

Another key part of this study is to exploit the signal decomposition (source separation) capabilities of Nonnegative Matrix Factorization (NMF) for noise-robustness, a technology which has led to considerable success in the ASR domain. The basic principle of NMF-based audio processing, as will be explained in detail in Section 2, is to find a locally optimal factorization of a spectrogram into two factors, of which the first one represents the spectra of the *acoustic events* occurring in the signal and the second one their *activation* over time. This factorization can be computed by iteratively minimizing cost functions resembling the perceptual quality of the product of the factors, compared with the original spectrogram. In this context, several studies have shown the advantages of NMF for speech denoising [14–16] as well as the related task of isolating speakers in a mixture ("cocktail party problem") [17–19]. While these approaches use NMF as a preprocessing method, recently another type of NMF technologies has been proposed that exploits the *structure* of the factorization: when initializing the first factor with values suited to the problem at hand, the activations (second factor) can be used as a dynamic feature which corresponds to the degree that a certain spectrum contributes to the observed signal at each time frame. This principle has been successfully introduced to ASR [20, 21] and the classification of acoustic events [22], particularly the detection of nonlinguistic vocalizations in speech [23]; yet it remains an open question whether it can be exploited within AER.

There do exist some recent studies on NMF features for emotion recognition from speech. In [24], NMF was proposed as an effective method to extract relevant spectral information from a signal by reducing the spectrogram to a single column, to which emotion classification can be applied; yet, this study lacks comparison to more conventional feature extraction methods. In [25], NMF as a *feature space reduction* method was reported being superior to related techniques such as Principal Components Analysis (PCA) in the context of AER. However, both these studies were carried out on clean speech with acted emotions; in contrast, our technique aims to augment NMF feature extraction in noisy conditions by making use of the intrinsic source separation capabilities of NMF. In this respect, it directly evolves from our previous research on robust ASR [20], where we proposed a "semisupervised" approach that detects spoken letters in noise by classifying the time-varying gains of corresponding spectra while simultaneously estimating the characteristics of the additive background noise. Transferring this paradigm to the emotion recognition domain, we propose to measure the amount of "emotional activation" in speech by NMF and show how this paradigm can improve state-of-the-art AER "in the wild".

The remainder of this paper is structured as follows. First, we introduce the mathematical background of NMF and its use in signal processing in Section 2. Second, we describe our feature extraction procedure based on NMF in Section 3. Third, we describe the data sets based on the INTERSPEECH 2009 Emotion Challenge task that we used for evaluation in Section 4 and show the results of our experiments on reverberated and noisy speech, including different microphone conditions, in Section 5 before concluding in Section 6.

2. Nonnegative Matrix Factorization

2.1. Definition. The mathematical specification of the NMF algorithm is as follows: given a matrix $\mathbf{V} \in \mathbb{R}_+^{m \times n}$ and a constant $r \in \mathbb{N}$, it computes two matrices $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$, such that

$$\mathbf{V} \approx \mathbf{WH}. \tag{1}$$

In case that $(m + n)r < mn$, NMF performs information reduction (*incomplete* factorization); otherwise, the factorization is called *overcomplete*. Incomplete and overcomplete factorizations require different algorithmic approaches [26]; we constrain ourselves to incomplete factorization in this study.

As a method of information reduction, it fundamentally differs from other methods such as PCA by using nonnegativity constraints: it does not merely aim at a mathematically optimal basis for describing the data, but at a decomposition into its actual parts. To this end, it finds a locally optimal representation where only additive—never subtractive—combinations of the parts are allowed. There is evidence that this type of decomposition corresponds to the human perception of images [27] and human language acquisition [28].

2.2. NMF-Based Signal Processing. NMF in signal processing is usually applied to spectrograms that are obtained by short-time Fourier transformation (STFT). Basic NMF approaches assume a linear signal model. Note that (1) can be written as follows (the subscripts $:, t$ and $:, j$ denote the tth and jth matrix columns, resp.):

$$\mathbf{V}_{:,t} \approx \sum_{j=1}^{r} \mathbf{H}_{j,t} \mathbf{W}_{:,j}, \quad 1 \leq t \leq n. \tag{2}$$

Thus, supposing $\mathbf{V}$ is the magnitude spectrogram of a signal (with short-time spectra in columns), the factorization from (1) represents each short-time spectrum $\mathbf{V}_{:,t}$ as a linear combination of spectral basis vectors $\mathbf{W}_{:,j}$ with nonnegative coefficients $\mathbf{H}_{j,t}$ $(1 \leq j \leq r)$. In particular, the ith row of the $\mathbf{H}$ matrix indicates the amount that the spectrum in the ith column of $\mathbf{W}$ contributes to the spectrogram of the original signal. This fact is the basis for our feature extraction approach, which will be explained in Section 3.

When there is no prior knowledge about the number of spectra that can describe the source signal, the number of components r has to be chosen empirically, depending on the application. As will be explained in Section 3, in the context of NMF feature extraction, this parameter also influences the number of features. The actual number of components used for our experiments will be described in Section 5 and was defined based on our previous experience with NMF-based source separation and feature extraction of speech and music [23, 29].

In concordance with recent NMF techniques for speech processing [17, 21], we apply NMF to Mel spectra instead of directly using magnitude spectra, in order to integrate a psychoacoustic measure and to reduce the computational complexity of the factorization. As common for feature extraction in speech and emotion recognition, the Mel filter bank had 26 bands and ranged from 0 to 8 kHz.

2.3. Factorization Algorithms. A factorization according to (1) is usually achieved by iterative minimization of a cost function c:

$$(\mathbf{W}, \mathbf{H}) = \arg\min_{\mathbf{W}', \mathbf{H}'} c(\mathbf{W}', \mathbf{H}'). \tag{3}$$

Several recent studies in NMF-based speech processing [15, 16, 18–20] use cost functions based on a modified version of Kullback-Leibler (KL) divergence such as

$$c_d(\mathbf{W}, \mathbf{H}) = \sum_{ij} \left(\mathbf{V}_{ij} \log \frac{\mathbf{V}_{ij}}{(\mathbf{WH})_{ij}} - (\mathbf{V} - \mathbf{WH})_{ij} \right). \tag{4}$$

Particularly, in our previous study on NMF feature extraction for detection of nonlinguistic vocalizations in speech [23], this function has been shown to be superior to a metric based on Euclidean distance, which matches the results of the comparative study carried out in [30].

For minimization of (4), we implemented the algorithm by Lee and Seung [31], which iteratively modifies $\mathbf{W}$ and $\mathbf{H}$ using "multiplicative update" rules. With matrix-matrix multiplication being its core operation, the computational cost of this algorithm largely depends on the matrix dimensions: assuming a naive implementation of matrix-matrix multiplication, the cost per iteration step is $O(mnr)$ for the minimization of c_d from (4). However, in practice, computation time can be drastically reduced by using optimized linear algebra routines.

As for any iterative algorithm, initialization and termination must be specified. While $\mathbf{H}$ is initialized randomly with the absolute values of Gaussian noise, for $\mathbf{W}$ we use an approach tailored to the problem at hand, which will be explained in detail later. As to termination, a convergence-based stopping criterion could be defined, measured in terms of the cost function [30, 32]; however, several previous studies, including [20, 21, 23, 29], proposed to run a fixed number of iterations. We used the latter approach for two reasons: first, from our experience, the error in terms of c_d that is left after a few hundred iterations is not significantly reduced by further iterations [29]. Second, for a signal processing system in real-life use, this does not only reduce the computational complexity—as the cost function does not have to be evaluated after each iteration—but also ensures a predictable response time. During the experiments carried out in this study, the number of iterations remained fixed at 200.

2.4. Context-Sensitive Signal Model. Various extensions to the basic linear signal model have been proposed to address a fundamental limitation. In (2), the acoustic events are characterized only by an instantaneous spectral observation, rather than a sequence; hence, NMF cannot exploit any context information which might be relevant to discriminate classes of acoustic events. In particular, an extension called Nonnegative Matrix Deconvolution (NMD) has been proposed [33, 34] where each acoustic event is modeled by a spectrogram of fixed length T and is obtained by a modified version of the NMF multiplicative update algorithm; however, this modification implies that variations of the original NMF algorithm—such as minimization of different types of cost functions—cannot immediately be transferred to the NMD case [32]. In this paper, we use an NMD-related approach [21] where the original spectrogram $\mathbf{V}$ is converted to a matrix $\mathbf{V}'$ such that every column of $\mathbf{V}'$ is the row-wise concatenation of a sequence of short-time spectra (in the form of row vectors). Mathematically speaking, given a sequence length T and the original spectrogram $\mathbf{V}$, we compute a modified matrix $\mathbf{V}'$ defined by

$$\mathbf{V}' := \begin{bmatrix} \mathbf{V}_{:,1} & \mathbf{V}_{:,2} & \cdots & \mathbf{V}_{:,n-T+1} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{V}_{:,T} & \mathbf{V}_{:,T+1} & \cdots & \mathbf{V}_{:,n} \end{bmatrix}. \tag{5}$$

That is, the columns of $\mathbf{V}'$ correspond to overlapping sequences of spectra in $\mathbf{V}$. This method reduces the problem of context-sensitive factorization of $\mathbf{V}$ to factorization of $\mathbf{V}'$; hence, it will allow our approach to be easily extended by using a variety of available NMF algorithms. In our experiments, the parameter T was set to 10.

3. NMF Feature Extraction

3.1. Supervised NMF. Considering (2) again, one can directly derive a concept for feature extraction: by keeping the columns of $\mathbf{W}$ constant during NMF, it seeks a minimal-error representation of the signal using a given set of spectra with nonnegative coefficients. In other words, the algorithm is given a set of acoustic events, described by (a sequence of) spectra, and its task is to find the activation pattern of these events in the signal. The activation patterns for each of the predefined acoustic events then yield a set of time-varying features that can be used for classification. This method will subsequently be called *supervised NMF*, and we call the resulting features "NMF activations".

This approach requires a set of acoustic events that are known to occur in the signals to be processed. However, it can be argued that this is generally the case for speech-related tasks: for instance, in our study on NMF-based spelling recognition [20], the events corresponded to spelt letters; in [21], spectral sequences of spelt digits were used. In the emotion recognition task at hand, they could consist of manifestations of certain emotions. Still, a key question that remains to be answered is how to compute the spectra that are used for initialization. For this study, we chose to follow a paradigm that led to considerable success in source separation [17, 34, 35] as well as NMF feature extraction [20, 23] tasks: here, NMF itself was used to reduce a set of training samples for each acoustic event to discriminate

into a set of characteristic spectra (or spectrograms). More precisely, our algorithm for initialization of supervised NMF builds a matrix $\mathbf{W}$ as follows, assuming that we aim to discriminate K different classes of acoustic events. For each class $k \in \{1, \ldots, K\}$,

(1) concatenate the corresponding training samples,

(2) compute the magnitude spectrogram $\mathbf{V}_k$ by STFT,

(3) from $\mathbf{V}_k$ obtain matrices $\mathbf{W}_k$, $\mathbf{H}_k$ by NMF.

Intuitively speaking, the columns of each $\mathbf{W}_k$ contain "characteristic" spectra of class k. As we are dealing with modified spectrograms (5), we will subsequently call the columns of $\mathbf{W}$ "characteristic sequence". More precisely, these are the observation sequences that model all of the training samples belonging to class k with the least overall error. From the $\mathbf{W}_k$ we build the matrix $\mathbf{W}$ by column-wise concatenation:

$$\mathbf{W} := [\mathbf{W}_1 \mathbf{W}_2 \cdots \mathbf{W}_K]. \tag{6}$$

3.2. Semisupervised NMF. If supervised NMF is applied to a signal that cannot be fully modeled with the given set of acoustic events—for instance, in the presence of background noise—the algorithm will produce erroneous activation features. Hence, in [20, 22] a *semisupervised* variant was proposed: here, the matrix $\mathbf{W}$ containing characteristic spectra is extended with additional columns that are randomly initialized. By updating only these columns during the iteration, the algorithm is "allowed" to model parts of the signal that cannot be explained using the predefined set of spectra. In particular, these parts can correspond to noise: in both the aforementioned studies, a significant gain in noise-robustness of the features could be obtained by using semisupervised NMF. Thus, we expect that semisupervised NMF features could also be beneficial for recognition of emotion in noise, especially for mismatched training and test conditions. As the feature extraction method can isolate (additive) noise, it is expected that the activation features are less degraded, and less dependent on the type of noise, than those obtained from supervised NMF, or more conventional spectral features such as MFCC. In contrast, it is not clear how semisupervised NMF features, and NMF features in general, behave in the case of reverberated signals; to our knowledge, this kind of robustness issue has not yet been explicitly investigated. We will deal with the performance of NMF features in reverberation as well as additive noise in Sections 5.3 and 5.4.

Finally, as semisupervised NMF can actually be used for arbitrary two-class signal separation problems, it could be useful for emotion recognition in clean conditions as well. In this context, one could initialize the $\mathbf{W}$ matrix with "emotionless" speech and use an additional random component. Then, it could be assumed that the activations of the random component are high if and only if there are signal parts that cannot be adequately modeled with nonemotional speech spectra. Thus, the additional component in semisupervised NMF would estimate the degree of emotional activation in the signal. We will derive and evaluate a feature extraction algorithm based on this idea in Section 5.2.

3.3. Processing of NMF Activations. Finally, a crucial issue is the postprocessing of the NMF activations. In this study, we constrain ourselves to static classification using segmentwise functionals of time-varying features, as the performance of static modeling is often reported as superior for emotions [36] and performs very well in classification of nonlinguistic vocalizations [37], particularly using NMF features [23]. In the latter study, the Euclidean length of each row of the activation matrix was taken as a functional. We extend this technique by adding first-order regression coefficients as well as other functionals of the NMF activations, exactly corresponding to those computed for the INTERSPEECH 2009 Emotion Challenge baseline (see Table 2), to ensure best comparability of results.

As to normalization of the NMF activations, in [23] the functionals were normalized to sum to unity. Also in [21], the columns of the "activation matrix" $\mathbf{H}$ were normalized to unity after factorization. Normalization was not an issue in [20], as the proposed discrete "maximum activation" feature is invariant to the scale of $\mathbf{H}$. In our preliminary experiments on NMF feature extraction for emotion recognition, we found it inappropriate to normalize the NMF activations, since the unnormalized matrices contain some sort of energy information which is usually considered very relevant for the emotion recognition task; furthermore, in fact an optimal normalization method for each type of functional would have to be determined. In contrast, we did normalize the initialized columns of $\mathbf{W}$, each corresponding to a characteristic sequence, such that their Euclidean length was scaled to unity, in order to prevent numerical problems.

For best transparency of our results, the NMF implementation available in our open-source NMF toolkit "openBliSSART" was used (which can be downloaded at http://openblissart.github.io/openBliSSART/). Functionals were computed using our openSMILE feature extractor [38, 39] that provided the official feature sets for the INTERSPEECH 2009 Emotion Challenge [12] and the INTERSPEECH 2010 Paralinguistic Challenge [40].

3.4. Relation to Information Reduction Methods. NMF has been proposed as an information reduction method in several studies on audio pattern recognition, including [24, 25, 41]. One of its advantages is that there are no requirements on the data distribution other than nonnegativity, unlike, for example, for PCA which assumes Gaussianity. On the other hand, nonnegativity is the only asserted property of the basis $\mathbf{W}$—in contrast to PCA or Independent Component Analysis (ICA).

Most importantly, our methodology of NMF feature extraction goes beyond previous approaches for information reduction, including those that use NMF. While it also gains a more compact representation from spectrograms, it does so by finding coefficients that minimize the error induced by the dimension reduction for each individual instance. This is a fundamental difference to, for example, the extraction of Audio Spectral Projection (ASP) features proposed in the MPEG-7 standard [41], where the spectral observations are simply projected onto a basis estimated

by some information reduction method, such as NMF or PCA. Furthermore, traditional information reduction methods such as PCA cannot be straightforwardly extended to semisupervised techniques that can estimate residual signal parts, as described in Section 3.2—this is a specialty of NMF due to its nonnegativity constraints which allow a part-based decomposition.

Laying aside these theoretical differences, it still is of practical interest to compare the performance of our supervised NMF feature extraction against a dimension reduction by PCA. We apply PCA on the extended Mel spectrogram $\mathbf{V}$ (5), as PCA on the logarithm of the Mel spectrogram would result in MFCC-like features which are already covered by the IS feature set. To rather obtain a feature set comparable to the NMF features, the same functionals of the according projections on this basis are taken as in Table 2. While the PCA basis could be estimated class-wisely, in analogy to NMF (6), we used all available training instances for the computation of the principal components, as this guarantees pairwisely uncorrelated features. We will present some key results obtained with PCA features in Section 5.

4. Data Sets

The experiments reported in this paper are based on the FAU Aibo Emotion Corpus and four of its variants.

4.1. FAU Aibo Emotion Corpus. The German FAU Aibo Emotion Corpus [42] with 8.9 hours of spontaneous, emotionally colored children's speech comprises recordings of 51 German children at the age of 10 to 13 years from two different schools. Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder. The sampling rate of the signals is 48 kHz; quantization is 16 bit. The data is downsampled to 16 kHz.

The children were given five different tasks where they had to direct Sony's dog-like robot Aibo to certain objects and through a given "parcours". The children were told that they could talk to Aibo the same way as to a real dog. However, Aibo was remote-controlled and followed a fixed, predetermined course of actions, which was independent of what the child was actually saying. At certain positions, Aibo disobeyed in order to elicit negative forms of emotions. The corpus is annotated by five human labelers on the word level using 11 emotion categories that have been chosen prior to the labeling process by iteratively inspecting the data. The units of analysis are not single words, but semantically and syntactically meaningful chunks, following the criteria defined and evaluated in [43] (18 216 chunks, 2.66 words per chunk on average, cf. [42]). Heuristic algorithms are used to map the decisions of the five human labelers on the word level onto a single emotion label for the whole chunk [42]. The emotional states that can be observed in the corpus are rather nonprototypical, emotion-related states than "pure" emotions. Mostly, they are characterized by low emotional intensity. Along the lines of the INTERSPEECH 2009 Emotion Challenge [12], the complete corpus is used for the experiments reported in this paper, that is, no balanced subsets were defined, no rare states and no ambiguous states are removed—all data had to be processed and classified (cf. [44]). The same 2-class problem with the two main classes *negative valence* (NEG) and the default state *idle* (IDL, i.e., neutral) is used as in the INTERSPEECH 2009 Emotion Challenge. A summary of this challenge is given in [45].

Table 1: Number of instances in the FAU Aibo Emotion Corpus. The partitioning corresponds to the INTERSPEECH 2009 Emotion Challenge, with the training set split into a training and development set ("devel").

(a) close-talk microphone (CT), additive noise (BA = babble, ST = street)

#	NEG	IDL	$\sum$
train	1 541	3 380	4 921
devel	1 817	3 221	5 038
test	2 465	5 792	8 257
$\sum$	5 823	12 393	**18 216**

(b) room microphone (RM), artificial reverberation (CTRV)

#	NEG	IDL	$\sum$
train	1 483	3 103	4 586
devel	1 741	2 863	4 604
test	2 418	5 468	7 886
$\sum$	5 642	11 434	**17 076**

As the children of one school were used for training and the children of the other school for testing, the partitions feature speaker independence, which is needed in most real-life settings, but can have a considerable impact on classification accuracy [46]. Furthermore, this partitioning provides realistic differences between the training and test data on the acoustic level due to the different room characteristics, which will be specified in the next section. Finally, it ensures that the classification process cannot adapt to sociolinguistic or other specific behavioral cues. Yet, a shortcoming of the partitioning originally used for the challenge is that there is no dedicated development set. As our feature extraction and classification methods involve a variety of parameters that can be tuned, we introduced a development set by a stratified speaker-independent division of the INTERSPEECH 2009 Emotion Challenge training set. To allow for easy reproducibility, we chose a straightforward partitioning into halves. That is, the first 13 of the 26 speakers (speaker IDs 01–08, 10, 11, 13, 14, and 16) were assigned to our training set, and the remaining 13 (speaker IDs 18–25, 27–29, 31, and 32) to the development set. This partitioning ensures that the original challenge conditions can be restored by jointly using the instances in the training and development sets for training.

Note that—as it is typical for realistic data—the two emotion classes are highly unbalanced. The number of instances for the 2-class problem is given in Table 1(a). This version, which also has been the one used for the INTERSPEECH 2009 Emotion Challenge, will be called "close-talk" (CT).

4.2. Realistic Noise and Reverberation. Furthermore, the whole experiment was filmed with a video camera for documentary purposes. The audio channel of the videos is reverberated and contains background noises, for example, the noise of Aibo's movements, since the microphone of the video camera is designed to record the whole scenery in the room. The child was not facing the microphone, and the camera was approximately 3 m away from the child. While the recordings for the training set took place in a normal, rather reverberant class room, the recording room for the test set was a recreation room, equipped with curtains and carpets, that is, with more favorable acoustic conditions. This version will be called "room microphone" (RM). The amount of data that is available in this version (17 076 chunks) is slightly less than in the close-talk version due to technical problems with the video camera that prevented a few scenes from being simultaneously recorded on video tape. See Table 1(b) for the distribution of instances in the RM version. To allow for comparability with the same choice of instances, we thus introduce the set CT_{RM}, which contains only those close-talk segments that are also available in the RM version, in addition to the full set CT.

4.3. Artificial Reverberation. The third version [47] of the corpus was created using artificial reverberation: the data of the close-talk version was convolved with 12 different impulse responses recorded in a different room using multiple speaker positions (four positions arranged equidistantly on one of three concentric circles with the radii $r \in \{60\,\text{cm}, 120\,\text{cm}, 240\,\text{cm}\}$) and alternating echo durations $T_{60} \in \{250\,\text{ms}, 400\,\text{ms}\}$ spanning 180°. The training, development, and test set of the CT_{RM} version were evenly split in twelve parts, of which each was reverberated with a different impulse response. The same impulse response was used for all chunks belonging to one turn. Thus, the distribution of the impulse responses among the instances in the training, development, and test set is roughly equal. This version will be called "close-talk reverberated" (CTRV).

4.4. Additive Nonstationary Noise. Finally, in order to create a corpus which simulates spontaneous emotions recorded by a close-talk microphone (e.g., a headset) in the presence of background noise, we overlaid the close-talk signals from the FAU Aibo Emotion Corpus with noises corresponding to those used for the Aurora database [13], which was designed to evaluate performance of noise-robust ASR. We chose the "Babble" (BA) and "Street" (ST) noise conditions, as these are nonstationary and frequently encountered in practical application scenarios. The very same procedure as in creating the Aurora database [13] was followed: first, we measured the speech activity in each chunk of the FAU Aibo Emotion Corpus by means of the algorithm proposed in the ITU-T P.56 recommendation [48], using the original software provided by the ITU. Then, each chunk was overlaid with a random noise segment whose gain was adjusted in such a way that the signal-to-noise ratio (SNR), in terms of the speech activity divided by the long-term (RMS) energy of the noise segment, was at a given level. We repeated this procedure for the SNR levels −5 dB, 0 dB, 5 dB, and 10 dB, similarly to the Aurora protocol.

In other words, the ratio of the perceived loudness of voice and noise is constant, which increases the realism of our database: since persons are supposed to speak louder once the level of background noise increases (Lombard effect), it would not be realistic to mix low-energy speech segments with a high level of background noise. This is of particular importance for the FAU Aibo Emotion Corpus, which is characterized by great variance in the speech levels. To avoid clipping in the audio files, the linear amplitude of both speech and noise was multiplied with 0.1 prior to mixing. Thus, for the experiments with additive noise, the volume of the clean database had to be adjusted accordingly. Note that at SNR levels of 0 dB or lower, the performance of conventional automatic speech recognition on the Aurora database decreases drastically [13]; furthermore, our previous study on emotion recognition in the presence of additive noise [11] indicates that an SNR of 0 dB poses a challenge even for recognition of *acted* emotions.

5. Results

The structure of this section is oriented on the different variants of the FAU Aibo Emotion Corpus as introduced in the last section—including the original INTERSPEECH 2009 Emotion Challenge setting.

5.1. Classification Parameters. As classifier, we used Support Vector Machines (SVM) with a linear kernel on normalized features, which showed better performance than standardized ones in a preliminary experiment on the development set. Models were trained using the Sequential Minimal Optimization (SMO) algorithm [49]. To cope with the unequal distribution of the IDL and NEG classes, we always applied the Synthetic Minority Oversampling Technique (SMOTE) [50] prior to classifier training, as in the Challenge baselines. For both oversampling and classification tasks, we used the implementations from the Weka toolkit [51], in line with our strategy to rely on open-source software to ensure the best possible reproducibility of our results, and utmost comparability with the Challenge results. Thereby parameters were kept at their defaults except for the kernel complexity parameter, as we are dealing with feature vectors of different dimensions and distributions. Hence, this parameter was fine-tuned on the development set for each training condition and type of feature set, with the results presented in the subsequent sections.

5.2. INTERSPEECH 2009 Emotion Challenge Task. In a first step, we evaluated the performance of NMF features on the INTERSPEECH 2009 Emotion Challenge task, which corresponds to the 2-class problem in the FAU Aibo Emotion Corpus (CT version) to differentiate between "idle" and "negative" emotions. As the two classes are highly unbalanced (cf. Table 1)—with over twice as much "idle" instances as "negative" ones—we consider it more appropriate to measure performance in terms of unweighted average recall

Table 2: INTERSPEECH 2009 Emotion Challenge feature set (IS): low-level descriptors (LLD) and functionals.

LLD (16 · 2)	Functionals (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, rel. position, range
(Δ) MFCC 1–12	linear regression: offset, slope, MSE

Table 3: Summary of NMF feature sets for the Aibo 2-class problem. # IDL: number of characteristic sequences from IDL training instances; # NEG: number of characteristic sequences from NEG instances; # free: number of randomly initialized components; Comp: indices of NMF components whose functionals are taken as features; Dim: dimensionality of feature vectors. For N30/31-1, no "free" component is used for training instances of clean speech. As explained in the text, the $N31_I$ set is not considered for the experiments on additive noise.

Name	# IDL	# NEG	# free	Comp	Dim
$N31_I$	30	0	1	1–31	744
N30	15	15	0	1–30	720
N31	15	15	1	1–31	744
N30/31-1	15	15	0/1	1–30	720
N31-1	15	15	1	1–30	720

(UAR) than weighted average recall (WAR). Furthermore, UAR was the metric chosen for evaluating the Challenge results.

As a first baseline feature set, we used the one from the classifier subchallenge [12], which is shown in Table 2. Next, as NMF features are essentially spectral features with a different basis, we also compared them against Mel spectra and MFCCs, to investigate whether the choice of "characteristic sequences" as basis, instead of frequency bands, is superior.

Based on the algorithmic approaches laid out in Section 3, we applied two variants of NMF feature extraction, whereby factorization was applied to Mel spectrograms (26 bands) obtained from STFT spectra that were computed by applying Hamming windows of 25 ms length at 10 ms frame shift. First, semisupervised NMF was used, based on the idea that one could initialize the algorithm with manifestations of "idle" emotions and then estimate the degree of negative emotions in an additional, randomly initialized component. Thus, in contrast to the application of semisupervised NMF in noise-robust speech recognition [20], where the activations of the randomly initialized component are ignored in feature extraction, in our case we consider them being relevant for classification. 30 characteristic sequences of idle emotions were computed from the INTERSPEECH 2009 Emotion Challenge training set according to the algorithm from Section 3.1, whereby a random subset of approximately 10% (in terms of signal length) was selected to cope with memory requirements for the factorization, as in [17, 23]. including functionals, is denoted by "$N31_I$" (cf. Table 3).

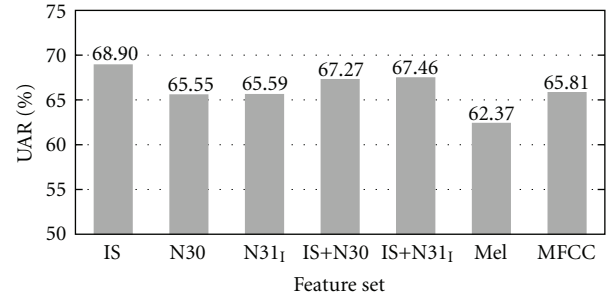


Figure 1: Results on the INTERSPEECH 2009 Emotion Challenge task (FAU Aibo 2-class problem, close-talk speech = CT). "UAR" denotes unweighted average recall. "IS" is the baseline feature set from the challenge; "N30" and "$N31_I$" are supervised and unsupervised NMF features (cf. Table 3); "+" denotes the union of feature sets. "Mel" are functionals of 26 Mel frequency bands and "MFCC" functionals of the corresponding MFCCs (1–12). Classification was performed by SVM (trained with SMO, complexity $C = 0.1$).

As another method, we used supervised NMF, that is, without a randomly initialized component, and predefining characteristic spectrograms of negative emotion as well, which were computed from the NEG instances in the INTERSPEECH 2009 Emotion Challenge training set (again, a random subset of about 20% was selected). In order to have a feature set with comparable dimension, 15 components per class (IDL, NEG) were used for supervised NMF, yielding the feature set "N30" (Table 3).

As an alternative method of (fully) supervised NMF that could be investigated, one could compute characteristic sequences from all available training data, instead of restricting the estimation to class-specific matrices. While this is an interesting question for further research, we did not consider this alternative due to several reasons: first, processing all training data in a single factorization would result in even larger space complexity, which is, speaking of today, already an issue for the classwise estimation (see above). Second, our N30 feature set contains the same amount of discriminative features for each class, while the training set itself is unbalanced (cf. Table 1). Finally, while it could theoretically occur that the same, or very similar, characteristic sequences are computed for both classes, and thus redundant features would be obtained, we found that this was not a problem in practice, as in the extracted features no correlation could be observed, neither within the features corresponding to the IDL or NEG classes, nor in the NMF feature space as a whole. Note that in NMF feature extraction using a cost function that purely measures reconstruction error, such as (4), statistical properties of the resulting features can never be guaranteed.

Results can be seen in Figure 1. NMF features clearly outperformed "plain" Mel spectra and deliver a comparable UAR in comparison to MFCCs. Still, it turned out that they could not outperform the INTERSPEECH 2009 feature set; even a combination of the NMF and IS features (IS+N30, IS+$N31_I$) could not yield a performance gain over the baseline. Considering the performance of different variants of NMF,

no significant differences can be seen according to a one-tailed t-test ($P > 0.05$), which will be the test we refer to in the subsequent discussion. Note that the baseline in Figure 1 is higher than the one originally presented for the challenge [12], due to the SMO complexity parameter being lowered from 1.0 to 0.1.

To complement our extensive experiments with NMF, we further investigated information reduction by PCA. To that end, PCA features were extracted using the first 30 principal components of the extended spectrograms of the training set as transformation, as described in Section 3.4, and computing functionals of the transformed extended spectrograms of the test set. This type of features will be referred to as "P30", in analogy to "N30", in all subsequent discussions. However, the observed UAR of 65.33% falls clearly below the baseline features, and also below both types of NMF features considered. Still, as the latter difference is not significant ($P > 0.05$), we further considered PCA features for our experiments on reverberation and noise, as will be pointed out in the next sections.

5.3. Emotion Recognition in Reverberated Speech. Next, we evaluated the feature extraction methods proposed in the last section on the reverberated speech from the FAU Aibo Emotion Corpus (RM and CTRV versions). The same initialization as for the NMF feature extraction on CT speech was used, thus the NMF feature sets for the different versions are "compatible".

Our evaluation methodologies are inspired by techniques in the noise-robust ASR domain, taking into account *matched condition*, *mismatched condition*, and *multicondition* training. Similar procedures are commonly performed with the Aurora database [13] and were also partly used in our previous study on noise-robust NMF features for ASR [20].

In particular, we first consider a classifier that was trained on CT_{RM} speech only and evaluate it across the three test conditions available (CT_{RM}, RM, and CTRV). Next, we join the training instances from all three conditions and evaluate the same three test conditions (multicondition training). Lastly, we also consider the case of "noise-corrupted" models, that is, classifiers that were, respectively, trained on RM and CTRV data. Note that for the multicondition training, upsampling by SMOTE was applied prior to joining the data sets, to make sure that each combination of class and noise type is equally represented in the training material. Thereby we optimized the complexity parameter C for the SMO algorithm on the development set to better take into account the varying size and distribution of feature vectors depending on (the combination of) features investigated. In Figure 2, we show the mean UAR over all test conditions on the development set, depending on the value of C for each of the different training conditions. Different parameter values of $C \in \{10^{-3}, 2 \cdot 10^{-3}, 5 \cdot 10^{-3}, 10^{-2}, 2 \cdot 10^{-2}, 5 \cdot 10^{-2}, 10^{-1}, 0.2, 0.5, 1\}$ were considered. The general trend is that on one hand, the optimal parameter seems to depend strongly on the training condition and feature set; however, on the other hand, it turned out that N30 and N31 can be treated with similar complexities, as can IS + N30 and IS+N31. Thus, we exemplarily show the IS, N31, and IS+N31 feature sets in the graphs in Figure 2 and leave out N30.

Table 4: Results on the Aibo 2-class problem (7 886 test instances in each of the CT_{RM}, RM, and CTRV versions) for different training conditions. All results are obtained with SVM trained by SMO with complexity parameter C, which was optimized on the development set (see Figure 2). "UAR" denotes unweighted average recall. "IS" is the baseline feature set (INTERSPEECH 2009 Emotion Challenge) while "N30" and "N31_{I}" are NMF features obtained using supervised and semisupervised NMF (see Table 3). "+" denotes the union of feature sets. "Mean" is the arithmetic mean over the three test conditions. The best result per column is highlighted.

(a) Training with close-talk microphone (CT_{RM})

UAR [%]	C	CT_{RM}	RM	CTRV	Mean
IS	1.0	**67.62**	**60.51**	**53.06**	**60.40**
N30	1.0	65.48	52.36	50.23	56.02
N31_{I}	1.0	65.54	53.10	50.36	56.33
IS + N30	0.5	67.37	49.15	51.62	56.05
IS + N31_{I}	1.0	67.15	56.47	51.95	58.52

(b) Multicondition training ($CT_{RM} + RM + CTRV$)

UAR [%]	C	CT_{RM}	RM	CTRV	Mean
IS	0.01	**67.72**	59.52	66.06	64.43
N30	0.05	66.73	**67.55**	52.66	62.31
N31_{I}	0.2	65.81	64.61	63.32	64.58
IS + N30	0.005	67.64	62.64	**66.78**	**65.69**
IS + N31_{I}	0.005	67.07	61.85	65.92	64.95

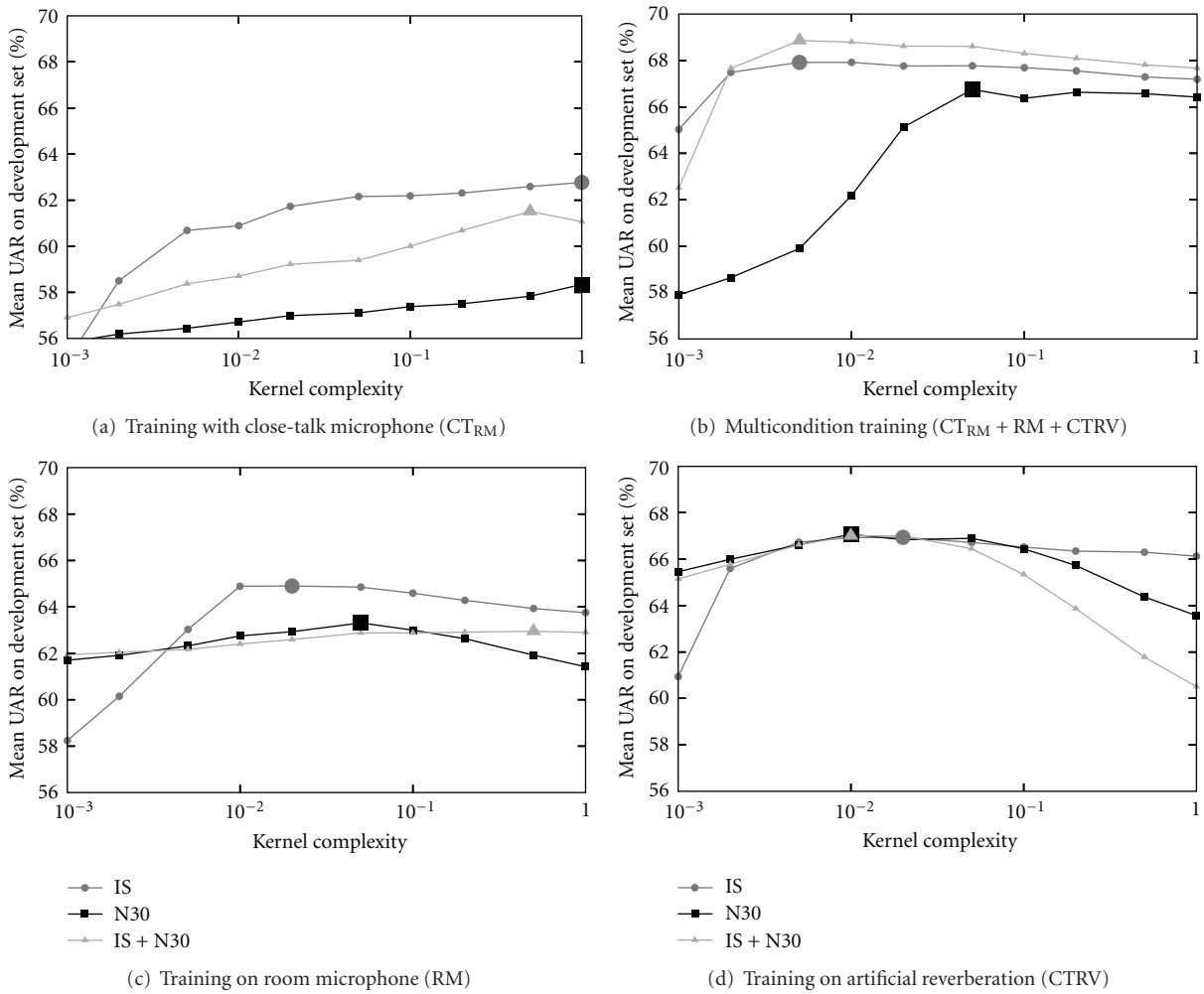
(c) Training on room microphone (RM)

UAR [%]	C	CT_{RM}	RM	CTRV	Mean
IS	0.02	61.61	62.72	**62.10**	62.14
N30	0.2	53.57	65.61	54.87	58.02
N31_{I}	0.5	54.50	**66.54**	56.20	59.08
IS + N30	0.05	**65.13**	66.26	60.39	**63.93**
IS + N31_{I}	0.05	64.68	66.34	59.54	63.52

(d) Training on artificial reverberation (CTRV)

UAR [%]	C	CT_{RM}	RM	CTRV	Mean
IS	0.02	60.64	59.29	66.35	62.09
N30	0.05	60.73	**68.19**	62.72	**63.88**
N31_{I}	0.02	60.94	64.40	64.30	63.21
IS + N30	0.01	**61.70**	49.17	**66.68**	59.18
IS + N31_{I}	0.02	61.61	63.03	66.56	63.73

After obtaining an optimized value of C for each training condition, we joined the training and development sets and used these values for the experiments on the CT_{RM}, RM, and CTRV versions of the test set; the results are given in Table 4. First, it has to be stated that NMF features can outperform the baseline feature set in a variety of scenarios involving room-microphone (RM) data. In particular, we obtain a significant ($P < 0.001$) gain of almost 4% absolute for matched condition training, from 62.72% to 66.54% UAR. Furthermore, a multicondition trained classifier using



(a) Training with close-talk microphone (CT_{RM})

(b) Multicondition training (CT_{RM} + RM + CTRV)

(c) Training on room microphone (RM)

(d) Training on artificial reverberation (CTRV)

Figure 2: Optimization of the SMO kernel complexity parameter C on the mean unweighted average recall (UAR) on the development set of the FAU Aibo Emotion Corpus across the CT_{RM}, RM, and CTRV conditions. For the experiments on the test set (Table 4), the value of C that achieved the best performance on average over all test conditions (CT_{RM}, RM, and CTRV) was selected (depicted by larger symbols). The graphs for the $N31_I$ and IS + $N31_I$ sets are not shown for the sake of clarity, as their shape is roughly similar to N30 and IS + N30.

the N30 feature set outperforms the baseline by 8% absolute; in the case of a classifier trained on CTRV data, the improvement by using N30 instead of IS features is even higher (9% absolute, from 59.29% to 68.19%). On the other side, NMF features seem to lack robustness against the more diverse reverberation conditions in the CTRV data, which generally results in decreased performance when testing on CTRV, especially for the mismatched condition cases. Still, the difference on average across all test conditions for multicondition trained classifiers with IS + N30 (65.69% UAR), respectively, IS features (64.43% UAR) is significant ($P < 0.002$). Considering semisupervised versus fully supervised NMF, there is no clear picture, but the tendency is that the semisupervised NMF features ($N31_I$) are more stable. For example, consider the following unexpected result with the N30 features: in the case of training with CTRV and testing with RM, N30 alone is observed 9% absolute *above* the baseline, yet its combination with IS falls 10% *below* the baseline.

As the multicondition training case has proven most promising for dealing with reverberation, we investigated the performance of P30 features in this scenario. On average over the three test conditions, the UAR is 62.67%; thus comparable with supervised NMF (N30, 62.31%), but significantly ($P < 0.001$) below semisupervised NMF ($N31_I$, 64.58%). Thereby the complexity was set to $C = 1.0$, which had yielded the best mean UAR on the development set. In turn, P30 features suffer from the same degradation of performance when CT training data is used in mismatched test conditions: in that case, the mean UAR is 56.17% (again, at the optimum of $C = 1.0$), which does not differ significantly ($P > 0.05$) from the result achieved by either type of NMF features (56.02% for N30, 56.33% for $N31_I$).

5.4. Emotion Recognition in Noisy Speech. The settings for our experiments on emotion recognition in noisy speech correspond to those used in the previous section—with the disturbances now being formed by purely additive noise,

not involving reverberation. Note that the clean speech and multicondition training scenarios now exactly match the "Aurora methodology" (test set A from [13]). Additionally, we consider mismatched training with noisy data as in our previous study [20] or the test case "B" from the Aurora database [13]. In correspondence with Aurora, all SNR levels from −5 dB to 10 dB were considered as testing condition, while the −5 dB level was excluded from training. Thus, the multicondition training, as well as training with BA or ST noise, involves the union of training data corresponding to the SNR levels 0 dB, 5 dB, and 10 dB.

As in the previous sections, the baseline is defined by the IS feature set. For NMF feature extraction, we used semisupervised NMF with 30 predefined plus one uninitialized component, but this time with a different notion: now, the additional component is supposed to model primarily the additive noise, as observed advantageous in [20]. Hence, both the idle and negative emotions should be represented in the preinitialized components, with 15 characteristic spectrograms for each—the "N31" feature set is now used instead of $N31_I$ (cf. Table 3).

It is desirable to compare these semisupervised NMF features with the procedure proposed in [20]. In that study, supervised NMF was applied to the clean data, and semisupervised NMF to the noisy data, which could be done because neither multicondition training was followed nor were models trained on clean data tested in noisy conditions, due to restrictions of the proposed classifier architecture. However, for a classifier in real-life use, this method is mostly not feasible as the noise conditions are usually unknown. On the other hand, using *semi*supervised NMF feature extraction both on clean and noisy signals, the following must be taken into account: when applied to clean speech, the additional component is expected to be filled with speech that cannot be modeled by the predefined spectra; however, it is supposed to contain mostly noise once NMF is applied to noisy speech. Thus, it is not clear how to best handle the activations of the uninitialized component in such a way that the features in the training and test sets remain "compatible", that is, that they carry the same information: we have to introduce and evaluate different solutions, as presented in Table 3.

In detail, we considered the following three strategies for feature extraction. First, the activations of the uninitialized component can be ignored, resulting in the "N31-1" feature set; second, we can take them into account ("N31"). A third feature set, subsequently denoted by "N30/31-1", finally provides the desired link to our approach introduced in [20]: here, the activations for the clean training data were computed using fully supervised NMF; in contrast, the activations for the clean and noisy test data, as well as the noisy training data, were computed using semisupervised NMF with a noise component (without including its activations in the feature set).

Given that the noise types considered are nonstationary, one could think of further increasing the number of uninitialized components for a more appropriate signal modeling. Yet, we expect that this would lead to more and more speech being modeled by the noise components, which is a known drawback of NMF—due to the spectral overlap between noise and speech—if no further constraints are imposed on the factorization [15, 16]. Hence, an undesired amount of randomness would be introduced to the information contained in the features.

We experimented with all three of the N31, N31-1, and N30/31-1 sets, and their union with the IS baseline feature set. First, Table 5(a) shows the recognition performance for the clean training case. The result is twofold: on the one hand, for both cases of noise they outperform the baseline, particularly in the case of babble noise, where the mean UAR across the SNR levels is 60.79% for IS and 63.80% for N31-1. While this effect is lower for street noise, all types of NMF features outperform the IS baseline on average over all testing conditions. The difference in the mean UAR achieved by N31-1 (63.75%) compared with the IS (62.34%) is significant with $P < 0.001$. On the other hand, for neither of the NMF feature sets could a significant improvement be obtained by combining them with the baseline feature set; still, the union of IS and N31-1 exhibits the best overall performance (63.99% UAR). This, however, comes at a price: comparing N31 to IS for the clean test condition, a performance loss of about 5% absolute from 68.47% to 63.65% UAR has to be accepted, which can only partly be compensated by joining N31 with IS (65.63%). In summary, the NMF features lag considerably behind in the clean testing case (note that the drop in performance compared to Figure 1 is probably due to the different type of Semisupervised NMF as well as the complexity parameter being optimized on the mean).

A counterintuitive result in Table 5(a) deserves some further investigation: while the UAR obtained by the IS features gradually decreases when going from the clean case (68.47%) to babble noise at 10, 5, and 0 dB SNR (57.71% for the latter), it considerably increases for −5 dB SNR (64.52%). Still, this can be explained by examining the confusion matrices, as shown in Table 6. Here, one can see that at decreasing SNR levels, the classifier more and more tends to favor the IDL class, which results in lower UAR; this effect is however reversed for −5 dB, where more instances are classified as NEG. This might be due to the energy features contained in IS; generally, higher energy is considered to be typical for negative emotion. In fact, preliminary experiments indicate that when using the IS set without the energy features, the UAR increases monotonically with the SNR but is significantly below the one achieved with the full IS set, being at chance level for −5 dB (BA and ST) and at 66.31% for clean (CT) testing. The aforementioned unexpected effect also occurs—in a subdued way—for the NMF features, which, as explained before, also contain energy information. As a final note, when considering the WAR, that is, the accuracy instead of the UAR, as usually reported in studies on noise-robust ASR where balancing is not an issue, there is no unexpected drop in performance from −5 to 0 dB for the BA testing condition: indeed, the WAR is 69.44% at −5 dB and 71.41% at 0 dB, respectively. For the ST testing condition, the WAR drops below chance level (49.22%) for −5 dB, then monotonically raises to 62.44, 69.70, and 70.58% at increased SNRs of 0, 5, and 10 dB.

Table 5: Results on the Aibo 2-class problem with additive noise (8 257 test instances) for different training conditions. The following test conditions were considered: CT (clean), BA (babble noise at −5–10 dB SNR), and ST (street noise at −5–10 dB SNR). All results are obtained with SVM trained by SMO with complexity parameter C, which was optimized on the development set (see Figure 3). "UAR" denotes unweighted average recall. "IS" is the baseline feature set (INTERSPEECH 2009 Emotion Challenge); NMF features ("N31" etc.) were obtained using supervised and semisupervised NMF (see Table 3). "+" denotes the union of feature sets. "Mean" is the arithmetic mean over the nine test conditions for Tables 5(a) and 5(b), and the mean over all SNRs for Tables 5(c) and 5(d). Note that the "N30/31-1" set differs from "N31-1" only in the case that clean speech occurs in the training material. The best result per column is highlighted. Note that the UAR does not uniformly increase with SNR, as could be expected—this is partly due to the imbalanced test set, as explained in the text.

(a) Clean training (CT)

UAR [%]	C	CT	BA				ST				Mean
			−5 dB	0 dB	5 dB	10 dB	−5 dB	0 dB	5 dB	10 dB	
IS	0.2	**68.47**	64.52	57.71	57.73	63.20	60.60	64.19	62.47	62.20	62.34
N31-1	0.001	62.85	65.79	**64.06**	**62.84**	62.49	63.82	64.94	**63.77**	63.18	63.75
N30/31-1	0.002	62.23	65.64	63.18	61.71	61.90	**64.11**	64.63	63.21	62.37	63.22
N31	0.001	63.65	65.78	63.25	62.24	63.03	64.01	64.77	62.90	62.68	63.59
IS + N31-1	0.002	65.24	**65.93**	63.13	62.45	63.46	63.00	**65.39**	63.75	**63.53**	63.99
IS + N30/31-1	0.005	64.20	65.22	61.51	60.95	61.85	63.51	65.11	62.23	61.93	62.95
IS + N31	0.002	65.63	65.73	62.32	61.74	**63.90**	63.20	65.33	63.00	63.25	**63.79**

(b) Multicondition training (CT + BA + ST)

UAR [%]	C	CT	BA				ST				Mean
			−5 dB	0 dB	5 dB	10 dB	−5 dB	0 dB	5 dB	10 dB	
IS	0.2	**66.96**	65.78	66.36	66.60	**66.57**	64.79	65.87	65.58	65.59	66.01
N31-1	0.5	64.36	66.11	66.25	65.61	65.49	65.27	65.64	65.64	65.86	65.58
N30/31-1	1.0	63.40	66.17	66.30	65.65	65.29	65.19	65.61	65.35	65.59	65.39
N31	0.2	65.37	**66.48**	65.86	66.04	65.93	65.72	65.50	65.72	65.76	65.82
IS + N31-1	0.02	66.61	66.00	**66.43**	**66.69**	**66.57**	65.60	**66.48**	**66.48**	66.22	**66.34**
IS + N30/31-1	0.02	66.28	66.10	66.51	**66.69**	66.42	65.58	66.52	66.41	66.06	66.29
IS + N31	0.05	66.69	66.13	66.02	66.66	66.52	**65.75**	66.38	66.07	**66.27**	66.28

(c) Training on babble noise (BA)

UAR [%]	C	CT	BA				Mean	ST				Mean
			−5 dB	0 dB	5 dB	10 dB		−5 dB	0 dB	5 dB	10 dB	
IS	1.0	62.17	66.15	66.04	66.16	65.62	65.99	61.26	65.57	66.05	64.95	64.46
N31-1	0.2	62.95	66.38	65.88	65.20	65.03	65.62	**65.37**	65.58	65.20	64.94	65.27
N31	0.5	**65.81**	66.59	66.35	66.07	65.96	66.24	64.54	65.70	65.85	65.89	65.50
IS + N31-1	0.02	63.32	67.16	**67.26**	66.48	65.99	66.72	61.82	**66.56**	**67.20**	**66.78**	**65.59**
IS + N31	0.02	64.38	**67.57**	67.22	**66.95**	**66.37**	**67.03**	61.55	66.53	67.17	66.47	65.43

(d) Training on street noise (ST)

UAR [%]	C	CT	BA				Mean	ST				Mean
			−5 dB	0 dB	5 dB	10 dB		−5 dB	0 dB	5 dB	10 dB	
IS	1.0	61.33	62.20	63.03	63.61	62.22	62.77	65.15	65.44	65.67	65.20	65.37
N31-1	0.5	62.84	**65.40**	**65.61**	65.33	64.78	**65.28**	65.56	65.55	65.00	65.48	65.40
N31	0.2	**65.55**	64.91	64.78	**65.69**	**65.71**	65.27	65.56	65.74	65.06	66.03	65.60
IS + N31-1	0.2	61.51	64.02	64.50	64.86	64.09	64.37	**66.00**	66.02	66.18	**66.28**	66.12
IS + N31	0.1	63.43	63.60	64.14	65.07	64.94	64.44	65.95	**66.29**	**66.34**	66.16	**66.19**

Next, Table 5(b) evaluates multicondition training with the aforementioned feature sets. Again, the union of IS and N31-1 shows the best mean UAR (66.34%), but the gain with respect to the IS baseline (66.01%) is not significant; however, the aforementioned performance loss in the clean test condition is avoided. As is expected, the mean UAR for multicondition training is higher than for clean training, which is true for all feature sets, and with the IS + N30/31-1 feature set profiting the most (over 3% absolute on average). From both Tables 5(a) and 5(b), one can see that the "N30/31-1" feature set inspired by [20] is inferior to the other two kinds of semisupervised NMF

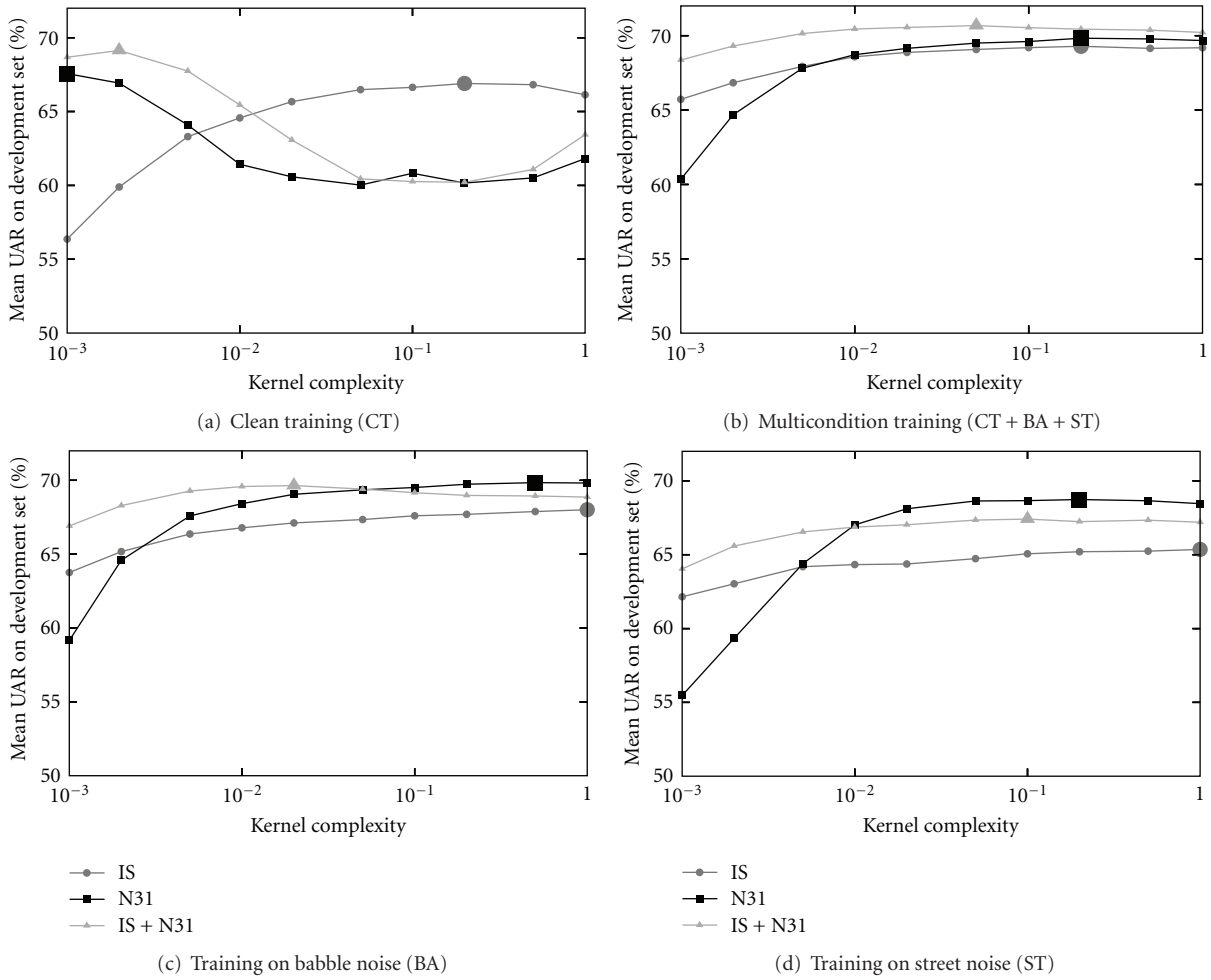


Figure 3: Optimization of the SMO kernel complexity parameter C on the mean unweighted average recall (UAR) on the development set of the FAU Aibo Emotion Corpus across the CT, BA, and ST test conditions, including all available SNRs from −5 to 10 dB. For the experiments on the test set (Table 5), the value of C that achieved the best performance on average over all test conditions (CT, BA, and ST) was selected (depicted by larger symbols). The graphs for the N30 and IS + N30 sets are not shown for the sake of clarity, as their shape is roughly similar to N31 and IS + N31.

Table 6: Confusion matrices on the Aibo 2-class problem with additive noise (8 257 test instances) for clean training, using the IS feature set and an SVM classifier with complexity parameter $C = 0.2$, as in Table 5(a). The following test conditions were considered: CT (clean), BA (babble noise at −5–10 dB SNR), and ST (street noise at −5–10 dB SNR).

#	CT		BA								ST							
			−5 dB		0 dB		5 dB		10 dB		−5 dB		0 dB		5 dB		10 dB	
	IDL	NEG	IDL	NEG	IDL	NEG	IDL	NEG	IDL	NEG	IDL	NEG	IDL	NEG	IDL	NEG	IDL	NEG
IDL	4 195	1 597	4 445	1 347	5 311	481	5 228	564	4 718	1 074	1 874	3 918	3 467	2 325	4 657	1 135	4 808	984
NEG	875	1 590	1 176	1 289	1 880	585	1 844	621	1 357	1 108	275	2 190	776	1 689	1 367	1 098	1 445	1 020

feature sets; the difference between IS + N30/31-1 and IS + N31-1 for the clean training case is even significant with $P < 0.01$.

Finally, Tables 5(c) and 5(d) evaluate training on noisy data, with matched and mismatched test condition. In this context, it is especially notable that the NMF features outperform the IS feature set for the clean test condition, with N31 (65.81%) being more than 3% absolute over the baseline (62.17%) for BA training and over 4% for ST training (N31: 65.55%; IS: 61.33%). Both these differences are significant with $P < 0.001$.

Additionally, comparing mismatched and matched *noisy* test conditions in Tables 5(c) and 5(d), one can see that the improvement by NMF features is generally higher for mismatched conditions, providing evidence for the claim from Section 3.2. Particularly, in the case of ST training and testing on BA, we observed a gain of 2.5% absolute over the baseline (62.77%) by both the N31-1 (65.28%) and N31 (65.27%)

feature sets, on average over the four SNRs. Still, both these sets also improve the results for matched condition training (by almost 2% absolute, from 63.76% to 65.60% for N31). On the other hand, in the case of BA-matched condition training, a significant gain is only obtained by combining NMF with IS; yet, this feature set provides the overall best mean UAR on babble noise (67.03%) among all training conditions. Again, for mismatched condition (testing on ST), there is an improvement of about 1.0% absolute comparing N31 (65.50%) to IS (64.46%).

To complement the discussion of our results, we conducted several experiments using the P30 features to deal with additive noise. As in the last section, we considered multicondition training, since overall, this scenario yielded the most stable results across all testing conditions considered. Again, it turned out that all three types of NMF features were superior to P30, which was evaluated at a complexity parameter of $C = 0.02$ that was found to be optimal on the development set and yielded a mean UAR of 65.08% across all nine testing conditions. Finally, in an experiment with BA training using the P30 features, we found that both in the mismatched test conditions (CT, ST) and in matched condition, the P30 features fell clearly behind NMF features: on average over the BA respectively ST conditions, the mean UAR (65.21%/64.88%) is significantly ($P < 0.05$) below the performance of N31 (66.24%/65.50%), and also falls behind N31-1 (65.62%/65.27%). While in clean testing, P30 (with an UAR of 65.31%) can outperform N31-1 (62.84%), it is still is slightly below N31 (65.55%), which gives the overall best result for BA training and clean testing.

5.5. Discussion. Summarizing the results from both Tables 4 and 5, it can be seen that especially the $N31_I$ and N31 feature sets are promising for robust emotion recognition: while they are sometimes inferior to other NMF features, in almost all cases, they increase the performance when added to the baseline feature set, and in some cases, they even outperform the baseline alone. The latter observation is particularly remarkable when taking into account that NMF features are computed by a purely heuristic algorithm on spectral information, while the baseline was specifically engineered for emotion recognition.

While in case of multicondition training on realistic noise and reverberation, a significant gain could be obtained by adding NMF features to the baseline, this was not true for multicondition training on *additive* noise. Still, in a scenario where the classifier was trained on one (additive) noise type and tested in clean and other noisy conditions, NMF features have led to significantly better performance than the baseline; hence, it will be an interesting topic for future research to evaluate NMF features in multicondition training with mismatched noises—such as in the "Aurora" test case "B" [13]. In summary, we conclude that in application scenarios "in the wild", the information contained in the NMF features seems to complement traditional AER features considerably well.

In contrast, for clean testing conditions, that is, in the absence of noise and reverberation, including the original INTERSPEECH 2009 Challenge task, we could not achieve a performance improvement over the baseline by NMF features. It is actually a frequently encountered phenomenon that while methods tailored to noise-robust speech processing, such as NMF, are valuable for deteriorated signals, they result in slightly lower performance on clean signals; similar conclusions have been recently drawn in, for example, [21].

Concerning the different notions of supervised NMF for AER that we proposed, no clear tendency can be observed when comparing the $N31_I$ feature set which is supposed to measure the degree of negative emotion in the random component of semisupervised NMF, with the supervised NMF feature set N30. Hence, we conclude that both approaches are valid and should be considered in further research.

Finally, when comparing the various solutions to extract features from noisy speech by Semisupervised NMF, including our previous approach [20], it is notable that ignoring the activation of the noise component in classification (as done for N31-1) is not necessarily the best choice, as could be assumed in the first place. In fact, the additional features in N31 considerably increase mean UAR over all test conditions for BA as well as ST training, while they do not contribute to robustness for clean and multicondition training. The best result for both CT and multicondition training is, however, achieved by the union of IS and N31-1. Notably, the feature set N30/31-1 corresponding to our previous approach [20] lags considerably behind the other types of NMF features in the case of clean testing, both for clean and multicondition training: it is 1.4% and 2.0% below N31, respectively, which is significant with $P < 0.05$.

6. Conclusion

The experiments dealt with in this paper were motivated by the considerable mismatch between the ASR and AER—or the linguistic and paralinguistic—domains, regarding the techniques and evaluation methodologies for enhanced robustness. Hence, we did not only present our results in a manner that resembles the well-known Aurora training and test scenarios, but also integrated NMF as a novel noise-robust signal processing method. Further, in contrast to many current studies that perform subject dependent percentage splits or cross-validations, we strictly enforced speaker independence. Finally, we focused on exact reproducibility by relying on open-source software for all major steps in the feature extraction and classification procedure, and most importantly by using clearly defined training, development, and test sets based on publicly available corpora. In fact, a deficiency that shows in a number of studies is that they do not explicitly mention the parts of data used for optimizing parameters. On the other hand, classifier parameters tend to have great influence on the recognition rates, as we have clearly demonstrated in this paper.

From our experimental results, we conclude that the overall performance of NMF features is remarkable, especially compared to our previous study on NMF in the paralinguistic domain [23], where performance of NMF

features themselves was observed considerably below the MFCC baseline; in this paper, NMF features were often observed on par with the well-tuned INTERSPEECH 2009 Emotion Challenge feature set. Yet, the most noticeable tendency that we find in our results is that a gain by NMF can be obtained exactly in the most realistic conditions: that is, in the presence of realistic noise and reverberation, and to some extent in the (simulated) presence of babble or street noise. Note that while it is a common phenomenon in noise-robust ASR that performance monotonically increases with SNR, and this type of behavior could be reproduced for AER in [10], other studies, such as [9] suggest that this might not always be the case. Given the fact that there is still a lack of comprehensive studies on noise-robust AER, this issue may be worth further investigation in the future.

Caution must be exercised when comparing recognition rates on spontaneous, nonprototypical emotions, as those reported in this paper, to the ones from other studies on AER, which are typically carried out on corpora of acted emotions. While much research work has been invested into tuning performance on the INTERSPEECH 2009 Emotion Challenge task, the best result in terms of UAR still remains at 71.2% for the two-class problem, which is obtained by fusing the individual classification engines of the challenge participants [45]. This clearly indicates that the "open-microphone" setting in which the FAU Aibo Emotion Corpus was generated still poses a hard challenge to today's AER systems, yielding recognition rates that are considerably lower than it could be expected for a two-class problem consisting of acted emotions.

On the other hand, the promising results concerning robustness that were reported in this paper motivate a lot of further research in the domain. First, we might consider overcomplete NMF that is initialized with a large set of spectral sequences that correspond to different emotion classes—inspired by the "exemplar-based" recognition architecture introduced in [21], which delivered excellent results in noise-robust ASR, and which particularly marks a departure from the information reduction paradigm found in traditional NMF approaches. Second, a novel technique could perform adaptation to noise on the NMF feature extraction level by measuring the activations of spectra from different noise conditions. Concerning evaluation, we have not yet adopted the various feature enhancement techniques developed in years of ASR research, such as Histogram Equalization or Switching Models (cf. [1]), which could be beneficial both for conventional as well as NMF features. Further, evaluation of noise-robust techniques should be carried out taking into account a greater variety of noise conditions—a task that we are now ready to address after having defined the basic methodologies. In this context, we will also strive at a more detailed investigation of the proposed feature extraction approach in comparison to more traditional information reduction methods, building on the preliminary experiments with PCA reported in this paper, and further including ICA.

Finally, we are confident that our paradigms can be extended to other fields of the paralinguistic domain. Hence, we will consider further application scenarios for NMF feature extraction, for instance, the INTERSPEECH 2010 Paralinguistic Challenge [40] task to recognize the level of interest in spontaneous speech.

Acknowledgments

This work was partly funded by the Federal Republic of Germany through the German Research Foundation (DFG) under the Grant no. SCHU 2508/2-1 ("Nonnegative Matrix Factorization for Robust Feature Extraction in Speech Processing"), the European Union in the projects PF-STAR under Grant IST-2001-37599, and HUMAINE under Grant IST-2002-50742. The responsibility lies with the authors.

References

[1] B. Schuller, M. Wllmer, T. Moosmayr, and G. Rigoll, "Recognition of noisy speech: a comparative survey of robust model architecture and feature enhancement," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, Article ID 942617, 2009.

[2] E.-H. Kim, K.-H. Hyun, and Y.-K. Kwak, "Robustemotion recognition feature, frequency range of meaningful signal," in *Procceings of the IEEE International Workshopon Robots and Human Interactive Communication(RO-MAN '05)*, Nashville, Tenn, USA, 2005.

[3] A. Tawari and M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *Proceedings of the International Conference on Pattern Recognition (ICPR '10)*, pp. 4605–4608, Istanbul, Turkey, August 2010.

[4] K.-K. Lee, Y.-H. Cho, and K.-S. Park, "Robust feature extraction for mobile-based speech emotion recognition system," in *Intelligent Computing in Signal Processing and Pattern Recognition*, Lecture Notes in Control and Information Sciences, pp. 470–477, Springer, Berlin, Germany, 2006.

[5] W.-J. Yoon, Y.-H. Cho, and K.-S. Park, "A study of speech emotion recognition and its application tomobile services," in *Ubiquitous Intelligence and Computing*, Lecture Notes In Computer Science, pp. 758–766, Springer, Berlin, Germany, 2007.

[6] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Manifolds based emotion recognition in speech," *Computational Linguistics and Chinese Language Processing*, vol. 12, no. 1, pp. 49–64, 2007.

[7] M. Grimm, K. Kroschel, H. Harris et al., "On the necessity and feasibility of detecting a driver's emotional state while driving," in *Affective Computing and Intelligent Interaction*, A. Paiva, R. Prada, and R. W. Picard, Eds., pp. 126–138, Springer, Berlin, Germany, 2007.

[8] M. Lugger, B. Yang, and W. Wokurek, "Robust estimation of voice quality parameters under real world disturbances," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 1, pp. 1097–1100, Toulouse, France, 2006.

[9] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," in *Procceings of the IEEE International Conference on Multimedia and Expo (ICME '06)*, pp. 1653–1656, Toronto, Canada, July 2006.

[10] B. Schuller, D. Arsić, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *Proceedings of the Speech Prosody*, Dresden,Germany, 2006.

[11] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech,"

in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 4, pp. 941–944, Honolulu, Hawaii, USA, April 2007.

[12] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Proceedings of the INTERSPEECH of Conference of the International Speech Communication Association (ISCA '09)*, pp. 312–315, Brighton, UK, September 2009.

[13] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '00)*, Beijing, China, October 2000.

[14] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Efficient model-based speech separation and denoising using non-negative subspace analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 1833–1836, Las Vegas, Nev, USA, 2008.

[15] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proceedings of the INTERSPEECH*, Brisbane, Australia, 2008.

[16] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 4029–4032, Las Vegas, Nev, USA, April 2008.

[17] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proceedings of the INTERSPEECH of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, pp. 2614–2617, September 2006.

[18] T. Virtanen and A. T. Cemgil, "Mixtures of gamma priorsfor non-negative matrix factorization based speech separation," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA '09)*, pp. 646–653, Paraty, Brazil, 2009.

[19] T. Virtanen, "Spectral covariance in prior distributions of non-negative matrix factorization based speech separation," in *Proceedings of the European Signal Processing Conference (EUSIPCO '09)*, Glasgow, Scotland, 2009.

[20] B. Schuller, F. Weninger, M. Wöllmer, Y. Sun, and G. Rigoll, "Non-negative matrix factorization as noiserobust feature extractor for speech recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal (ICASSP '10)*, pp. 4562–4565, Dallas, Tex, USA, March 2010.

[21] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal (ICASSP '10)*, Dallas, Tex, USA, March 2010.

[22] Y.-C. Cho, S. Choi, and S.-Y. Bang, "Non-negative component parts of sound for classification," in *Proceedings of the International Symposium on Signal Processing and Information Technology (ISSPIT '03)*, pp. 633–636, Darmstadt, Germany, 2003.

[23] B. Schuller and F. Weninger, "Discrimination of speech and non-linguistic vocalizations by non-negative matrix factorization," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '10)*, pp. 5054–5057, Dallas, Tex, USA, March 2010.

[24] K. Jeong, J. Song, and H. Jeong, "NMF features for speech emotion recognition," in *Proceedings of the International Conference on Hybrid Information Technology (ICHIT '09)*, pp. 368–374, ACM, New York, NY, USA, 2009.

[25] D. Kim, S.-Y. Lee, and S.-I. Amari, "Representative and discriminant feature extraction based on NMF foremotion recognition in speech," in *Proceedings of the of the 16th International Conference on Neural Information Processing (ICONIP '09)*, pp. 649–656, Springer, Berlin, Germany, 2009.

[26] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proceedings of the Neural Networks*, vol. 4, pp. 2529–2533, Dalian, China, 2004.

[27] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[28] L. ten Bosch, J. Driesen, H. van hamme, and L. Boves, "On a computational model for language acquisition:Modeling cross-speaker generalisation," in *Text, Speech and Dialogue*, V. Matoušek and P. Mautner, Eds., vol. 5729 of *Lecture Notes in Computer Science*, pp. 315–322, Springer, Berlin, Germany, 2009.

[29] B. Schuller, A. Lehmann, F. Weninger, F. Eyben, and G. Rigoll, "Blind enhancement of the rhythmic and harmonic sections by NMF: does it help?" in *Proceedings of the International Conference on Acoustics (NAG/DAGA '09)*, pp. 361–364, DEGA, Rotterdam, The Netherlands, 2009.

[30] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[31] D. D. Lee and H. S. Seung, "Algorithms for nonnegative matrix factorization," in *Proceedings of the Neural Information Processing Systems (NIPS '01)*, pp. 556–562, Vancouver,Canada, 2001.

[32] W. Wang, A. Cichocki, and J. A. Chambers, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared euclidean distance," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2858–2864, 2009.

[33] P. Smaragdis, "Discovering auditory objects through non-negativity constraints," in *Proceedings of the Workshop on Statistical and Perceptual Audition (SAPA '04)*, Jeju, Republic of Korea, 2004.

[34] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[35] P. D. O'Grady and B. A. Pearlmutter, "Discovering convolutive speech phones using sparseness and non-negativity," in *Proceedings of the International Workshop on Independent Component Analysis (ICA '07)*, London, UK, 2007.

[36] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing," in *Affective Computing and Intelligent Interaction*, A. Paiva, R. Prada, and R. W. Picard, Eds., pp. 139–147, Springer, Berlin, Germany, 2007.

[37] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Proceedings of the 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems (PIT '08)*, pp. 99–110, Springer, Berlin, Germany, 2008.

[38] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit," in *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII '09)*, Amsterdam, The Netherlands, September 2009.

[39] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE—the Munich versatile and fast open-source audio feature

extractor," in *Proceedings of the ACM International Conference on Multimedia (MM '10)*, pp. 1459–1462, ACM, Florence, Italy, October 2010.

[40] B. Schuller, S. Steidl, A. Batliner et al., "The INTERSPEECH 2010 paralinguistic challenge," in *Proceedings of the INTERSPEECH of Conference of the International Speech Communication Association (ISCA '10)*, pp. 2794–2797, Makuhari, Japan, September 2010.

[41] H.-G. Kim, J. J. Burred, and T. Sikora, "How efficientis MPEG-7 for general sound recognition?" in *Proceedings of the International Conference Convention of the Audio Engineering Society (AES '04)*, London, UK, June 2004.

[42] S. Steidl, *Automatic classification of emotion-related user states in spontaneous children's speech*, Ph.D. thesis, Logos, Berlin, Germany, 2009, FAU Erlangen-Nuremberg.

[43] A. Batliner, R. Kompe, A. Kießling, M. Mast, H. Niemann, and E. Nöth, "M = Syntax + Prosody: a syntactic-prosodic labelling scheme for large spontaneous speech databases," *Speech Communication*, vol. 25, no. 4, pp. 193–222, 1998.

[44] S. Steidl, B. Schuller, A. Batliner, and D. Seppi, "The hinterland of emotions: facing the open-microphone challenge," in *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII '09)*, pp. 690–697, IEEE, Amsterdam, The Netherlands, September 2009.

[45] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect inspeech: state of the art and lessons learnt from the first challenge," to appear in *Speech Communication, Special Issue on "Sensing Emotion and Affect—Facing Realism in Speech Processing"*.

[46] V. Sethu, E. Ambikairajah, and J. Epps, "Speaker dependency of spectral features and speech production cues for automatic emotion classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07)*, pp. 4693–4696, Taipei, Taiwan, 2009.

[47] A. Maier, C. Hacker, S. Steidl, E. Nöth, and H. Niemann, "Robust parallel speech recognition in multiple energy bands," in *Proceedings of the Annual Symposium of the German Association for Pattern Recognition (DAGM '05)*, W. G. Kropatsch, R. Sablatnig, and A. Hanbury, Eds., Lecture Notes in Computer Science, pp. 133–140, Vienna, Austria, August 2005.

[48] *ITU-T Recommendation P.56: Objective Measurementof Active Speech Level*, International Telecommunications Union (ITU), Geneva, Switzerland, 1993.

[49] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advancesin Kernel Methods: Support Vector Learning*, pp. 185–208, MIT Press, Cambridge, Mass, USA, 1999.

[50] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[51] I. H. Witten and E. Frank, *Data mining: Practical machine Learning Tools and Techniques*, Morgan Kauffmann, San Francisco, Calif, USA, 2nd edition, 2005.