The Effect of Narrow-Band Transmission on Recognition of Paralinguistic Information From Human Vocalizations

SASCHA FRÜHHOLZ[1,2,3], ERIK MARCHI[4,5,6], (Member, IEEE), AND BJÖRN SCHULLER[5,6,7], (Senior Member, IEEE)

[1]Department of Psychology, University of Zurich, Zürich 8006, Switzerland
[2]Neuroscience Center Zurich, University of Zurich, Zürich 8006, Switzerland
[3]Zurich Center for Integrative Human Physiology, University of Zurich, Zürich 8006, Switzerland
[4]Machine Intelligence and Signal Processing Group, Institute for Human-Machine Communication, Technische Universität München, Munich 80333, Germany
[5]Complex and Intelligent Systems, University of Passau, Passau 94032, Germany
[6]audEERING GmbH, Gilching 82205, Germany
[7]Department of Computing, Imperial College London, London SW7 2AZ, U.K.

Corresponding author: E. Marchi (erik.marchi@tum.de)

This work was supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013) under Grant 338164 (ERC Starting Grant iHEARu), Grant (RIA ARIA VALUSPA), and Grant 688835 (RIA DE-ENIGMA) and in part by the Swiss National Science Foundation under Grant SNSF PP00P1 1574091.

ABSTRACT Practically, no knowledge exists on the effects of speech coding and recognition for narrow-band transmission of speech signals within certain frequency ranges especially in relation to the recognition of paralinguistic cues in speech. We thus investigated the impact of narrow-band standard speech coders on the machine-based classification of affective vocalizations and clinical vocal recordings. In addition, we analyzed the effect of speech low-pass filtering by a set of different cut-off frequencies, either chosen as static values in the 0.5–5-kHz range or given dynamically by different upper limits from the first five speech formants (F1–F5). Speech coding and recognition were tested, first, according to short-term speaker states by using affective vocalizations as given by the Geneva Multimodal Emotion Portrayals. Second, in relation to long-term speaker traits, we tested vocal recording from clinical populations involving speech impairments as found in the Child Pathological Speech Database. We employ a large acoustic feature space derived from the Interspeech Computational Paralinguistics Challenge. Besides analysis of the sheer corruption outcome, we analyzed the potential of matched and multicondition training as opposed to miss-matched condition. In the results, first, multicondition and matched-condition training significantly increase performances as opposed to mismatched condition. Second, downgrades in classification accuracy occur, however, only at comparably severe levels of low-pass filtering. The downgrades especially appear for multi-categorical rather than for binary decisions. These can be dealt with reasonably by the alluded strategies.

INDEX TERMS Speech analysis, speech coding, emotion recognition, computational paralinguistics.

I. INTRODUCTION

Although data links are increasing in bandwidth and are becoming faster, there is a growing need for bandwidth conservation for wireless cellular and satellite communication networks. Voice-related applications, which are now integrated in the majority of mobile devices, require that the speech signal is processed, stored, or transmitted in the most costless and efficient manner. Speech Coding is the art of creating a minimally redundant representation of the speech signal that can be efficiently transmitted or stored in digital media [1]. The effects of speech coding in machine-based speech and speaker recognition systems have been broadly investigated over the last two decades. In Automatic Speech Recognition (ASR), several studies report the effects of different speech coders[1] on speech recognition performances [2]–[5]. In particular, given the interest in translating speech recognition technologies to mobile

[1]A speech coder converts a digitized speech signal into a coded representation [1].

environments, the recognition of coded speech transmitted over wireless channels in network-based speech recognition (NSR) and distributed speech recognition (DSR) systems is reported [6]–[8]. Further analyses have also been made on the influence of the sampling rate and bit rate changes with different narrowband and wideband codecs on the speech sounds [9], [10]. The effects of speech compression algorithms has been widely investigated also on speaker recognition systems. In fact, Dunn and colleagues [11] employed standard speech coders (GSM, G.729, G.723, MELP) in order to evaluate speaker recognition performances under matched and mismatched training and testing conditions.[2] A detailed analysis of speech coding techniques and bit rate effects on compressed-domain automatic speaker recognition is reported by Petracca and colleagues [12]. Moreover, the influence of speech coding on features dedicated to speaker identification such as formant frequencies and fundamental frequency (F0) trajectories is reported in previous studies [13], [14].

Thus, it is clear that for Automatic Speech Recognition and Automatic Speaker Recognition a plethora of studies have been carried out with respect to speech coding. Until now, however, no study has directly investigated the effects of speech coding for narrow-band transmission of human vocalizations with respect to the recognition of paralinguistic cues which seem a highly important issue, as in the last few year Computational Paralinguistic tasks such as automatic affect recognition from speech, are increasingly gaining commercial attention due to the rapid growing interest in social intelligence [15]–[17] and multi-modal user profiling [18]–[20], e. g., for interactive speech systems [21], [22], or serious games [23]–[25] and fun applications, such as the love detector by Nemesysco Ltd.[3] Many applications also exist in the public health sector. Hearing-impaired persons can profit, because cochlear implant processors typically alter the spectral cues, which are crucial for the perception of paralinguistic information [26]. Children with autism may profit from the analysis of emotional cues as they may have difficulties understanding or displaying them [27]–[32]. Further, health related applications could be the monitoring of elderly people living on their own [33], or could be related to diseases and speech disorders [34] such as occuring in Parkinson's disease [35], or cancer, cleft lip, and palate [36].

Affect recognition from speech thus seems not only highly relevant for techniques concerning digital transmission and specific health-related applications, but they also concern social communication and sharing of social data between human agents that are mediated by these techniques. This digital communication and social information exchange concerns, for example, wearable wireless devices [37] as well as mobile social networks [38], [39]. The latter aims at exchanging and integrating relevant information, especially for the integrated use of heterogeneous networks with diverse cell sizes and radio access technologies [40]. A specific application of such digital communication is healthcare communication, for example for patients living in remote areas or for elderly patients with limited access to primary health institutions [41], [42]. A major issue with conventional remote computer-assisted healthcare was the lack of the dimensions of emotional care next to a basic medico-physical care. New developments of wearable computer-assisted emotional interactions [37] for the emotional care of patients might improve disease outcome to a significant degree [43], especially if real-time healthcare communication is applied [44]. Thus, developing techniques for the appropriate and efficient transmission and decoding of emotional cues from the acoustic pattern of speech and vocalizations in digital communication is a timely issue.

Human speech and vocalizations have a specific pattern of acoustic features with a strong harmonic and formant structure and high energy power in low frequency bands [45]. This acoustic pattern is evident for human voices in general, but specifically for human speech and especially for speech melody as the most important paralinguistic cue. Speech melody is considerably influenced by the affective states of the speaker and provides powerful paralinguistic cues from which listeners infer the emotional states of other individuals. Concerning the power of spectral frequencies, the aforementioned high power in low frequency ranges is especially evident for neutral human vocalizations. Computing the alpha level of voices as the ratio between the power in high frequency bands (>1 kHz) and the power in low frequency bands (<1 kHz) usually reveals low alpha levels for neutral voices [46]. However, emotional compared with neutral vocalizations usually show an increased alpha level [46] or an increased proportion of high frequency power [47], [48], indicating that high energy in upper frequency bands is a reasonable cue for emotional voices. Especially, angry and happy voices show an increased alpha level compared with neutral voices [46], [48], while sad voices usually have a decreased alpha level [47]. These high frequency cues might therefore be an important voice feature, which might serve the discrimination of different vocal expressions by human listeners [49].

Apart from an increased power in high frequency bands, human speech in general and affective vocalizations in specific are characterized by a strong harmonic and formant structure [45]. The human voice is mostly characterized by its fundamental frequency (F0) and the first five formant frequency bands (F1–F5). The latter result from resonances in the oral cavity. Recently, we have shown that certain neural subregions in the cortical auditory system are actually sensitive to these acoustic features in vocal expressions, especially to the F0, which mainly contributes to the pitch perception in voices [50]. This decoding of vocal features from affective vocalizations is presumably accomplished by the orchestrated functioning of brain network involving the auditory cortex in the temporal lobe, the frontal lobes [51]

[2]Matched condition refers to use only data with exactly the same conditions for training and testing. Multi-condition uses a variety of different condition to generate a more generalised model.

[3]http://www.nemesysco.com/

and subcortical structures such as the amygdala [52]. Voice detection restricted to the F0–F1 should be sufficient to detect and classify human voices [53] and the F1–F2 seem to be sufficient to classify speech vowels [54]. Emotional voices are sometimes strongly detected and discriminated from the third formant (F3; see [55]). Voices consist of temporal formant dynamics evolving over time, and not all of them [56] and only those with a limited modulation rate [57] might be important to detect human voices and speech.

However, though high energy in upper frequency bands as well as the formant structure and their temporal dynamics might be valuable cues for the detection and recognition of affective vocalizations, many different vocal expressions actually share the same paralinguistic cues. Listeners often use these same cues for detecting several vocal emotions, such as high power in upper frequency bands for expressions of anger and desperation, or high power in lower frequency bands for the expression of sadness and elation [49]. Furthermore, the amplitude level between the F1 and the F2 can help to differentiate some, but not all affective vocalizations [46]. These shared cues between vocal emotions can impair their discrimination and might lead to a perceived confusing between vocalizations [58]. Thus, it seems necessary to exactly determine the different roles of the relative power in high and low frequency bands as well as of the speech formant structure for the classification of different affective vocalizations.

Besides these more transient and short-term emotional states, which can considerably influence speech and paralinguistic clues during affective vocalizations and which provide important cues for classifying the emotional states of the speaker, human speech and vocalizations are also influenced by more long-term speaker traits, such as neurological [59] or psychiatric disorders [60], [61]. Specifically, speech impairments in children with pervasive developmental disorders have a strong impact on perceiving as well as expressing speech in conversational contexts. In relation to the expression of vocalizations, recent studies for example have shown that children, which were diagnosed with a disorder from the autistic spectrum, show an impaired expression of paralinguistic cues in speech [61], [62]. Furthermore, children with language disorders, such as specific language impairments (SLI) or dyslexia, show an impaired expression of paralinguistic cues in speech [63]. Thus, these impairments in expressing paralinguistic cues in speech might be one indicator for the discrimination of 'normal' expression of vocalidizations from impaired expressions in clinical disorders related to the transmission of speech signals in certain frequency bands. An automated and machine-based analysis and classification of these vocalizations in normally developing children and in children with developmental disorders might help to obtain objective measures of diagnosing specific developmental conditions [64] based on vocal analysis on a large set of vocal features.

In this light, we investigated the impact of narrow-band standard speech coders (such as G.711, G.726, G.728, GSM, G.723.1, LPC10, and codec2) on the machine based classification of affective vocalization and clinical vocal recordings. Additionally we analysed the effect of low-pass filtering of human vocalizations recorded both in healthy and in clinical populations. We used a set of different cut-off frequencies for low-pass filtering. These are either chosen as static values in the 0.5 – 5 kHz range or given dynamically by different upper formant limits from F1 to F5. For the purpose of machine-based classifications of affective vocalizations as well as for the classification of impaired vocalizations in clinical populations we trained a computer-based classifier on two large-scale databases of vocal recordings. For short-term and transient emotional speaker states we used recordings of affective vocalizations as given by the Geneva Multimodal Emotion Portrayals (GEMEP) [65]. For long-term speaker traits related to clinical relevant disorders we used vocal recordings as found in the Child Pathological Speech Database (CPSD) [66].

The remainder of this contribution is structured as follows: First, a description of the databases and methods is given in Section II. Then, the experiments and results are described in Section III, before discussing and drawing conclusions on the evaluation of obtained results in Section IV.

II. MATERIAL AND METHODS

A. DATABASES

For the purpose of this study we used two different large-scale databases of human vocalizations. One database was used for the purpose of classifying affective vocalizations according to the emotional valence, the arousal and the category of the vocalizations. The second database included recordings from clinical populations in children.

1) THE GEneva MULTIMODAL EMOTION PORTRAYALS (GEMEP)

The GEMEP database [65] contains 1.2 k instances of emotional speech from ten professional actors (five female) in 18 categories. The database contains prompted emotional speech of an 'artificial language' to ensure emotionally neutral semantics of underlying speech, comprising sustained vowel phonations, as well as two meaningless phrases with two different intended sentence modalities ("ne kal ibam soud molen!" = phrase #1, "koun se mina lod belam?" = phrase #2), each expressed by each actor according to different emotional qualities (emotional valence) and in various degrees of regulation (emotional intensity or arousal) ranging from 'high' to 'masked' (hiding the true emotion).

Given this layout, a partitioning that is both text and speaker independent is not feasible. Hence, the following strategy was followed. Vowels and phrase #2 are used for training and development, subdividing by speaker ID, and phrase #1 is used for testing. Masked regulation utterances are only contained in the test set in order to alleviate potential model distortions. By this partitioning, one obtains text independence. Since six of the 18 emotional categories are

extremely sparse ($\leq$ 30 instances in the entire GEMEP database), the twelve most frequent ones are used in a multi-class classification task according to the emotional valence (referred to as 'categorization task'). Besides this multi-class valence classification we also performed a binary 'valence task' as well as an 'arousal task'. For the later two tasks, mappings are only defined for selected categories such as to obtain a balanced distribution of positive / negative arousal and valence among the categories. The resulting partitioning is shown in Table 1. As meta data, speaker IDs, prompts, and intended regulation are provided for the training and development sets. For the two dimensions arousal and valence (binary tasks) we provide results training on these binary targets and mapping from the twelve categories after classification.

TABLE 1. Partitioning of the GEMEP database into train(ing), dev(elopment), and test sets for 12-way classification by emotion category, and binary classification by pos(itive)/neg(ative) arousal (A) and valence (V). $^+$: Mapped to 'other' and excluded from evaluation in 12-class task. *: Mapped to 'undefined' and excluded from evaluation in binary tasks.

#	train	dev	test	A	V	Σ
admiration$^+$	20	2	8	pos	pos	30
amusement	40	20	30	pos	pos	90
anxiety	40	20	30	neg	neg	90
cold anger	42	12	36	neg	neg	90
contempt$^+$	20	6	4	neg	neg	30
despair	40	20	30	pos	neg	90
disgust$^+$	20	2	8	–*	–*	30
elation	40	12	38	pos	pos	90
hot anger	40	20	30	pos	neg	90
interest	40	20	30	neg	pos	90
panic fear	40	12	38	pos	neg	90
pleasure	40	20	30	neg	pos	90
pride	40	12	38	pos	pos	90
relief	40	12	38	neg	pos	90
sadness	40	12	38	neg	neg	90
shame$^+$	20	2	8	pos	neg	30
surprise$^+$	20	6	4	–*	–*	30
tenderness$^+$	20	6	4	neg	pos	30
Σ	602	216	442			1 260

2) THE CHILD PATHOLOGICAL SPEECH DATABASE (CPSD)

The CPSD database [66] provides speech recorded in two university departments of child and adolescent psychiatry, located in Paris, France (Université Pierre et Marie Curie/ Pitié-Salpêtière Hospital and Université René Descartes/ Necker Hospital). The dataset used here contains 2.5 k instances of speech recordings from 99 children aged 6 to 18 years. 35 of these children show Pervasive Development Disorders either of the autistic spectrum (PDD, 10 male, 2 female), specific language impairment such as dysphasia (DYS, 10 male, 3 female) or PDD Non-Otherwise Specified (NOS, 9 male, 1 female) according to the criteria of the Diagnostic and Statistical Manual of Psychiatric Disorders, Version 4 (DSM-IV). A monolingual healthy control group consisted of 64 further children (TYP, 52 male, 12 female). The French speech includes prompted sentence imitation of 26 sentences representing different modalities (declarative, exclamatory, interrogative, and imperative) and four types of intonations (descending, falling, floating, and rising).

Two evaluation tasks have been performed here: a binary 'typicality task' (typically vs. atypically developing children), and a four-way 'diagnosis task' (classifying into the above named categories: PDD, DYS, NOS, TYP). Partitioning into training, development and test data is done by order of speaker ID, stratified by age and gender of the children, and speaker-independently. The class distribution is given in Table 2.

TABLE 2. Partitioning of the child pathological speech database into train(ing), dev(elopment), and test sets for four-way classification by diagnosis, and binary classification by typical/atypical development. Diagnosis classes: TYPically developing, pervasive developmental disorders (PDD), pervasive developmental disorders non-otherwise specified (NOS), and specific language impairment such as DYSphasia.

#	train	dev	test	Σ
Typically developing				
TYP	566	543	542	1651
Atypically developing				
PDD	104	104	99	307
NOS	104	68	75	247
DYS	129	104	104	337
Σ	903	819	820	2542

B. PROCESSING OF VOCAL RECORDINGS

In this study we processed vocal recordings with standard speech coders and with two different low pass filtering techniques.

1) SPEECH CODERS

The speech coders tested cover a wide range of bit rates from 64 to 1.2 kbit/s, including:

- 64 kbit/s ITU G.711 mu-law PCM
- 40, 32, 24 kbit/s ITU G.726 ADPCM
- 16 kbit/s ITU G.728 LD-CELP
- 13 kbit/s ETSI GSM-FR RPE-LTP
- 6.3, 5.3 kbit/s ITU G.723.1 MPLPC, CELP
- 2.4 kbit/s DDVPC FS1015 LPC-10e
- 1.2 kbit/s open source codec2

The first two of these are traditional speech waveform coders: 64 kbit/s G.711 Pulse-coded Modulation (PCM) and 40, 32, 24 kbit/s G.726 Adaptive Differential PCM (ADPCM). These are toll-quality standards defined by the International Telecommunications Union (ITU) and widely deployed throughout the conventional telephone network. The next three coders are based on Code Excited Linear Prediction (CELP), a more sophisticated speech-specific waveform coding technology providing near-toll quality at medium bit rates. One of these are widely used in the cellular telephony systems: 13 kbit/s GSM-FR. The G.728 finds application in teleconferencing. The lower rates 6.3, 5.3 kbit/s G.723.1 CELP coder is used in VoIP applications. The 2.5 kbit/s LPC-10 is used in secure communication, whereas the lowest rate 1.2 kbit/s codec2 is an open source codec for speech over HF/VHF digital radio. Table 3 shows the different objective measures per speech coder. We can observe that

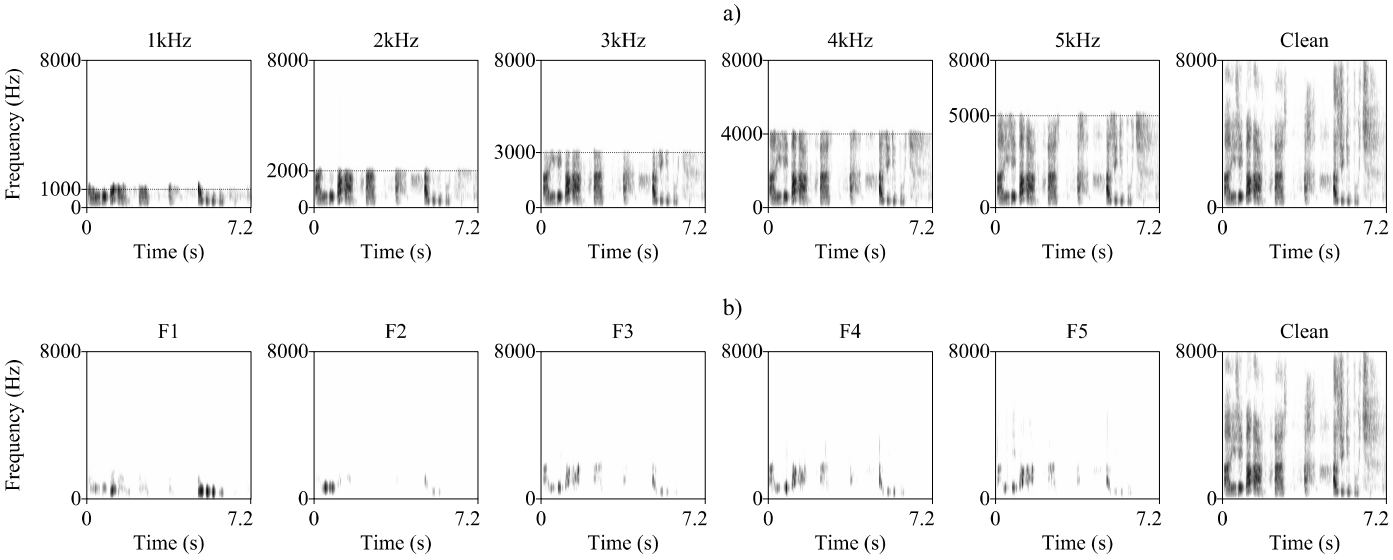

FIGURE 1. **Spectrograms of an excerpt from the CPSD database. Shown are: a) static low-pass filtering at different cut-off frequencies (from left to right: 1, 2, 3, 4, 5 kHz) and b) dynamic low-pass filtering at different upper formants (from left to right: F1, F2, F3, F4, F5).**

TABLE 3. **Objective measures averaged over the GEMEP and CPSD datasets per each speech codec. Segmental signal to noise ratio (SSNR), Itakura-Saito distance (IS), and perceptual evaluation of speech quality (PESQ).**

$codec_{[kbit/s]}$	GEMEP			CPSD		
	SSNR	IS	PESQ	SSNR	IS	PESQ
$G711_{64}$	29.4	0.10	4.3	34.8	0.01	4.4
$G726_{40}$	27.1	0.11	4.2	30.7	0.02	4.3
$G726_{32}$	24.6	0.14	4.0	26.7	0.04	4.2
$G726_{24}$	19.4	0.33	3.6	21.2	0.12	3.9
$G728_{16}$	15.6	0.21	4.0	16.2	0.21	4.0
GSM_{13}	10.2	0.40	3.4	11.6	0.40	3.4
$G7231_{6.3}$	-2.2	1.44	3.4	-2.5	1.58	3.2
$G7231_{5.3}$	-2.0	2.20	3.2	-2.4	2.24	3.1
$LPC10_{2.4}$	-3.1	29.08	2.4	-3.7	26.55	1.9
$codec2_{1.3}$	-2.8	3.57	2.1	-3.1	3.05	2.1

low bit-rate codecs show a strong decrease in Segmental Signal to Noise Ration (SNR) and Perceptual Evaluation of Speech Quality (PESQ) standardised as ITU-T P.862. In particular LPC10 produce the highest distortion with an Itakura-Saito (IS) distance of 29.08 and 26.55 for the GEMEP and CPSD datasets respectively.

2) LOW-PASS FILTERING

To test the contribution of acoustic information in high frequency bands for the recognition of human vocalizations we applied low-pass filtering to the stimuli using the Praat software [67]. We low-pass filtered the stimuli using a 'static' as well as as a 'dynamic' method of filtering. Both types correspond to acoustic filtering as it occurs in natural and social conditions. Degraded speech signals resulting from static low-pass filtering usually occurs in daily environments, when voices are heard through distance or through walls, for example. The extraction of the dynamic pattern of certain formants usually happens by the cognitive and neural decoding of human speech in the brain of the listener as outlined in the introduction. While the first method is speech-unspecific since it simply supresses any signal above a certain cut-off frequency, the second filtering method takes into account the spectral and dynamical properties of the speech signal in terms of formant patterns as they are relevant for human speech and vocalization recognition.

a: STATIC LOW-PASS-FILTERING WITH CUT-OFF FREQUENCIES

We used six different cut off frequencies, namely 500 Hz, 1 kHz, 2 kHz, 3 kHz, 4 kHz, and 5 kHz. Low-pass filters were constructed as Hann bandpass filters between 0 Hz and the different cut-off frequencies with a roll-off bandwidth of 100 Hz. Figure 1a shows the spectrograms of an excerpt taken from the CPSD dataset at different cut-off frequencies from 1-5kHz. These static cut-off frequencies were chosen because of two reasons: first, they correspond to a specific feature that is frequently reported in studies on human perception of vocal emotions [68], and which is referred to as alpha ratio [46] or as Hammerberg index [49]. This ratio or index reflects the relative amount of spectral energy in high versus low frequency bands, and the static cut-off frequency for calculating this ratio is variably reported in the literature [46], [49]. Second, the selection of the cut-off frequencies used in the present study roughly follows the spectral distribution of the mean formant frequencies across several different expressions of vocal emotions [49].

b: DYNAMIC LOW-PASS FILTERING WITH SPEECH FORMANT CONTOURS

For this formant filtering procedure we extracted the first one to five formants (F1–F5) for each vocal stimulus. We created stimuli for five different conditions. For the F1 condition we

extracted the first formant (F1) for each stimulus using a temporal analysis window of 25 ms and a spectral pre-emphasis of 50 Hz to create a flatter spectrum. The original voice then was filtered with the extracted temporal F1 pattern, resulting in a stimulus with nearly zero energy in the spectral frequency bands above the F1. For the F2 condition we extracted the first two formants and applied the combined F1–F2 pattern to filter to the original stimulus resulting in a stimulus with a spectral profile including the F1 and the F2, but no spectral energy above F2. This procedure was repeated for the F3 condition, the F4 condition, and the F5 condition with adapted procedures according to the number of formants. Figure 1b shows the spectrograms of an excerpt taken from the CPSD dataset filtered at different upper formant frequencies. For each condition we set maximum frequencies for the formant detection to 600 Hz, 1600 Hz, 2600 Hz, 3600 Hz, and 5500 Hz, for the F1 to F5 conditions, respectively. Table 4 gives the mean and standard deviation of dynamically ranged values corresponding to different upper formants from F1 to F5. The difference in particular for F1 among the databases is owing to CPSD consisting of children voices as opposed to the adult voices contained in GEMEP. The formant filtering was applied only to voiced segments (cf. Table 5) leading to a lower bitrate.

TABLE 4. Mean and standard deviation of dynamic filtering values corresponding to different upper formants from F1 to F5. Values for the GEMEP and the CPSD databases.

[kHz]	F1	F2	F3	F4	F5
			GEMEP (5m, 5f)		
Male	0.286	0.746	1.763	2.671	4.443
	±0.074	±0.150	±0.112	±0.130	±0.210
Female	0.294	0.760	1.781	2.696	4.452
	±0.076	±0.132	±0.116	±0.142	±0.188
			CPSD (81m, 18f)		
Male	0.417	0.714	1.698	2.622	4.394
	±0.038	±0.101	±0.170	±0.191	±0.208
Female	0.418	0.734	1.736	2.629	4.383
	±0.038	±0.094	±0.156	±0.187	±0.186

TABLE 5. Voiced and unvoiced time in seconds in the GEMEP and the CPSD databases.

[s]	Voiced	Unvoiced
GEMEP	1681	1088
CPSD	1335	370

Table 6 shows the different objective measures per low-pass filtering constellation. We can observe that low bit-rate filtering show a strong decrease in SSNR and PESQ. In particular $\text{DYNAMIC}_{\text{F1}}$ produce the lowest SSNR on both datasets as expected. In general the dynamic filtering, produced lower SSNR than static filtering given that it was applied only to voiced segments. The same trend can be observed for PESQ where lower values are shown in the case of dynamic filtering. IS is in general very high given the high spectral difference between the original signal and the filtered ones.

TABLE 6. Objective measures averaged over the GEMEP and CPSD datasets in the case of static and dynamic filtering. Segmental signal to noise ratio (SSNR), Itakura-Saito distance (IS), and perceptual evaluation of speech quality (PESQ).

$\text{filtering}_{[\text{kHz}]}$	GEMEP			CPSD		
	SSNR	IS	PESQ	SSNR	IS	PESQ
STATIC_5	27.3	87.8	4.5	28.2	99.7	4.5
STATIC_4	25.4	90.3	4.5	24.7	99.8	4.5
STATIC_3	23.0	92.4	4.3	21.2	99.9	4.3
STATIC_2	19.7	94.0	3.9	17.9	99.9	3.9
STATIC_1	14.3	96.1	3.5	13.2	99.9	3.6
$\text{STATIC}_{0.5}$	10.4	97.2	3.0	4.7	99.9	3.2
$\text{DYNAMIC}_{\text{F5}}$	-0.6	96.1	2.2	0.8	99.6	2.5
$\text{DYNAMIC}_{\text{F4}}$	-1.1	97.4	2.1	0.2	99.7	2.5
$\text{DYNAMIC}_{\text{F3}}$	-1.4	97.4	2.2	-0.1	99.7	2.5
$\text{DYNAMIC}_{\text{F2}}$	-1.5	97.7	2.3	-0.5	99.7	2.6
$\text{DYNAMIC}_{\text{F1}}$	-2.6	96.2	2.6	-1.0	99.8	2.9

TABLE 7. Applied functionals. [1]: arithmetic mean of LLD / positive Δ LLD. [2]: not applied to voice related LLD except F_0. [3]: only applied to F0.

Functionals applied to LLD / Δ LLD
quartiles 1–3, 3 inter-quartile ranges
1 % percentile (≈ min), 99 % percentile (≈ max)
percentile range 1 %–99 %
position of min / max, range (max – min)
arithmetic mean[1], root quadratic mean
contour centroid, flatness
standard deviation, skewness, kurtosis
relative duration LLD is above 25 / 50 / 75 / 90% range
relative duration LLD is rising
relative duration LLD has positive curvature
gain of linear prediction (LP), LP Coefficients 1–5
mean, max, min, stdandard deviation of segment length[2]
Functionals applied to LLD only
mean value of peaks
mean value of peaks – arithmetic mean
mean / stdandard deviation of inter peak distances
amplitude mean of peaks, of minima
amplitude range of peaks
mean / stdandard deviation of rising / falling slopes
linear regression slope, offset, quadratic error
quadratic regression a, b, offset, quadratic error
percentage of non-zero frames[3]

C. ACOUSTIC FEATURES

We based our analysis on the ComParE acoustic feature set [69], which is an improved version of the INTERSPEECH 2012 Speaker Trait Challenge baseline feature set [20]. The features are a brute-force set of 6 373 acoustic features, where numerous functionals (such as but not limited to mean, standard deviation, regression coefficients; cf. Table 7) are applied to a large set of commonly used low-level descriptors (LLDs) and their delta coefficients as indicated in Table 8. The set includes energy, spectral, cepstral (MFCC), voicing related low-level descriptors (LLDs) as well as voice quality features (jitter and shimmer) including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, psychoacoustic spectral sharpness.[4] Note that the LLDs do not contain formant-related features given the testing conditions of different formant positions. The functionals

[4]Note that the LLDs do not contain formant-related features. This is reasonable given the testing of cutting at different formant positions.

TABLE 8. Low-level descriptors (LLD).

4 energy related LLD
Sum of auditory spectrum (loudness).
Sum of RASTA-style filtered auditory spectrum.
RMS Energy.
Zero-Crossing Rate.
55 spectral LLD
RASTA-style auditory spectrum, bands 1–26 (0–8 kHz).
MFCC 1–14.
Spectral energy 250–650 Hz, 1 k–4 kHz.
Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90.
Spectral Flux, Centroid, Entropy, Slope
Psychoacoustic Sharpness, Harmonicity, Variance, Skewness, Kurtosis.
6 voicing related LLD
F0 by SHS + Viterbi smoothing, Probability of voicing
logarithmic HNR, Jitter (local, delta), Shimmer (local).

are applied over one utterance resulting in one 6 373 dimensional feature vector for every utterance, regardless of its length. The features are extracted with our open-source audio and paralinguistics analysis toolkit openSMILE [70].

D. MACHINE-BASED TRAINING AND CLASSIFICATION

Thus, besides the analysis of the corruption outcome, we analysed the potential of matched and multi-condition training as opposed to miss-matched condition.

Based on the extraction of features as outlined in the previous section, a system has to be designed to map these features to the target variable of interest, such as affective vocalization (GEMEP database) or impaired vocalization from clinical population (CPSD database). In order to map our feature vector to the target domain, we choose static classifiers given our frame-wise or supra-segmental features. With static classification we refer to the process of assigning a discrete class label to an unknown feature vector of fixed dimensionality. For transparency and reproducibility, we used an open-source classifier implementation from the WEKA data mining toolkit [72]. Linear kernel Support Vector Machines (SVM) were used, which are known to be robust against over-fitting and which were also used for the baseline calculation in ComParE 2013 [69]. SVM are originally designed for binary classification, however, we used a multi-class SVM [71] by combining several binary classifiers trained in a 'one-against-one' fashion. As a training algorithm, we used Sequential Minimal Optimisation (SMO). The complexity parameter C was set to the values that achieved best UAR on the development set as reported by Schuller and colleagues [69]. For the GEMEP corpus C was set to $C = 0.01$ for arousal, $C = 0.1$ for valence and $C = 1.0$ for category. For the CPSD corpus C was set to $C = 0.01$ for typicality and $C = 0.001$ for diagnosis. The optimization of the hyperparameters on the development set is a traditional procedure to avoid overfitting.

Techniques from ASR for acoustic pre-processing and signal enhancement or multi-condition training have typically been applied to boost performances in degraded acoustic conditions caused by additive noise [73], reverberation, and also speech codec compression [11]. Thus, besides the analysis of the sheer corruption outcome, we analysed the potential of matched and multi-condition training as opposed to miss-matched condition. Matched conditions learning is used, which refers to training on data that is filtered at the same cut-off frequency or upper formant as the test data. Mismatched conditions training refers to training on the original, clean data and testing on filtered data. It investigates the performance of a generic model in varying filtering conditions. A third alternative is multi-condition training, which combines benefits from matched and mismatched conditions training. Multiple copies of the data filtered at different cut-off frequencies (or formants) are used during training. Thus, a generic model is generated, which is expected to work well in a variety of filtered conditions.

Here we evaluate for the first time the impact of speech coders and low-pass filtering by a set of different cut-off frequencies under mismatched, matched, and multi-condition learning for speaker state (affective vocalizations) and trait analysis (clinical populations). As primary evaluation measure, we retain the choice of unweighted average recall (UAR) as used in ComParE 2013 and broadly in the field [74]. In the given case of two classes ('X' and 'NX'), it is calculated as (Recall(X)+Recall(NX))/2, i. e., the number of instances per class is ignored by intention. The motivation to consider *unweighted* rather than weighted average recall ('conventional accuracy') is that it is also meaningful for highly unbalanced distributions of instances among classes, as is given for CPSD. In the case of equal distribution, UAR and 'usual accuracy' naturally resemble each other.

To cope with imbalanced class distribution in the CPSD set, up-sampling is applied on the training data. The under-represented categories (PDD, NOS, DYS) in the four-way diagnosis task are up-sampled by using a factor of five. In the binary typicality task a factor of two is applied. No re-sampling of the training set is done for the GEMEP set which appears sufficiently equally distributed.

We only show the evaluation on the test set, thus we re-train the models using the training and development set, applying re-sampling as described above.

III. RESULTS

The results are summarised in Table 9 for the standard speech coders, in Table 10 for the classification of affective vocalizations and in Table 11 for the classification of vocalizations from the clinical population. In Table 10 and Table 11, we show results for the static filtering ('Cut-off') and for the dynamic formant filtering ('Formants') scenarios in the three training methods (mismatched condition, matched condition, and multi-condition training). In multi-condition training we join the training sets of different formant (or static) filtering conditions (e. g., clean, F1, F2, F3, F4, F5) and evaluate on all the test sets. In the case of mismatched evaluations we train on the unfiltered clean training sets and evaluate on all test sets. Finally, in the case of matched condition we

TABLE 9. Emotion in the GEMEP set and Autism in the CPSD set – Detailed results for clean (*mi*), matched condition (*ma*) and multi-condition (*mu*) training with eleven codecs. Unweighted average recall (UAR) for the 'arousal task', the 'valence task', the 'category task', the 'typicality task', and the 'diagnosis task' on all coded conditions. For comparison: baseline results [69] ('clean'). The asterisk * indicates, which improvements w.r.t. the mismatched system are significant (one-tailed t-test, $p < 0.05$).

UAR[%]	GEMEP									CPSD						Avg.
	Arousal			Valence			Emotion			Typicality			Diagnosis			
	mi	*ma*	*mu*	*mi*	*ma*	*mu*	*mi*	*ma*	*mu*	*mi*	*ma*	*mu*	*mi*	*ma*	*mu*	
clean	75.0	75.0	72.9	61.6	61.6	62.3	40.9	40.9	40.4	90.7	90.7	89.9	67.1	67.1	62.9	66.6
PCM_{128}	71.2	74.6	74.2	60.1	63.2	60.7	34.3	36.8	38.4	85.7	89.4	90.9	56.6	63.4	62.2	64.1
$G711_{64}$	70.9	74.1	74.9	60.3	60.0	62.5	33.1	37.4	39.7	85.7	89.9	91.0	57.3	62.5	62.6	64.1
$G726_{40}$	71.3	73.4	74.2	58.7	61.1	62.6	33.7	37.6	38.4	85.6	89.2	90.2	58.0	62.7	62.2	63.9
$G726_{32}$	71.6	72.5	75.1	58.9	58.8	62.5	34.0	35.6	40.3	85.5	88.8	89.5	58.1	62.1	61.3	63.6
$G726_{24}$	67.5	74.8	73.0	58.7	57.6	61.4	26.0	33.9	36.4	85.2	90.7	89.9	56.3	60.7	60.6	62.2
$G728_{16}$	74.3	75.5	75.1	60.8	56.9	62.5	32.1	38.3	40.3	85.7	89.8	89.5	56.8	62.4	61.3	64.1
GSM_{13}	72.0	74.7	73.0	58.9	59.5	61.4	34.0	34.5	36.4	84.4	89.2	89.9	57.5	62.6	60.6	63.2
$G7231_{6.3}$	73.0	74.5	74.5	62.4	60.9	65.1	31.0	36.0	36.9	83.5	89.1	90.2	52.3	61.6	61.7	63.5
$G7231_{5.3}$	73.2	75.4	71.4	58.0	61.6	65.8	30.7	33.9	33.9	82.1	90.0	89.9	51.9	61.2	61.5	62.7
$LPC10_{2.4}$	63.0	73.4	72.5	56.1	61.9	62.1	25.8	32.6	31.0	73.3	86.1	82.9	40.1	64.2	56.3	58.8
$codec2_{1.3}$	61.1	74.5	73.1	56.0	56.9	63.0	25.4	35.3	36.3	80.1	89.7	87.6	51.2	62.4	56.3	60.6
Avg.	69.9	74.3	73.7	59.0	59.9	62.7	30.9	35.6	37.1	83.3	89.3	89.2	54.2	62.3	60.6	62.8

TABLE 10. Emotion in the GEMEP set – Detailed results for clean (*mi*), matched condition (*ma*) and multi-condition (*mu*) training with five static cut-off frequencies (in the 0.5–5 kHz range) and five different upper formants from F1 to F5. Unweighted average recall (UAR) for the 'arousal task', the 'valence task' and the 'category task' on all low-pass filtered conditions. For comparison: baseline results [69] ('clean'). The asterisk * indicates, which improvements w.r.t. the mismatched system are significant (one-tailed t-test, $p < 0.05$).

UAR[%]	Cut-off [kHz]							Formants						
	0.5	1	2	3	4	5	Avg.	F1	F2	F3	F4	F5	Avg.	Clean
Arousal *(mi)*	61.1	70.9	74.4	73.4	72.0	75.7	71.3	51.9	49.3	47.9	48.2	48.8	53.5	75.0
Arousal *(ma)*	*74.2	73.8	74.0	75.2	73.9	74.3	74.2	*75.0	*72.6	*73.1	*74.2	*72.9	73.8	75.0
Arousal *(mu)*	*73.6	73.7	74.0	75.8	74.1	75.3	74.4	*72.4	*74.1	*71.9	*71.9	*71.4	72.5	73.6
Valence *(mi)*	56.0	57.0	61.7	62.6	61.0	63.3	60.3	56.0	57.1	59.6	58.9	53.5	57.8	61.6
Valence *(ma)*	61.6	60.5	59.3	60.7	61.8	61.9	61.0	60.0	62.8	62.8	61.6	*63.4	62.0	61.6
Valence *(mu)*	61.1	62.3	62.3	62.3	61.6	60.7	61.7	60.0	*64.6	64.4	64.2	*63.9	63.5	60.2
Category *(mi)*	16.7	21.1	27.0	32.9	33.7	37.8	28.2	17.7	15.3	15.2	17.3	15.3	20.3	40.9
Category *(ma)*	31.0	34.0	37.0	33.0	36.1	37.5	34.8	28.6	*32.1	*32.5	*33.6	*30.3	33.0	40.9
Category *(mu)*	31.0	33.3	35.4	36.9	37.2	37.7	35.3	32.4	*33.4	*34.8	*36.0	*36.9	34.8	36.2

TABLE 11. Autism in the CPSD set – Detailed results for clean (*mi*), matched condition (*ma*) and multi-condition (*mu*) training with five static cut-off frequencies (in the 0.5–4 kHz range) and five different upper formants from F1 to F5. Unweighted average recall (UAR) for the 'typicality task', and the 'diagnosis task' on all low-pass filtered conditions. For comparison: baseline results [69] ('clean'). The asterisk * indicates, which improvements w.r.t. the mismatched system are significant (one-tailed t-test, $p < 0.05$).

UAR[%]	Cut-off [kHz]							Formants						
	0.5	1	2	3	4	5	Avg.	F1	F2	F3	F4	F5	Avg.	Clean
Typicality *(mi)*	72.7	86.8	88.5	89.3	84.4	87.5	84.9	68.5	64.3	68.2	71.5	74.5	73.0	90.7
Typicality *(ma)*	*86.0	88.7	90.2	90.6	*89.5	90.0	89.2	*86.8	*85.2	*86.7	*85.3	*87.2	87.0	90.7
Typicality *(mu)*	*86.3	87.1	89.7	89.1	*89.4	88.0	88.3	*86.2	*86.0	*86.2	*84.4	*88.1	86.8	91.1
Diagnosis *(mi)*	45.4	51.7	55.7	57.3	58.1	61.4	54.9	36.7	36.7	36.2	40.2	39.8	42.8	67.1
Diagnosis *(ma)*	*55.6	53.1	58.1	60.0	62.2	62.3	58.6	*56.4	*49.9	*57.3	*55.9	*61.6	58.0	67.1
Diagnosis *(mu)*	54.4	52.8	56.7	59.1	60.2	60.1	57.2	*53.5	*55.1	*59.4	*54.8	*55.3	56.4	61.2

train on a certain filtered training set (e. g., F2) and evaluate on the related test partition (e. g., F2). The same applies for the static filtering scenario. Note that, accordingly the training set in multi-condition training is 7 (for static filtering) and 6 (for formant filtering) times the size of the original (clean) training set. The results for clean testing are given in the 'clean' column. Note that 'clean' here represents non-filtered speech – the term was chosen owing to its common usage when dealing with noisy or even reverberated speech.

A. CLASSIFICATION ACCURACY OF AFFECTIVE VOCALIZATIONS

1) SPEECH CODERS

The results in Table 9 ('GEMEP' column) show the general performance trends for the Arousal, Valence, and Category tasks over the three training methods in the eleven speech coding scenarios.

We observed that with clean training (mismatched) there was a drop in performances that became more evident for lower bit-rate coders. With codec2 – showing the lowest

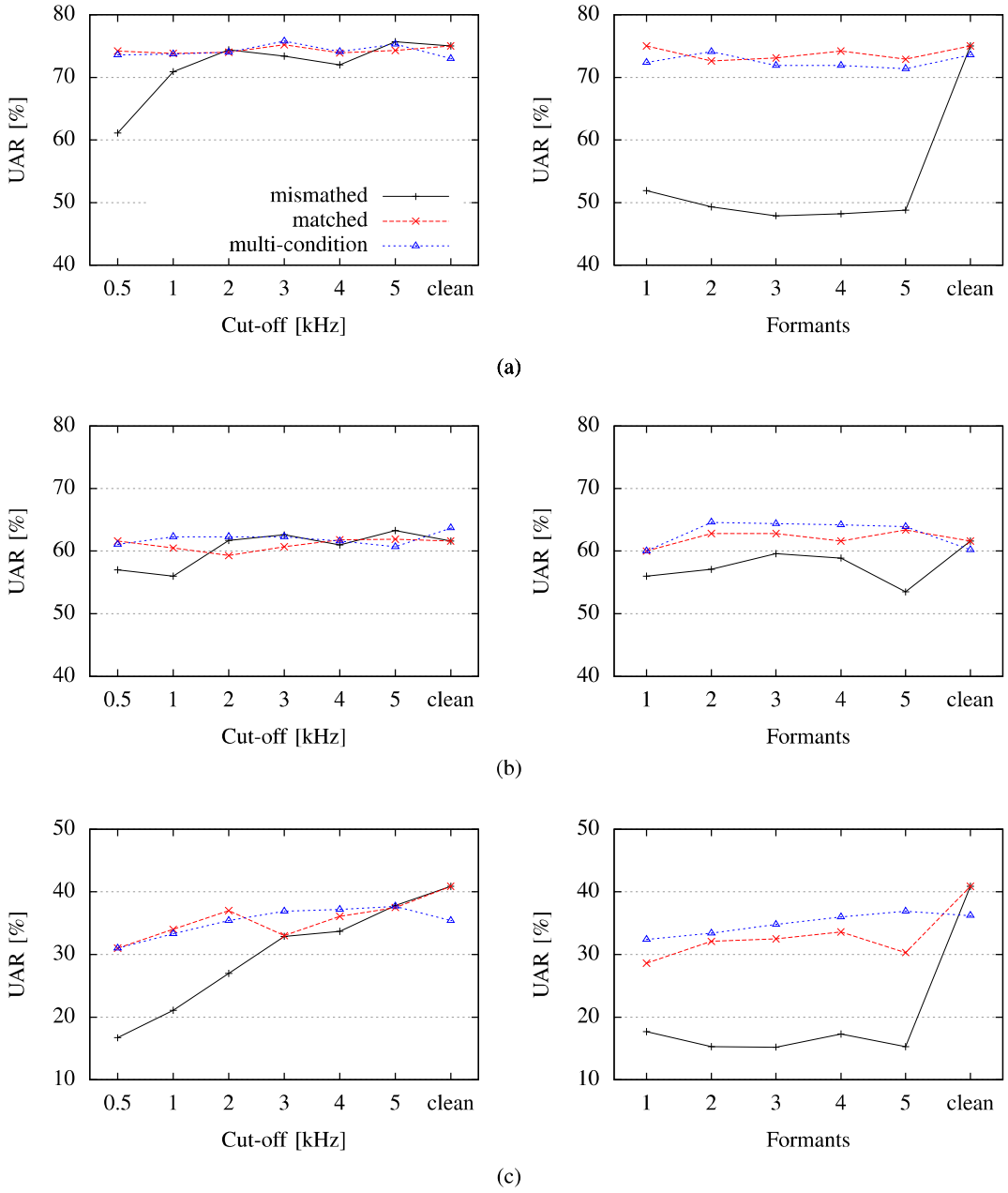

FIGURE 2. **Detailed results for clean (mismatched), matched condition and multi-condition training with five static cut-off frequencies (left) and five different upper formants from F1 to F5 (right). Unweighted average recall (UAR) for the (a) 'arousal task', (b) the 'valence task', and (c) the 'category task' on all low-pass filtered conditions (cf. also Table 10 for details). For comparison: baseline results [69] ('clean').**

bit-rate of 1.3 kbit/s – we obtained the lowest results for all the three classification tasks (61.1 %, 56.0 %, and 25.4 % UAR for the arousal, the valence, and category task, respectively), dropping considerably below the baseline reported by Schuller and colleagues [69] and indicated in the 'clean' column (75.0 %, 61.6 %, and 40.9 % UAR for the arousal, the valence, and the category task, respectively). Matched and multi-condition training performed notably better. In fact, for codec2 with matched training we obtained the best result (74.5 % UAR) for the arousal. With multi-condition training we achieved on average the best performances over all the different test conditions. In particular, for the arousal and the valence task we even outperformed the baseline results by achieving respectively 75.1 % and 65.8 % UAR when testing on $G.728_{16}$ and $G.7213_{5.3}$ signals respectively. Note that, for the valence task we also outperformed the baseline when testing on 'clean', i. e., non-filtered, speech (62.3% UAR).

2) LOW-PASS FILTERING

The results in Table 10 and Figure 2 show the general performance trends for the Arousal, Valence, and Category tasks over the three training methods in the two filtering scenarios.

a: STATIC LOW-PASS FILTERING WITH CUT-OFF FREQUENCIES

By looking at the static filtering scenario ('Cut-off' columns), we observed that with clean training (mismatched) there was a drop in performances that became more evident for lower cut-off frequencies. At 500 Hz we obtained the lowest results (in the static filtering scenario) for all the three classification tasks (61.1 %, 56.0 %, and 16.7 % UAR for the arousal, the valence, and category task, respectively). Matched and multi-condition training performed notably better. With matched training we obtained the best results at 500 Hz (74.2 %, 61.6 %, and 31.0 % UAR for the arousal, the valence, and the category task respectively). It is particularly interesting to observe that for the valence task we reached the baseline performance. This means that the 500 Hz low-pass filtered signals contained sufficient and relevant information for the valence classification task. With multi-condition training we achieved on average the best performances over all the different test conditions. In particular, for the arousal and the valence task we even outperformed the baseline results by achieving respectively 75.8 % and 62.3 % UAR when testing on 3 kHz low-pass filtered signals. For the comparison between the binary valence task (positive vs. negative valence) and the category task (12 emotional categories) we found that the performance for the mismatched conditions drops only for the lowest cut-off frequencies (0.5 and 1 kHz), whereas performance already drops for the 2 kHz conditions for the category task including multiple categories.

b: DYNAMIC LOW-PASS FILTERING WITH SPEECH FORMANTS

For the dynamic filtering scenario ('Formants' columns), the performance trends were similar to the static filtering scenario, but the drop of performance already happened with higher levels of low-pass filtering. We again found that with clean training (mismatched) the performances are strongly dropping below the baseline in all of the five testing conditions (47.9 %, 53.5 %, and 15.3 % UAR for the arousal, the valence, and the category task, respectively). Again, matched and multi-condition training performed notably better. For the arousal task, matched training performed slightly better than multi-condition by reaching the baseline (75.0 % UAR) when testing on the F1-filtered test partition. For the valence and the category task, multi-condition training was evidently boosting the performances across all the different test conditions. It is interesting to note that one even outperforms the baseline by reaching 64.6 % UAR for the valence task when testing on the F2-filtered condition.

B. CLASSIFICATION ACCURACY OF VOCALIZATIONS FROM CLINICAL POPULATIONS

1) SPEECH CODERS

With respect to the machine-based classification of vocalizations from the clinical population, Table 9 ('CPSD' column) show the general performance trends for the typicality and diagnosis tasks over the three training methods in the eleven speech coding scenarios.

With clean training (mismatched) there was a drop in performances that became more evident for lower bit-rate coders. In particular, with LPC10 we obtained the lowest results for the two classification tasks (73.3 %, and 40.1 % UAR for the typicality, and diagnosis task). The results are significantly below the baseline reported in [69] and indicated in the 'clean' column (90.7 %, 67.1 % UAR for the typicality and the diagnosis task, respectively). This means that codecs with very low PESQ and SSNR such as LPC10 are not containing sufficient information the typicality and diagnosis tasks. Matched and multi-condition training performed notably better. With matched training we obtained the best results with LPC10 (86.1 %, and 64.2 % UAR for the typicality, and the diagnosis task respectively). With matched training we achieved on average the best performances over all the different test conditions.

2) LOW-PASS FILTERING

Table 11 and Figure 3 show the general performance trends for the typicality and diagnosis tasks considering the three training methods in the two filtering scenarios.

a: STATIC LOW-PASS FILTERING WITH CUT-OFF FREQUENCIES

In the static filtering scenario ('Cut-off' columns), we found that clean training (mismatched) leads to lower performances. The lower the cut-off frequency, the lower the performance is. Similar as for the emotion tasks described above, at 500 Hz we observed the lowest results for the two tasks (72.7 %, and 45.4 % UAR for the typicality and the diagnosis task, respectively). The results are significantly below the baseline. Applying matched and multi-condition training increased the results mostly on the lower cut-off frequencies scenarios. The two training strategies performed similarly over the different testing condition. With a matched training condition we obtained slightly better results on average. In particular, for the typicality task UAR is achieved very close to the baseline when testing at 2, 3 and 5 kHz cut-off frequencies (90.2 %, 90.6 %, and 90.0 % UAR, respectively). This indicates that relevant information for this task is contained in lower frequency bands. For the diagnosis task, we obtained a maximum of 62.3 % UAR when testing on 5 kHz-filtered signals. In this case we were far from the baseline meaning that relevant information is still present at 'even higher' frequencies.

b: DYNAMIC LOW-PASS FILTERING WITH SPEECH FORMANTS

Concerning the dynamic filtering scenario ('Formants' columns), the performance trends are partially worse in comparison with the static filtering scenario. With clean training (mismatched) the performances strongly dropped below the baseline on all the five testing conditions (64.3 % and 36.2 % UAR for the typicality and the diagnosis task, respectively).

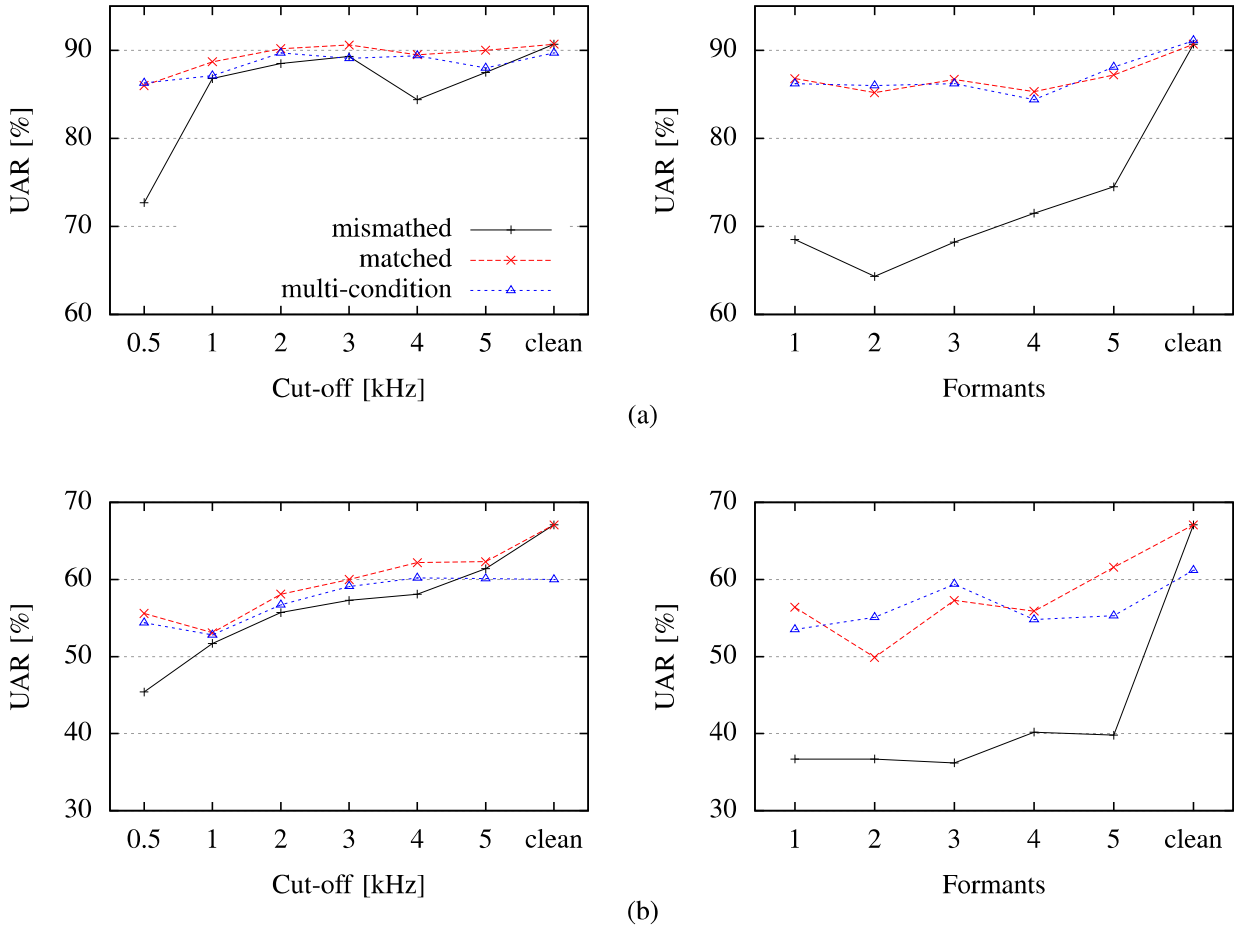

FIGURE 3. **Detailed results for clean (mismatched), matched condition and multi-condition training with five static cut-off frequencies (left) and five different upper formants from F1 to F5 (right). Unweighted average recall (UAR) for the (a) 'typicality task', and (b) the 'diagnosis task' on all low-pass filtered conditions (cf. also Table 11 for details). For comparison: baseline results [69] ('clean').**

In this case the performances were even below the ones obtained in the static filtering scenario. With matched and multi-condition training again better results are obtained. For the typicality task we measured performances close to the baseline only when testing with a higher number of formants (e. g., we obtain 88.1 % UAR when testing on F5). It is interesting to see that we outperformed the baseline in multi-condition training when testing on clean signals (91.1 % UAR). For the diagnosis task, we obtained a maximum of 61.6 % UAR when testing on F5-filtered conditions.

IV. DISCUSSION

Considering that practically no knowledge exists on the effects of machine-based speech coding in narrow-band transmission with regard to the recognition of paralinguistic cues in human vocalizations, we evaluated the impact of speech degradation applying several standard speech coders and by low-pass filtering of the speech signal from different sources (affective, clinical) by a set of different static and dynamic cut-off frequencies. The aim was to test the performance of a computer-based speech classifier on degraded speech input by investigating the sensitivity of the classifier to different acoustic information contained in the different frequency bands. For a static condition of low-pass filtering we used values in the 0.5–5 kHz range, whereas for the condition of dynamic filtering we used different upper speech formants from F1 to F5. We were mainly interested in the influence of three major factors on the classification performance of the classifier: (1) the type of paralinguistic vocalizations (affective, clinical), (2) the low-pass filtering method (static, dynamic), and (3) the training method of the classifier (clean/mismatched, matched, multi-condition).

We thus first evaluated the impact of speech degradation using low bit-rate speech coders ranging from 64 to 1.3 kbit/s and we interestingly found that even with very low bit-rate the arousal, the valence, and the typicality tasks showed a slight performance degradation. However, more complex task such as the category, and the diagnosis seemed to be more sensitive to lower bit-rates.

We thus then first to gain more insight into the impact of speech coding of different paralinguistic cues originating from different vocalizations. For the tasks concerning

affective vocalizations, namely the arousal, the valence, and the category task we found that relevant information is still contained in the bandwidth-limited signals and that the performance of the classifier generally stayed astonishingly invariant (e. g., category task) or even increases (e. g., arousal and valence task) at first in case of static filtering. In general the machine-based classification accuracy was close to the baseline condition for the different tasks even for a considerable low-pass filtering with static cut-off frequencies when using matched or multi-condition training. The performance of the classifier only substantially dropped for extreme levels of low-pass filtering with 0.5 kHz or in case of dynamic filtering without matched or multi-condition training. This drop of performance was evident for each task performed on the affective vocalizations. However, there were also some specific performance differences between the tasks. For example, while a binary valence classification of vocalizations seems only to become more inaccurate with extreme low levels of low-pass filtering, a more complex multi-categorical decision already shows a drop in performance with intermediate low-pass filtering, especially during the mismatched training condition. Thus, the more alternatives for the categorical decision exist, the more the accuracy drops with decreasing levels of low frequency information.

Unlike the classification accuracy on affective vocalizations using cut-off filtering, we found a much more impaired classification accuracy with dynamic formant filtering, again especially during the mismatched training condition. All levels of formant filtering revealed considerably decreased classification accuracy. This is indicative that the speech formants provide an important paralinguistic cue, even in higher formant ranges. This drop in performance was specifically evident for the binary arousal and the categorization tasks with multiple categories. Thus, the speech formants seem to carry a lot of information about the arousal level of affective vocalization and they seem to support the discrimination of multiple emotional categories, but only have small information for the distinction of negative and positive emotions.

For the tasks related to the clinical populations, we found that the classification performances remained more closer to the state-of-the-art performance (typicality task) or slightly deteriorate (diagnosis task) as compared to the tasks performed on affective vocalizations. This drop of performance was most evident again for the lowest cut-off frequency in the static filtering condition and especially for the mismatched training condition. Furthermore, the drop of performance was again greater for the condition of multiple classification categories (i. e., the typicality task). For the dynamic formant filtering condition we also found a drop of performance already with only filtering above F5, but this drop was less pronounced as for the tasks performed on affective vocalizations.

Besides the performance of the classifier according to the different speech databases and the different tasks performed on each dataset, the third interest of this study was to investigate the impact of different training procedures of the classifier. We analysed the potential of matched and multi-condition training as opposed to the miss-matched condition. The results corroborates common evidence that multi-condition and matched-condition training significantly increase performances as opposed to mismatched condition. This applies both to the classification of affective vocalizations as well as to the classification of vocalizations from clinical populations.

Future work might deal with extending the multi-condition learning approach to more diverse filtering types, as well as an in-depth analysis on acoustic features under bandlimit. Further, one can evaluate approaches as originally tailored for ASR for the transfer to the domain of Computational Paralinguistics. In particular, transfer learning and domain adaptation techniques can be investigated.

REFERENCES

[1] A. S. Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, no. 10, pp. 1541–1582, Oct. 1994.

[2] B. T. Lilly and K. K. Paliwal, "Effect of speech coders on speech recognition performance," in *Proc. 4th Int. Conf. Spoken Lang. (ICSLP)*, vol. 4. Oct. 1996, pp. 2344–2347.

[3] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 115–132, Jan. 1994.

[4] A. Gallardo-Antolin, C. Pelaez-Moreno, and F. Diaz-de-Maria, "Recognizing GSM digital speech," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1186–1205, Nov. 2005.

[5] N. B. Yoma, C. Molina, J. Silva, and C. Busso, "Modeling, estimating, and compensating low-bit rate coding distortion in speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 1, pp. 246–255, Jan. 2006.

[6] A. M. Gomez, A. M. Peinado, V. Sanchez, and A. J. Rubio, "Recognition of coded speech transmitted over wireless channels," *IEEE Trans. Wireless Commun.*, vol. 5, no. 9, pp. 2555–2562, Sep. 2006.

[7] W.-H. Hsu and L.-S. Lee, "Efficient and robust distributed speech recognition (DSR) over wireless fading channels: 2D-DCT compression, iterative bit allocation, short BCH code and interleaving," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1. May 2004, pp. I-69–I-72.

[8] D. S. Kim and H. K. Kim, "Voicing class dependent Huffman coding of compressed front-end feature vector for distributed speech recognition," in *Proc. 2nd Int. Conf. Future Generat. Commun. Netw. Symp. (FGCNS)*, vol. 3. Dec. 2008, pp. 51–54.

[9] A. V. Ramana, P. Laxminarayana, and P. Mythilisharan, "Investigation of speech coding effects on different speech sounds in automatic speech recognition," in *Perception and Machine Intelligence* (Lecture Notes in Computer Science), vol. 7143, M. Kundu, S. K. Mitra, D. Mazumdar, and S. K. Pal, Eds. Berlin, Germany: Springer, 2012, pp. 367–377.

[10] A. K. Vuppala, S. Chakrabarti, and K. S. Rao, "Effect of speech coding on recognition of consonant-vowel (CV) units," in *Contemporary Computing* (Communications in Computer and Information Science), vol. 94, S. Ranka *et al.*, Eds. Berlin, Germany: Springer, 2010, pp. 284–294.

[11] R. B. Dunn, T. F. Quatieri, D. A. Reynolds, and J. P. Campbell, "Speaker recognition from coded speech and the effects of score normalization," in *Proc. Conf. Rec. 35th Asilomar Conf. Signals, Syst. Comput.*, vol. 2. Nov. 2001, pp. 1562–1567.

[12] M. Petracca, A. Servetti, and J. C. De Martin, "Performance analysis of compressed-domain automatic speaker recognition as a function of speech coding technique and bit rate," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 1393–1396.

[13] M. Phythian, J. Ingram, and S. Sridharan, "Effects of speech coding on text-dependent speaker recognition," in *Proc. IEEE Region 10 Annu. Conf. Speech Image Technol. Comput. Telecommun. (TENCON)*, vol. 1. Dec. 1997, pp. 137–140.

[14] T. Maka and L. Bonikowski, "Speech coding influence on features dedicated to speaker identification," in *Proc. Int. Conf. Signals Electron. Syst. (ICSES)*, Sep. 2008, pp. 489–492.

[15] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[16] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, pp. 1743–1759, Nov. 2009.

[17] F. Eyben, F. Weninger, N. Lehment, G. Rigoll, and B. Schuller, "Violent scenes detection with large, brute-forced acoustic and visual feature sets," in *Proc. MediaEval Workshop*, Pisa, Italy, 2012, pp. 1–2.

[18] B. Schuller *et al.*, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1760–1774, Nov. 2009.

[19] C.-C. Lee *et al.*, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Proc. InterSpeech*, Makuhari, Chiba, Japan, Sep. 2010, pp. 793–796.

[20] B. Schuller *et al.*, "The INTERSPEECH 2012 speaker trait challenge," in *Proc. INTERSPEECH*, Portland, OR, USA, 2012, pp. 254–257.

[21] J. Pittermann, A. Pittermann, and W. Minker, "Emotion recognition and adaptation in spoken dialogue systems," *Int. J. Speech Technol.*, vol. 13, pp. 49–60, Mar. 2010.

[22] M. Schröder *et al.*, "Building autonomous sensitive artificial listeners," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 165–183, Apr./Jun. 2012.

[23] B. Schuller *et al.*, "Reilly, D. Pigat, P. Robinson, "Recent developments and results of ASC-inclusion: An integrated Internet-based environment for social inclusion of children with autism spectrum conditions," in *Proc. 3rd Int. Workshop Intell. Digit. Games Empowerment Inclusion (IDGEI), 20th ACM Int. Conf. Intell. Interfaces (IUI)*, Mar. 2015, p. 9.

[24] B. Schuller *et al.*, "Reilly, D. Pigat, P. Robinson, "The state of play of ASC-inclusion: An integrated Internet-based environment for social inclusion of children with autism spectrum conditions," in *Proc. 2nd Int. Workshop Digit. Games Empowerment Inclusion (IDGEI)*, Feb. 2014, pp. 1–8.

[25] B. Schuller *et al.*, "ASC-inclusion: Interactive emotion games for social inclusion of children with autism spectrum conditions," in *Proc. 1st Int. Workshop Intell. Digit. Games Empowerment Inclusion (IDGEI), 8th Found. Digit. Games (FDG)*, May 2013, pp. 1–8.

[26] Z. Massida *et al.*, "Voice discrimination in cochlear-implanted deaf subjects," *Hearing Res.*, vol. 275, nos. 1–2, pp. 120–129, May 2011.

[27] E. Marchi *et al.*, "Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Dresden, Germany, Sep. 2015, pp. 115–119.

[28] E. Marchi *et al.*, "Voice emotion games: Language and emotion in the voice of children with autism spectrum condition," in *Proc. 3rd Int. Workshop Intell. Digit. Games Empowerment Inclusion (IDGEI), 20th ACM Int. Conf. Intell. User Interfaces (IUI)*, Mar. 2015, p. 9.

[29] E. Marchi, B. Schuller, A. Batliner, S. Fridenzon, S. Tal, and O. Golan, "Emotion in the speech of children with autism spectrum conditions: Prosody and everything else," in *Proc. 3rd Workshop Child, Comput. Interaction (WOCCI)*, Portland, OR, USA, ISCA, Sep. 2012, pp. 1–8.

[30] E. Marchi, A. Batliner, B. Schuller, S. Fridenzon, S. Tal, and O. Golan, "Speech, emotion, age, language, task, and typicality: Trying to disentangle performance and feature relevance," in *Proc. 1st Int. Workshop Wide Spectr. Social Signal Process. (WS3P), ASE/IEEE SocialCom*, Amsterdam, The Netherlands, Sep. 2012, pp. 961–968.

[31] J. Demouy *et al.*, "Differential language markers of pathology in autism, pervasive developmental disorder not otherwise specified and specific language impairment," *Res. Autism Spectr. Disorders*, vol. 5, no. 4, pp. 1402–1412, Oct./Dec. 2011.

[32] E. Mower, M. P. Black, E. Flores, M. Williams, and S. Narayanan, "Rachel: Design of an emotionally targeted interactive agent for children with autism," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2011, pp. 1–6.

[33] P. Belin, S. Fillion-Bilodeau, and F. Gosselin, "The montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing," *Behavior Res. Methods*, vol. 40, no. 2, pp. 531–539, May 2008.

[34] J. Schoentgen, "Vocal cues of disordered voices: An overview," *Acta Acustica United Acustica*, vol. 92, no. 5, pp. 667–680, 2006.

[35] I. Rektorova, J. Barrett, M. Mikl, I. Rektor, and T. Paus, "Functional abnormalities in the primary orofacial sensorimotor cortex during speech in Parkinson's disease," *Movement Disorders*, vol. 22, no. 14, pp. 2043–2051, Oct. 2007.

[36] A. Maier *et al.*, "PEAKS—A system for the automatic evaluation of voice and speech disorders," *Speech Commun.*, vol. 51, no. 5, pp. 425–437, May 2009.

[37] M. Chen, Y. Zhang, Y. Li, M. M. Hassan, and A. Alamri, "AIWAC: Affective interaction through wearable computing and cloud technology," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 20–27, Feb. 2015.

[38] T. H. Luan, R. Lu, X. Shen, and F. Bai, "Social on the road: Enabling secure and efficient social networking on highways," *IEEE Wireless Commun.*, vol. 22, no. 1, pp. 44–51, Feb. 2015.

[39] Y. E. Sagduyu and Y. Shi, "Navigating a mobile social network," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 122–128, Oct. 2015.

[40] H. Dahrouj, A. Douik, O. Dhifallah, T. Y. Al-Naffouri, and M.-S. Alouini, "Resource allocation in heterogeneous cloud radio access networks: Advances and challenges," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 66–73, Jun. 2015.

[41] M. Rockoff *et al.*, "The social implications of health care communication systems," *IEEE Trans. Commun.*, vol. 23, no. 10, pp. 1085–1088, Oct. 1975.

[42] D. W. Conrath, E. V. Dunn, J. N. Swanson, and P. D. Buckingham, "A preliminary evaluation of alternative telecommunication systems for the delivery of primary health care to remote areas," *IEEE Trans. Commun.*, vol. 23, no. 10, pp. 1119–1126, Oct. 1975.

[43] J. Turner and B. Kelly, "Emotional dimensions of chronic disease," *Western J. Med.*, vol. 172, no. 2, pp. 124–128, Feb. 2000.

[44] M. Patel and J. Wang, "Applications, challenges, and prospective in emerging body area networking technologies," *IEEE Wireless Commun.*, vol. 17, no. 1, pp. 80–88, Feb. 2010.

[45] P. Belin, "Voice processing in human and non-human primates," *Philos. Trans. Roy. Soc. B, Biol. Sci.*, vol. 361, no. 1476, pp. 2091–2107, Dec. 2006.

[46] S. Patel, K. R. Scherer, and E. Björkner, and J. Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biol. Psychol.*, vol. 87, no. 1, pp. 93–98, Apr. 2011.

[47] D. I. Leitman, P. Laukka, P. N. Juslin, E. Saccente, P. Butler, and D. C. Javitt, "Getting the Cue: Sensory Contributions to Auditory Emotion Recognition Impairments in Schizophrenia," *Schizophrenia Bull.*, vol. 36, no. 3, pp. 545–556, 2010.

[48] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: Different channels, same code?" *Psychol. Bull.*, vol. 129, no. 5, pp. 770–814, Sep. 2003.

[49] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality Social Psychol.*, vol. 70, no. 3, pp. 614–636, 1996.

[50] S. Frühholz, L. Ceravolo, and D. Grandjean, "Specific brain networks during explicit and implicit decoding of emotional prosody," *Cerebral Cortex*, vol. 22, no. 5, pp. 1107–1117, 2012.

[51] S. Frühholz and D. Grandjean, "Towards a fronto-temporal neural network for the decoding of angry vocal expressions," *NeuroImage*, vol. 62, no. 5, pp. 1658–1666, Sep. 2012.

[52] S. Frühholz and D. Grandjean, "Amygdala subregions differentially respond and rapidly adapt to threatening voices," *Cortex*, vol. 49, no. 5, pp. 1394–1403, May 2013.

[53] M. Latinus and P. Belin, "Human voice perception," *Current Biol.*, vol. 21, no. 4, pp. R143–R145, Feb. 2011.

[54] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Amer.*, vol. 97, no. 5, p. 3099, 1995.

[55] T. Waaramaa, A.-M. Laukkanen, M. Airas, and P. Alku, "Perception of emotional valences and activity levels from vowel segments of continuous speech," *J. Voice*, vol. 24, no. 1, pp. 30–38, 2010.

[56] M. D. Pell and S. A. Kotz, "On the time course of vocal emotion recognition," *PLoS ONE*, vol. 6, no. 11, p. e27256, 2011.

[57] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comput. Biol.*, vol. 5, no. 3, p. e1000302, 2009.

[58] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *Quart. J. Experim. Psychol.*, vol. 63, no. 11, pp. 2251–2272, 2010.

[59] E. D. Ross and M. Monnot, "Neurology of affective prosody and its functional–anatomic organization in right hemisphere," *Brain Lang.*, vol. 104, no. 1, pp. 51–74, Jan. 2008.

[60] M. Hoekert, R. S. Kahn, M. Pijnenborg, and A. Aleman, "Impaired recognition and expression of emotional prosody in schizophrenia: Review and meta-analysis," *Schizophrenia Res.*, vol. 96, nos. 1–3, pp. 135–145, Nov. 2007.
[61] R. B. Grossman, R. H. Bemis, D. P. Skwerer, and H. Tager-Flusberg, "Lexical and affective prosody in children with high-functioning autism," *J. Speech, Lang. Hearing Res.*, vol. 53, no. 3, pp. 778–793, 2010.
[62] J. McCann, S. Peppé, F. E. Gibbon, A. O'Hare, and M. Rutherford, "Prosody and its relationship to language in school-aged children with high-functioning autism," *Int. J. Lang. Commun. Disorders*, vol. 42, no. 6, pp. 682–702, Nov./Dec. 2007.
[63] C. R. Marshall, S. Harcourt-Brown, F. Ramus, and H. K. J. van der Lely, "The link between prosody and language skills in children with specific language impairment (SLI) and/or dyslexia," *Int. J. Lang. Commun. Disorders*, vol. 44, no. 4, pp. 466–488, 2009.
[64] D. K. Oller *et al.*, "Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 30, pp. 13354–13359, 2010.
[65] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, no. 5, pp. 1161–1179, 2012.
[66] F. Ringeval *et al.*, "Automatic intonation recognition for the prosodic assessment of language-impaired children," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1328–1342, Jul. 2011.
[67] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, nos. 9–10, pp. 341–345, 2002.
[68] S. Frühholz *et al.*, "Neural decoding of discriminative auditory object features depends on their socio-affective valence," *Social Cognit. Affective Neurosci.*, Jun. 2016.
[69] B. Schuller *et al.*, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Lyon, France, Aug. 2013, pp. 148–152.
[70] F. Eyben and M. Wöllmer, and B. Schuller, "openSMILE: The munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia*, Florence, Italy, 2010, pp. 1459–1462.
[71] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations Newslett.*, vol. 11, no. 1, pp. 10–18, Jun. 2009.
[72] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
[73] M. Wöllmer, E. Marchi, S. Squartini, and B. Schuller, "Multi-stream LSTM-HMM decoding and histogram equalization for noise robust keyword spotting," *Cognit. Neurodyn.*, vol. 5, no. 3, pp. 253–264, 2011.
[74] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, nos. 9–10, pp. 1062–1087, Nov./Dec. 2011.

SASCHA FRÜHHOLZ received the degrees in science of education and in psychology in 2001 and 2006, respectively, and the Ph.D. degree in neural mechanisms of facial expressions from Bremen University, Germany, in 2008. He currently holds an SNF Professorship (PP00P1_157409/1) with the Psychology Department, University of Zurich, Switzerland. His current projects deal with dynamic connectivity patterns of local and remote brain regions during auditory emotion processing using high-resolution brain scans and specific connectivity modeling approaches for functional imaging data.

ERIK MARCHI (M'14) received the M.Sc. (*cum laude*) degree in electronic engineering from Università Politecnica delle Marche, Ancona, Italy, in 2011. He is currently pursuing the Ph.D. degree with the Machine Intelligence and Signal Processing Group, Technische Universität München, Munich, Germany. His research focuses on affective computing, speech recognition, and acoustic novelty detection. His further areas of involvement is centered on the EU-FP7 project ASC-Inclusion and on the EU-H2020 project DE-ENIGMA to realize robust, context-sensitive, multi-modal, and naturalistic human-robot interaction aimed at enhancing the social imagination skills of children with autism, where he is leading the development of automatic speech analysis and affect sensing in spontaneous speech and real-life acoustic environments.

He has co-authored more than 50 publications in peer-reviewed journals and conference proceedings. He is a member of the ACM.

BJÖRN SCHULLER (M'05–SM'15) received the Diploma degree, the Ph.D. degree with a focus on automatic speech and emotion recognition, and the Habilitation degree from Technical University of Munich (TUM), Munich, Germany, in 1999, 2006, and 2012, respectively, all in electrical engineering and information technology. He was with TUM in 2012, as an Adjunct Teaching Professor in signal processing and machine intelligence. He is currently a Tenured Full Professor heading the Chair of Complex Systems Engineering, University of Passau, Germany, and a Reader in Machine Learning with the Department of Computing, Imperial College London, London, U.K. He has co-authored five books and more than 550 publications in peer-reviewed books, journals, and conference proceedings leading to more than 11 000 citations (h-index 51). He is an elected member of the IEEE Speech and Language Processing Technical Committee, and a member of the ACM and ISCA.