Synthesized speech for model training in cross-corpus recognition of human emotion

Björn Schuller · Zixing Zhang · Felix Weninger · Felix Burkhardt

Abstract Recognizing speakers in emotional conditions remains a challenging issue, since speaker states such as emotion affect the acoustic parameters used in typical speaker recognition systems. Thus, it is believed that knowledge of the current speaker emotion can improve speaker recognition in real life conditions. Conversely, speech emotion recognition still has to overcome several barriers before it can be employed in realistic situations, as is already the case with speech and speaker recognition. One of these barriers is the lack of suitable training data, both in quantity and quality—especially data that allow recognizers to generalize across application scenarios ('cross-corpus' setting). In previous work, we have shown that in principle, the usage of synthesized emotional speech for model training can be beneficial for recognition of human emotions from speech. In this study, we aim at consolidating these first results in a large-scale cross-corpus evaluation on eight of most frequently used human emotional speech corpora, namely ABC, AVIC, DES, EMO-DB, eNTERFACE, SAL, SUSAS and VAM, covering natural, induced and acted emotion as well as a variety of application scenarios and acoustic conditions. Synthesized speech is evaluated standalone as well as in joint training with human speech. Our results show that the usage of synthesized emotional speech in acoustic model training can significantly improve recognition of arousal from human speech in the challenging cross-corpus setting.

B. Schuller (✉) · Z. Zhang · F. Weninger
Institute for Human-Machine Communication,
Technische Universität München, 80290 München, Germany
e-mail: schuller@tum.de

F. Burkhardt
Deutsche Telekom Laboratories, Berlin, Germany

1 Introduction

In the last years, the field of computational paralinguistics has emerged from loosely connected research in the more traditional disciplines of speech and speaker recognition. It deals with the generic problem to determine long-term speaker traits (e.g., identity, personality, age or gender) and medium-term or short-term states (e.g., emotion, mood, sleepiness, or intoxication) from the speech signal by means of acoustic and linguistic analysis. Such generic speech and speaker classification is immediately connected to a variety of relevant applications in surveillance, including detection of emotional stress, sleepiness or intoxication in high-risk environments, and forensics, by identifying speakers in audio recordings from their traits, without requiring an explicit speaker model (Schuller et al. 2012). Yet, these capabilities are also useful for emulating 'social competence' in technical systems such as dialogue systems or robots, i.e., using signal processing and machine learning to react appropriately to dialogue partners with respect to their traits and states, by adjusting the discourse strategy, or aligning to the dialogue partner. Among the fields of research in speaker state analysis, emotion recognition can be considered one of the most mature, with a first comparative evaluation campaign, the INTERSPEECH Emotion Challenge, held in 2009 (Schuller et al. 2011).

Interestingly, such speaker states can be recognized automatically from speech features such as cepstral coefficients, and methodologies such as universal background models (UBMs) often used for speaker recognition (Schuller et al. 2011, 2012; Bone et al. 2011). This suggests that speaker

state variation poses a major challenge to speaker recognition systems. In fact, downgrades in speaker recognition performance in emotional conditions have been demonstrated repeatedly in the literature (Scherer et al. 2000; Wu et al. 2006; Li and Yang 2007). First studies in that direction suggested the inclusion of emotional speech in the enrollment procedure for speaker verification (Scherer et al. 2000), which is however hardly feasible in practical applications. In Wu et al. (2006) it has been proposed to normalize UBM scores by emotion category in order to avoid that emotional speech is rejected in a speaker verification system. Recent results suggest that in principle, speaker state variation can be thought of as a 'channel effect' that can be modeled, e.g., by latent factor analysis (Li et al. 2012). However, from this study is not clear whether this unsupervised channel compensation method is sufficient to counter the effects of emotional variation on speaker recognition performance. Yet, in all of these previous results we can find significant evidence for interdependencies between emotion and speaker recognition.

As a matter of fact, it appears that building emotion recognizers that can operate in a speaker-independent fashion in various acoustic environments is challenging (Schuller et al. 2011). One of the barriers to overcome before emotion recognition can be employed in real-life systems is the scarcity of labeled training data to develop automatic recognition systems (Zeng et al. 2009; Schuller et al. 2011). Indeed, it is a common belief in the field of pattern recognition that "there is no data like more data". Yet, in emotion recognition the research community is still lacking publicly available databases of large size, which is in stark contrast to the field of automatic speech recognition (ASR) where typical recognizers are trained on hundreds of hours of transcribed speech. Recently, several methods have been proposed to alleviate the data scarcity problem, such as combining multiple training databases (Lefter et al. 2010) or employing unsupervised learning strategies to make use of large quantities of unlabeled emotional speech (Zhang et al. 2011; Mahdhaoui and Chetouani 2009). Another approach to remedy the problem is the use of artificially generated speech, i.e., speech synthesis. If such data are suitable for training or adapting models for the recognition of human emotional speech, countless options open up: Not only could training data be generated in virtually infinite quantities, but emotional speech could be produced for different target groups (e.g., by varying parameters of the synthesizer corresponding to different age or gender), for various and also under-resourced languages, and fitting to the spoken content at hand. The latter could help for the design of dialogue systems with specific vocabulary, but could also be promising to address the challenge of text-independent emotion recognition: Assuming reliable ASR, one could first recognize the phonetic content, and then re-produce this content in various emotional facets for adaptation of acoustic emotion models. The general feasibility of this idea has been repeatedly demonstrated: For example, Microsoft's Kinect sensor uses synthesized user models to provide for different body shapes, postures, etc. Concerning the field of audio processing, in Lee and Slaney (2006) improved recognition of chords in music is enabled by synthesis of training material from symbolic music using various sound fonts (sets of instrument samples). Finally, in Schuller and Burkhardt (2010), we have achieved tentative results showing that using synthesized speech for training benefits emotion recognition from human speech in a pair-wise cross-database evaluation using the eNTERFACE (Martin et al. 2006) and EMO-DB (Burkhardt et al. 2005) corpora, i.e., training on one database and testing on the other. There, using synthesized speech for training could often outperform training with human speech.

This article aims at consolidating these promising results by providing an extensive empirical evaluation on eight human emotional speech databases in a cross-corpus scenario. This cross-corpus setting aims to put the evaluation closer to real-life applications where data from the exact application domain might not be available. In fact, there is a large variability among the labeling schemes, languages, types of emotion elicitation and associated application scenarios found in typical databases of human emotional speech. Not only does this cause a mismatch in the feature distribution between training and testing instances, but it also necessitates a coercion of the original continuous valued or categorical labels to a generic scheme. In this study, we opt for a rather coarse binary labeling with positive/negative arousal and valence, for compatibility with our earlier work on cross-corpus emotion recognition (Zhang et al. 2011). This dimensional, but discrete approach is chosen because on the one hand, most categorical emotion labels (such as the 'Big Six' emotions joy, sadness, anger, fear, surprise and disgust) can be expressed as points in the arousal-valence space (Russell 1980); on the other hand, we use classification instead of regression since the majority of the considered databases is annotated by emotion categories instead of a more fine-grained, continuous annotation—this is mostly due to the type of emotion elicitation used for creating these databases. In Sect. 2 we will describe the recording setups and label mappings of the human emotional speech databases in detail.

In this study, synthesized speech is generated by two different phonemization components, namely TXT2PHO and OpenMary, in combination with Emofilt and Mbrola, as will be laid out in Sect. 3. The 6 k space of acoustic features extracted by our open source Emotion and Affect Recognition (openEAR) toolkit in compliance with our earlier work, as well as the classification procedures and results, are presented in Sect. 4 before concluding with future perspectives in Sect. 5.

2 Eight human emotional speech databases

2.1 Synopsis

In our selection of human emotional speech databases for evaluation, we choose frequently used and publicly available ones. These will be briefly introduced below (in alphabetical order), describing their recording setup and labeling. While many of them provide audiovisual data and transcription, we only use the audio tracks in our analysis.

The *Airplane Behavior Corpus* (ABC) (Schuller et al. 2007) is an audiovisual emotion database. It is crafted for the special target application of public transport surveillance. It is based on induced mood by pre-recorded announcements on a simulated vacation flight, consisting of several scenes such as start, serving of wrong food, turbulences, falling asleep, conversation with a neighbor, or touch-down.

The *Audiovisual Interest Corpus* (AVIC) (Schuller et al. 2009a) is another audiovisual emotion corpus. It consists of spontaneous speech and natural emotion. In its scenario setup, a product presenter leads subjects through a commercial presentation. AVIC is labeled in "level of interest" (loi) 1–3 corresponding to boredom, neutral and interested.

The *Danish Emotional Speech* (DES) database (Engbert and Hansen 1996) contains professionally acted nine Danish sentences, two words, and chunks that are located between two silent segments of two passages of fluent text. Speech is expressed in five emotional states: anger, happiness, neutral, sadness, and surprise.

The *Berlin Emotional Speech Database* (EMO-DB) (Burkhardt et al. 2005) features professional actors expressing six prototypical emotional states (anger, boredom, disgust, fear, joy, neutral, and sadness) and an emotionally neutral state in ten German sentences without emotional connotations.

The eNTERFACE (eNTER) (Martin et al. 2006) corpus is a further public audiovisual emotion database. It consists of recordings of naive subjects from 14 nations speaking pre-defined spoken content in English. The subjects listened to six successive short stories eliciting a particular emotion out of angry, disgust, fear, happy, sadness, and surprise.

The *Belfast Sensitive Artificial Listener* (SAL) data is part of the final HUMAINE database. This database was created in a Wizard-of-Oz scenario simulating the interaction of a human with an emotionally sensitive and expressive virtual agent (SAL) whose part was taken by a human operator. There are four different agents representing emotional stereotypes (the quadrants of the valence-arousal space) to elicit varying emotional responses from the user. The data has been labeled continuously in real time with respect to valence and activation using a system based on FEELtrace (Cowie et al. 2000). The 25 recordings have been split into turns using an energy based Voice Activity Detection. Labels for each obtained turn are computed by averaging over the complete turn.

The *Speech Under Simulated and Actual Stress* (SUSAS) database (Hansen and Bou-Ghazale 1997) serves as a reference for the first recordings of spontaneous emotional speech. In addition to the challenges introduced by the spontaneity, the speech is partly masked by field noise in the chosen actual stress speech samples recorded in subject motion fear and stress tasks (rollercoaster and free fall situations). The SUSAS content is restricted to 35 English aircommands in the speaker states high stress, medium stress, neutral, screaming, and fear.

Finally, the *Vera-Am-Mittag* (VAM) corpus (Grimm et al. 2008) consists of recordings taken from a German TV talk show. The audio recordings were manually segmented to the utterance level, whereby each utterance contained at least one phrase. The labeling bases on a discrete five point scale for valence, activation, and dominance. The evaluator weighted estimator (annotator mean weighted by reliability) (Grimm and Kroschel 2005) is used to create a quasi-continuous annotation. In our experiments, we only consider the valence and activation dimensions, the latter being regarded as equivalent to arousal.

Further details on the corpora are summarized in Table 1 and found in Schuller et al. (2009b). Overall, these corpora cover a broad variety of data from acted speech (DES, EMO-DB) to induced emotions (ABC, eNTERFACE) to spontaneous emotions (AVIC, SAL, SUSAS, VAM), and from strictly limited textual context (ABC, DES, EMO-DB, SUSAS) to more variation (eNTERFACE) to full variance (AVIC, SAL, VAM). Three languages (English, German, and Danish) belonging to the same family of Germanic languages are contained. Furthermore, the speaker characteristics, the recording conditions, as well as the number of annotators vary greatly among these databases.

2.2 Mapping and clustering

Since the eight human speech databases are annotated in various emotion categories and continuous valued dimensions, we map the diverse emotion groups into the quadrants of the two-dimensional arousal-valence space as in Schuller et al. (2009b): arousal (i.e., high vs. low) and valence (i.e., positive vs. negative) in order to generate a unified set of labels that can be used for cross-corpus experiments. These mappings are not straight forward—we favor better balance among target classes. For corpora labeled in continuous valued dimensions, we discretize the labels into the four quadrants (q) 1–4 of the arousal-valence plane. The mappings are given in Tables 2 and 3 for arousal and valence, respectively.

Table 1 Overview of the selected emotion corpora (Lab.: labelers, Rec.: recording environment, f/m: (fe-)male subjects), synth.: synthesized

Corpus	Language	Speech	Emotion	# Arousal		# Valence		# All	h:mm	# m	# f	# Lab.	Rec.	kHz
				−	+	−	+							
Human Speech (HS)														
ABC	German	Fixed	Induced	104	326	213	217	430	1:15	4	4	3	Studio	16
AVIC	English	Free	Natural	553	2449	553	2449	3002	1:47	11	10	4	Studio	44
DES	Danish	Fixed	Acted	169	250	169	250	419	0:28	2	2	–	Studio	20
EMO-DB	German	Fixed	Acted	248	246	352	142	494	0:22	5	5	–	Studio	16
eNTER	English	Fixed	Induced	425	852	855	422	1277	1:00	34	8	2	Studio	16
SAL	English	Free	Natural	884	808	917	779	1692	1:41	2	2	4	Studio	16
SUSAS	English	Fixed	Natural	701	2892	1616	1977	3593	1:01	4	3	–	Noisy	8
VAM	German	Free	Natural	501	445	875	71	946	0:47	15	32	6/17	Noisy	16
Synthesized Speech (SS)														
OpenMary	German	Fixed	Synth.	280	350	420	210	630	0:33	4	3	–	–	22
TXT2PHO	German	Fixed	Synth.	280	350	420	210	630	0:33	4	3	–	–	16

Table 2 Mapping the classes of various databases to a binary arousal (High or Low)

Corpus	High	Low
Eight Human Speech Databases		
ABC	Aggressive, cheerful, intoxicated, nervous	Neutral, tired
AVIC	loi2, loi3	loi1
DES	Angry, happy, surprise	Neutral, sad
EMO-DB	Anger, fear, joy	Boredom, disgust, neutral, sadness
eNTERFACE	Anger, fear, happiness, surprise	Disgust, sadness
SAL	q1, q4	q2, q3
SUSAS	High stress, medium stress, screaming, fear	Neutral
VAM	q1, q4	q2, q3
Two Synthesized Speech Databases		
OpenMary/ TXT2PHO	Despair, fear, happiness, hot anger, joy	Boredom, neutral, sadness, yawning

Table 3 Mapping the classes of various databases to a binary valence (Positive or Negative)

Corpus	Positive	Negative
Eight Human Speech Databases		
ABC	Cheerful, neutral, intoxicated	Aggressive, nervous, tired
AVIC	loi2, loi3	loi1
DES	Happy, surprise, neutral	Angry, sad
EMO-DB	Neutral, joy	Anger, boredom, disgust, sadness, fear
eNTERFACE	Happiness, surprise	Anger, fear, disgust, sadness
SAL	q1, q2	q3, q4
SUSAS	Medium stress, neutral	High stress, screaming, fear
VAM	q1, q2	q3, q4
Two Synthesized Speech Databases		
OpenMary/ TXT2PHO	Happiness, joy, neutral	Boredom, despair, fear, hot anger, sadness, yawning

3 Emotional speech synthesis

We now describe our approach for synthesis of emotional speech that is used to augment the human training data as detailed above.

3.1 Overview

Speech synthesis is usually done in a two step approach. First, the text is analyzed by a natural language processing (NLP) module and converted into a phonemic representation aligned with a prosodic structure, which is then passed to a digital speech processing (DSP) component in order to generate a speech signal. Both of these sub modules can be influenced by the emotion modeling component. Generally, there exist several main approaches to model synthetic speech: parametric systems like articulatory and formant synthesis, data-based synthesis like diphone and non-uniform unit-selection synthesis and hybrid systems, e.g., Hidden Markov Model (HMM) synthesis. (Steidl et al. 2012) evaluated articulatory/parametric synthesis by humans and an automatic emotion classifier. We used diphone synthesis, because it can be seen as a good compro-

mise between the flexibility of parametric and the naturalness of data-based synthesis. We developed an emotional speech synthesis system on the basis of Mbrola (Dutoit et al. 1996). In order to increase the variation in the synthesized data used for training, we used two different phonemization components, namely TXT2PHO (Portele 1999) and Mary (Schröder and Trouvain 2003) for NLP. Emofilt (Burkhardt 2005) acts as an 'emotional transformer' between the phonemization (Text2Pho or Mary) and the speech generation component (Mbrola).

The simulation of emotions is achieved by a set of parametrized rules that describe manipulation of the following aspects of a speech signal: pitch changes, duration changes, voice quality (simulation of jitter and support of multiple voice quality databases), and articulation (substitution of centralized/decentralized vowels).

For the experiments at hand we synthesized the ten sentences of the Berlin Emotional Speech Database (cf. Sect. 2) with TXT2PHO as well as with Mary and simulated eight target emotions (happiness, joy, boredom, yawning, fear, despair, hot anger, sadness) plus a neutral state with Emofilt, using all seven German voices for Mbrola (four female and three male), thus getting 1 260 samples ($10 \times 2 \times 9 \times 7$, cf. Table 1).

The remainder of this section firstly describes the modification rules in general and then specifies which rules were applied for the emotional states. The rules were motivated by descriptions of emotional speech found in the literature; a review can be found in Burkhardt (2000). Before the rules are applied by Emofilt, the input phoneme chain gets syllablized by an algorithm based on sonority hierarchy. In addition, stressed syllables are identified as those that carry local pitch contour maxima (Burkhardt 2005).

3.2 Pitch modification methods

The following modifications are provided for pitch feature changes.

Pitch level The overall level of the F_0 contour can be shifted by multiplying all values with a rate factor (rate $= 0$ resembles no change). This means that high values undergo stronger changes than low values and was chosen to conform with the human logarithmic hearing.

Pitch range The pitch range change was motivated by the peak-feature model mentioned in Bergmann et al. (1988) and is achieved by a shift of each F_0 value by a percentile of its distance to the mean F_0 value of the last syllable. If range $= 0$, all values become the last syllable's mean value. The shifting corresponds to a contrast change in image processing. Note that Emofilt currently assumes its input to consist of one 'utterance' in the sense of a *short* part of speech that shall be uttered emotionally. This might lead to problems if several sentences are given as input, because utterance-global values like, e.g., the 'mean pitch value of last syllable' are currently computed only once for the whole of the input phoneme sequence.

Pitch variation A pitch variation on the syllable level is achieved by the application of the pitch range algorithm on each syllable separately. The reference value in this case is the syllable's mean pitch.

Pitch contour (phrase) The pitch contour of the whole phrase can be designed as a rising, falling or straight contour. The contours are parametrized by a gradient (in semitones/sec). As a special variation for happy speech, the 'wave model' can be used where the main-stressed syllables are raised and the syllables that lie equally distanced in between are lowered. It is parametrized by the maximum amount of raising and lowering and connected with a smoothing of the pitch contour, because all F_0 values are linearly interpolated.

Pitch contour (syllables) A rising, falling or level contour can be assigned to each syllable type. Additionally, the last syllable can be handled separately.

Duration modification methods The speech rate can be modified for the whole phrase, specific sound categories or syllable stress types separately by changing the duration of the phonemes (given as a percentile). If the duration is shorter than the frame rate as a consequence of a length reduction, the phoneme is dropped.

3.3 Voice quality modification methods

We developed a formant synthesis based approach to simulate different voice quality types based on the Klatt synthesizer (Burkhardt 2009), but could not use the synthesizer for the problem at hand because it lacks a database for general speech synthesis. Because with Mbrola the voice quality of the speech is fixed within the diphone inventory, we had to restrict ourselves to the two following rules:

Jitter In order to simulate jitter (fast fluctuations of the F_0 contour) the F_0 values can be displaced by a percentile alternating down and up.

Vocal effort For the German language, there exist two voice databases that were recorded in three voice qualities: normal, soft, and loud (see Schröder and Grice 2003). The change of voice quality can be applied to the whole phrase or specific syllable stress types only.

Table 4 Modification rule description for the used emotional categories, sorted into pitch, duration, voice quality and articulation changes

Category	Pitch	Duration	Voice quality	Articulation
Boredom	Lower level (80), lower range (50 %), lower variability (80)	General slower (120), accented syllables (140)	Soft vocal effort	Vowel target undershoot
Yawning	Raise level (200), falling contour (50)	Like boredom	Like boredom	Like boredom
Despair	Raise level (200), lower range (50), lower variability (90), falling contour on stressed syllables (20)	Slower (120)	Jitter (20)	–
Happiness	Wave model (120)	Slower (120), voiceless fricatives slower (150)	–	–
Sadness	Lower variability (80) and range (80), straight contour on stressed syllables, last syllable rising contour (10)	Slower (140)	Vocal effort soft, jitter (10)	Vowel target undershoot
Joy	Level raised (150), range broader (200), rising contour for stressed syllables (70)	Faster (70), voiceless fricatives slower (150)	Vocal effort loud	–
Fear	Level raised (200), range broader (160), falling phrase contour (10), straight contour for stressed syllables (70), last syllable rising contour (10)	Faster (70), longer pauses (210)	Jitter (5)	Vowel target undershoot
Hot anger	Level raised (150), range broader (200), more variability (130), level of stressed syllables raised (130), falling contour for stressed syllables (30)	Faster (70), vowels faster (70)	Loud vocal effort, jitter (5)	Vowel target overshoot

3.4 Articulation modification

As a diphone synthesizer has a very limited set of phoneme realizations and does not provide for a way to do manipulations with respect to the articulatory effort, the substitution of centralized vowels with their decentralized counterparts and vice versa is possible as a work-around to change the *vowel precision*. This operation was inspired by Cahn (1989).

3.5 Simulating emotional states

As stated in the introduction, emotional modeling is usually either done by a categorical system that distinguishes between a specific set of emotion categories like *anger* or *boredom*, or by the use of dimensions like arousal, valence or dominance. It is easy to derive appropriate acoustic modification rules for the arousal dimension, because both are directly related to muscle tension, but such derivations are considerably more difficult for the other dimensions. Therefore we model emotional states with a categorical system, although we realize that dimensional systems are more flexible and better suited to model the impreciseness of the 'real world', in which 'full blown' emotions such as the 'Big 6' rarely occur. Table 4 lists the modifications for the eight emotion categories used in the oncoming experiments. As stated above, they were inspired by a literature review, manually fine tuned and partly verified by perception experiments (Burkhardt and Sendlmeier 2000). Most modifications are parametrized by a rate value which is stated in parenthesis. Emofilt is freely available (http://emofilt.sourceforge.net) and the reader is invited to reproduce the simulations.

The modifications work are applied in a cascading fashion; this means, for example, that if the overall speech rate is faster (e.g. 70) and vowel durations are also faster (e.g. 70), the new duration for vowels is shortened to 49 % (70 % of 70 %). Of course the emotional expression that is generated by these rules is very prototypical and only one possibility to display the target emotions. In order to get a higher variety, and hence to generate even more training data for future experiments, it would be possible to randomly shift the parameters for the modifications slightly, or use the Emofilt 'graded-emotion' function which generates stronger or weaker versions of the modification rules.

For the purpose of uniting both human and synthesized speech in the training of a single classifier, we mapped the emotional categories from Table 4 into the quadrants of the arousal-valence plane (cf. Tables 2 and 3, respectively).

3.6 Evaluation of synthesis performance

We evaluated the performance of the emotional synthesis in a forced choice listening test with 20 participants. The four basic emotions anger, joy, sadness and fear plus neutral stimuli were generated by the rules described above. Overall, 55.6 % accuracy could be achieved by the human subjects in assigning the stimuli to emotion classes (recall of neutral:

Table 5 56 frame-wise Low-Level Descriptors (LLD) extracted from audio signals for emotion classification. T_0: periodic time

Feature group	Features in group
Raw signal	Zero-crossing-rate
Signal energy	Logarithmic
Pitch	Fundamental frequency F_0 in Hz via Cepstrum and Autocorrelation (ACF) Exponentially smoothed F_0 envelope
Voice quality	Probability of voicing ($\frac{ACF(T_0)}{ACF(0)}$)
Spectral	Energy in bands 0–250 Hz, 0–650 Hz, 250–650 Hz, 1–4 kHz 25 %, 50 %, 75 %, 90 % roll-off point, centroid, flux, and rel. pos. max./min.
Mel-spectrum	Band 1–26
Cepstral	MFCC 0–12

Table 6 39 functionals applied to LLD contours

Functionals	#
Respective rel. position of max./min. value	2
Range (max./min.)	1
Max. and min. value—arithmetic mean	2
Arithmetic mean, Quadratic mean, Centroid	3
Number of non-zero values	1
Geometric, and quadratic mean of non-zero values	2
Mean of absolute values, Mean of non-zero abs. values	2
Quartiles and inter-quartile ranges	6
95 % and 98 % percentile	2
Std. deviation, variance, kurtosis, skewness	4
Zero-crossing rate	1
# of peaks, mean dist. btwn. peaks, arth. mean of peaks, arth. mean of peaks—overall arth. mean	4
Linear regression coefficients and error	4
Quadratic regression coefficients and error	5

60 %, anger: 55 %, joy: 35 %, sadness: 50 %, fear: 78 %). Joy was often confused with anger (28 %).

4 Experimental setup and results

We now proceed to describe the acoustic features extracted from the human and synthesized speech, as well as the details of the classifier setup and evaluation procedure. Note that in order to obtain emotion models from synthesized speech, we take a data-based approach as for the human speech, using the same set of acoustic parameters for compatibility; the classifier then learns automatically how our generation rules affect these parameters.

4.1 Acoustic features and classification

As acoustic features for classification, we employ the 'emolarge' set provided by our open source openEAR toolkit (Eyben et al. 2009). This set contains 6 552 ($39 \times 56 \times 3$) features obtained by systematically applying 39 statistical turn-level functionals to 56 frame-level Low-Level Descriptors (LLDs) and their first and second order discrete derivatives. The turn-level functionals serve to capture temporal variation of the LLDs in feature vectors of constant dimension independent of the utterance length. This kind of static modeling is often superior to dynamic modeling in emotion recognition (Schuller et al. 2011). The LLDs extracted (cf. Table 5) include prosodic features (zero-crossing rate, pitch, energy) as well as probability of voicing, spectral and cepstral features, all of which are commonly used in acoustic emotion recognition (Schuller et al. 2011). The considered spectral sub-bands emphasize on energy in the F_0 region (0–250 Hz), the first formant (F_1, 250–650 Hz, mostly vowels), both F_0 and F_1 (0–650 Hz), and higher order formants (1–4 kHz, mainly consonants). The LLD derivatives are computed by 'delta regression' as in the Hidden Markov Model Toolkit (HTK) (Young et al. 2006). Table 6 summarizes the statistical functionals which were applied to the LLDs shown in Table 5. These functionals cover extrema and moments of the LLD contours, outlier robust distribution features such as percentiles and inter-quartile ranges, functionals related to the occurrence of peaks, and parameters of linear and quadratic regression on the contours.

As classifier, we selected linear kernel Support Vector Machines (SVM) due to their good generalization properties in high dimensional feature spaces; at present, they are likely the most frequently used classifier in emotion recognition (Schuller et al. 2011). We chose an SVM complexity constant of 0.05, and binary class discrimination based on Sequential Minimal Optimization (Platt 1999). The implementation in the Weka toolkit (Hall et al. 2009) was used for maximum reproducibility. Note that we train separate classifiers for each dimension (arousal and valence).

4.2 Evaluation protocol

We evaluated our experiments in terms of unweighted accuracy (UA), i.e., the unweighted average of the recalls of the 'positive' and 'negative' classes, which has been the official competition measure of the first of its kind INTERSPEECH 2009 Emotion Challenge (Schuller et al. 2011). UA is evaluated separately for both binary arousal and valence discrimination. In a first baseline experiment, we employ pair-wise cross-corpus training and testing on the eight databases of human emotional speech (HS), i.e., for each test database, each of the remaining seven databases is used once as training set. This protocol results in $56 = 7 \times 8$ cross-corpus classifications for each dimension (arousal and va-

Table 7 Mean and maximum Unweighted Accuracy (UA) for varying training data in cross-corpus binary arousal and valence classification on eight test databases of human speech. Training with HS: single databases of human speech, SS: two databases of synthesized speech, HS + SS: all possible permutations of human speech and synthesized speech databases. EMO: EMO-DB

Train	Test								Mean
	ABC	AVIC	DES	EMO	eNTER	SAL	SUSAS	VAM	
Arousal									
Mean UA [%]									
HS	60.8	58.0	71.7	69.6	59.6	59.4	56.2	65.4	62.6
SS	65.7	66.8	69.9	77.3	55.9	57.2	59.7	66.2	64.8
HS + SS	64.0	61.7	74.1	76.8	59.0	59.7	58.5	67.6	65.2
Max UA [%]									
HS	66.1	64.2	80.3	71.0	64.0	64.7	60.6	73.2	69.3
SS	66.7	66.8	70.1	79.8	57.7	57.9	61.2	67.5	66.0
HS + SS	69.1	67.0	79.7	84.0	61.4	62.0	63.2	72.9	69.9
Valence									
Mean UA [%]									
HS	56.5	56.4	53.6	54.0	52.8	52.2	49.1	51.4	53.3
SS	48.3	51.8	55.0	54.2	56.9	50.8	38.5	47.7	50.4
HS + SS	55.2	59.1	54.2	54.5	54.1	52.1	42.1	49.1	52.6
Max UA [%]									
HS	60.6	66.1	57.7	58.8	58.4	57.4	56.0	58.5	59.2
SS	48.4	53.9	58.3	55.8	58.1	51.5	38.5	50.0	51.8
HS + SS	59.4	66.3	58.7	59.3	56.9	55.4	45.0	57.3	57.3

lence). Then, to assess the suitability of synthesized training data for analyzing human speech, we repeat the experiment by training with each of the two sets of synthesized emotional speech (TXT2PHO and Mary, SS), and testing on each of the human speech databases ($16 = 2 \times 8$ evaluations per dimension). Finally, to investigate the benefit of joint training with human and synthesized speech (HS + SS), we consider all $16 = 2 \times 8$ combinations of human and synthesized data sets for training, and evaluate on each of the seven human databases not found in the training data. This last experiment results in $112 = 16 \times 7$ combinations of training and test data per dimension. To generally enhance performance in cross-corpus emotion recognition, we employed feature standardization through linear scaling of each feature to zero mean and unit variance per corpus. This helps to reduce trivial cases of feature mismatch due to different microphone-to-mouth distances etc. We have shown that this 'z-normalization' strategy achieves better performance in cross-corpus emotion recognition than simple mean normalization or linear scaling to a certain feature range (Zhang et al. 2011).

4.3 Results and discussion

Table 7 shows the unweighted accuracies (UA) obtained for the two-class arousal and valence classification tasks when following the three above-mentioned evaluation protocols (HS, SS, and HS + *SS*). In summary, the results of the baseline experiment (cross-corpus training and testing on human speech, HS) corroborate the results of other cross-corpus emotion recognition studies, indicating that while arousal classification is somewhat stable, cross-corpus valence classification cannot be performed robustly using the acoustic features used in this study. In fact, results are often found below chance level UA (50 %) for valence. Furthermore, in arousal classification, we find that testing on highly prototypical emotions (e.g., EMO-DB or DES) generally leads to higher performance than testing on spontaneous emotions (e.g., in the SAL or SUSAS databases), which is expected. A notable exception from this general pattern is the comparably high UA (73.2 %) when testing on the VAM database; this can be attributed to the fact that while the emotions in this database are naturalistic, the talk show recording scenario is much more likely to elicit strong emotions than, e.g., the human-computer interaction scenario in the SAL database.

Comparing synthesized and human speech for training purposes, it is highly interesting that in the SS scenario (training on synthesized speech only) the average UA of binary arousal classification across all test databases (64.8 %) is significantly higher than in the HS scenario (training on human speech only, 62.6 %). In contrast, for valence, the performance of synthesized training data (50.4 % average UA) is observed significantly below the one of human data (53.3 % UA), and is near chance level UA (50 %). This in-

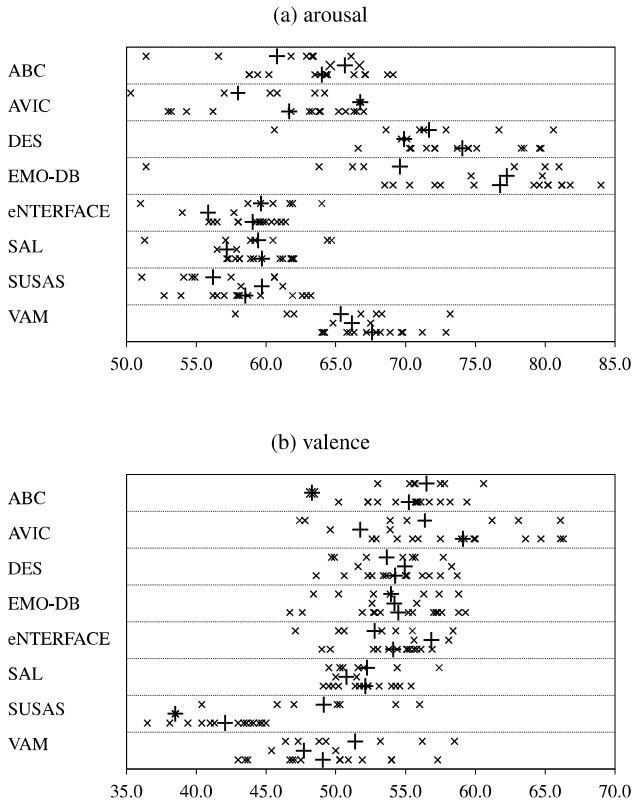

Fig. 1 Distributions of Unweighted Accuracies (UA) for cross-corpus binary arousal/valence classification of eight test databases: Single–database classifiers are depicted by crosses and the average of single–database classifiers by a plus sign. For each test database, the top row corresponds to training on human speech, the middle row to training with synthesized speech, and the bottom row to merging of one human database with one synthesized speech database

dicates a large mismatch between the features of the synthesized speech that is supposed to express negative valence, and the human utterances actually corresponding to negative valence (or being perceived as such by the human labelers). Generally, this result corroborates the well-known fact that variation of valence can only partly be modeled, and hence be generated, by variation of acoustic features.

Third, when considering the performance of merging 'HS' and 'SS' data in training (65.2 % average UA), we find a slight enhancement over training with only synthesized speech (64.8 % UA), and a significant gain of 3 % absolute across all databases with respect to training with only human speech (62.6 % UA). This performance enhancement by agglomeration of HS and SS training data is to be expected, since the performances of HS and SS on the individual databases suggest that they may have complementary strengths when used for model training (cf. Fig. 1 and the discussion below).

Besides these promising improvements in a large scale perspective, without doubt there are several noteworthy singular results which should not be overlooked. For example, we see that synthesized speech prevails over human training data when testing on the EMO-DB (77.3 % on average); this is probably a consequence of text dependency, due to the fact that the sentences from the EMO-DB were used to synthesize the emotional training speech. In the same vein, the overall best result on the EMO-DB (84.0 %) is achieved by joining the DES database of acted emotions with synthesized speech from Mary. This, however, should not suggest that synthesized speech is only useful when the textual content matches: On the spontaneous, free-text AVIC database, both variants of synthesized speech deliver 8.8 % absolute higher UA (66.8 %) than human speech on average (58.0 %), and are observed above the best single human speech database—which is, interestingly, the acted DES database (64.2 % UA). Looking at the maximum UA values in Table 7, we find other surprising cases of databases that seem to 'match' particularly well. For example, the best result in cross-corpus arousal classification on the DES database is achieved by using the spontaneous VAM database for training (80.3 % UA); even more notably, the same holds vice versa (training on DES and testing on VAM delivers the best single result of 73.2 % UA on VAM). This apparent similarity of DES and VAM is also reflected in the fact that both are 'equally hard' to classify by synthesized speech as opposed to human speech (max. UA of 70.1/80.3 % for DES, max. UA of 67.5/73.2 % for VAM). The latter also indicates that the evident mismatch between the synthesized speech and DES is not simply caused by different languages (Danish/German): VAM is in German as is the synthesized speech.

To give an overview of the performance and its variability of the various training and testing permutations, we visualize the distributions of the UA for the three kinds of training scenarios in Fig. 1 as one-dimensional scatter plots. For each human testing database, the top row shows results obtained by human speech training sets, the middle row corresponds to synthesized speech training sets, and the bottom row refers to training sets obtained by merging one human database with one synthesized speech database. The 'plus' symbols indicate the average performance per row. From Fig. 1, it is obvious that training with single databases of human speech results in greatly varying performance. This effect is most visible for the acted test databases (DES and EMO-DB), where the UA of binary arousal classification with training on human speech ranges from 60.6 % to 80.6 % (DES), and from 51.4 % to 81.0 % (EMO-DB). The latter is somewhat expected, since for databases with limited variation of content, there is a larger chance that some training databases will strongly 'mismatch' the testing data. We can see that especially for arousal classification (Fig. 1a), this variability can be partly compensated by adding synthetized data to the training set; this effect is clearly visible for all but the AVIC and SUSAS databases. For valence

classification (Fig. 1b), we cannot observe such decreases in variability, which can be attributed to the generally lower classification performance which is often near chance level.

5 Conclusions

In this article, we have presented a large scale study on the suitability of synthesized speech for model training in cross-corpus emotion recognition. We have proposed a label mapping and evaluation framework for this challenging cross-corpus scenario which involves various labeling schemes and recording conditions including different types of emotion elicitation. Investigating the performance of $56 + 16 + 112 = 184$ different combinations of human and synthesized speech in binary arousal and valence classification of eight popular human speech databases, we have found that combining human and synthesized speech increases the expected performance while decreasing the performance variability caused by training with 'matching' or 'mismatching' human speech databases. Furthermore, in many cases the training on synthesized speech alone has been shown to be at least competitive with training on disjoint human speech databases. The fact that we could not observe these trends for cross-corpus valence classification, and the generally disappointing performance in this task, show the difficulty not only of building generalizing models for acoustic valence classification, but also the difficulty of synthesizing speech that matches the human perception of positive/negative valence.

Overall, we believe that our results can add a new argument to the ever-lasting debate in the field of affective computing, whether to prefer data with controlled variation and stable ground truth (such as acted data) or data collected 'in the wild' that is subsequently annotated by human observers: Interestingly, based on our results, we can recommend the usage of synthesized data generated with full control of the 'emotional variation' for 'bootstrapping' acoustic models to be deployed in real-life emotion recognition systems, as this procedure evidently benefits classification of arousal on all of the four databases of natural emotion considered. This is all the more interesting as the problem of emotional speech synthesis appears far from being solved: A perception study revealed significant mismatch between the emotional categories associated with the acoustic parameters of the synthesized speech by humans, and the intended emotional categories.

Starting from the results presented in this article, future work should address using multi- or cross-lingual speech synthesis methods to benefit cross-lingual emotion recognition, and to develop synthesis methods to simulate different target groups of computer users, from children (e.g., the FAU Aibo Corpus used for the first INTERSPEECH Emotion Challenge) to the elderly, or even pathologic voices. If a meaningful synthesis of such voices can be established, it could be a major step forward to the generalization of emotion recognition to target groups which are nowadays overlooked by the lion's share of research in the (certainly justified) quest for stable results in 'controlled' evaluation scenarios involving healthy adult speech.

Besides, it is evident that the methodologies which we employ for emotional speech analysis and synthesis differ considerably in the signal modeling: Analysis is based on brute forcing of statistical functionals of generic acoustic features while synthesis involves a multi-stage approach partly using hand-crafted features and expert rules. Thus, we expect future research to provide insight into the fundamental question of whether we should continue working towards 'bridging the gap' between analysis and synthesis—by unifying modeling techniques to avoid first generating waveform data from one model and then extracting features to train another model, instead of directly adapting the parameters, as was proposed, e.g., in Li and Yang (2007) for emotional speaker recognition—or, whether complementary approaches for analysis and synthesis are needed.

Finally, one should continue aiming at an optimal exploitation of the obvious interdependencies between speaker states and traits recognition, for example, emotion recognition and speaker recognition. Using emotion specific background models, feature mapping, emotional factor compensation and so forth can be seen as a starting point to exploit emotion recognition for robust speaker identification. Into the other direction, for example, one could start by exploring multi-task learning of speaker and emotion identification or similar techniques, in order to allow semi-supervised adaptation to the differences in how emotion manifests in the vocal expression of the individual.

Acknowledgements The authors would like to thank Florian Eyben for helpful discussions. Zixing Zhang is supported by the Chinese Research Council.

References

Bergmann, G., Goldbeck, T., & Scherer, K. R. (1988). Emotionale Eindruckswirkung von prosodischen Sprachmerkmalen. *Zeitschrift für experimentelle und angewandte Psychologie*, *35*, 167–200.

Bone, D., Black, M. P., Li, M., Metallinou, A., Lee, S., & Narayanan, S. (2011). Intoxicated speech detection by fusion of speaker normalized hierarchical features and GMM supervectors. In *Proc. of INTERSPEECH*, Florence, Italy (pp. 3217–3220).

Burkhardt, F. (2000). *Simulation emotionaler Sprechweise mit Sprachsynthesesystemen*. Aachen: Shaker Verlag.

Burkhardt, F. (2005). Emofilt: The simulation of emotional speech by prosody transformation. In *Proc. Interspeech 2005*, Lisbon, Portugal.

Burkhardt, F. (2009). Rule-based voice quality variation with formant synthesis. In *Proc. Interspeech 2009*.

Burkhardt, F., & Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proc. of the ISCA workshop on speech and emotion*.

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

Cahn, J. E. (1989). The affect editor. *Journal of the American Voice I/O Society*, *8*, 1–19.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). Feeltrace: an instrument for recording perceived emotion in real time. In *Proceedings of the ISCA workshop on speech and emotion*, Newcastle, Northern Ireland (pp. 19–24).

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vreken, O. (1996). The MBROLA project: towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In *Proc. ICSLP*.

Engbert, I. S., & Hansen, A. V. *Documentation of the Danish emotional speech database des*. Tech. rep., Center for PersonKommunikation, Aalborg University, Denmark (2007). http://cpk.auc.dk/~tb/speech/Emotions/. Last visited 11/13/2007.

Eyben, F., Wöllmer, M., & Schuller, B. (2009). openEAR—introducing the Munich open-source emotion and affect recognition toolkit. In *Proc. affective computing and intelligent interaction (ACII)*, Amsterdam, The Netherlands. New York: IEEE.

Grimm, M., & Kroschel, K. (2005). Evaluation of natural emotions using self assessment manikins. In *Proc. of ASRU* (pp. 381–385).

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The "Vera am Mittag" German audio-visual emotional speech database. In *Proc. of the IEEE international conference on multimedia and Expo (ICME)*, Hannover, Germany (pp. 865–868).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*, 10–18.

Hansen, J., & Bou-Ghazale, S. (1997). Getting started with susas: a speech under simulated and actual stress database. In *Proc. EUROSPEECH-97*, Rhodes, Greece (Vol. 4, pp. 1743–1746).

Lee, K., & Slaney, M. (2006). Automatic chord recognition from audio using a supervised HMM trained with audio-from-symbolic data. In *Proceedings of the 1st ACM workshop on audio and music computing multimedia table of contents*, Santa Barbara, CA, USA (pp. 11–20). New York: ACM.

Lefter, I., Rothkrantz, L. J. M., Wiggers, P., & van Leeuwen, D. A. (2010). Emotion recognition from speech by combining databases and fusion of classifiers. In *Proc. of text, speech and dialogue*, Berlin, Germany.

Li, D. D., & Yang, Y. C. (2007). Affect-insensitive speaker recognition by feature variety training. In A. Paiva, R. Prada, & R. W. Picard (Eds.), *Proc. of ACII. Lecture notes in computer science: Vol. 4738* (pp. 743–744). Berlin: Springer.

Li, M., Metallinou, A., Bone, D., & Narayanan, S. (2012). Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling. In *Proc. of ICASSP*, Kyoto, Japan (pp. 1937–1940).

Mahdhaoui, A., & Chetouani, M. (2009). A new approach for motherese detection using a semi-supervised algorithm. In *Proc. of IEEE international workshop on machine learning for signal processing (MLSP)* (pp. 1–6).

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

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: support vector learning* (pp. 185–208). Cambridge: MIT Press.

Portele, T. TXT2PHO—a TTS front end for the German inventories of the MBROLA project (1999). http://www.sk.uni-bonn.de/forschung/phonetik/sprachsynthese/txt2pho.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Scherer, K. R., Johnstone, T., Klasmeyer, G., & Bänziger, T. (2000). Can automatic speaker verification be improved by training the algorithms on emotional speech. In *Proc. of ICSLP*, Beijing, China (pp. 807–810).

Schröder, M., & Grice, M. (2003). Expressing vocal effort in concatenative synthesis. In *Proc. 15th international conference of phonetic sciences*, Barcelona, Spain (pp. 2589–2592).

Schröder, M., & Trouvain, J. (2003). The german text-to-speech synthesis system MARY: a tool for research, development and teaching. *International Journal of Speech Technology* 365–377.

Schuller, B., & Burkhardt, F. (2010). Learning with synthesized speech for automatic emotion recognition. In *Proc. ICASSP 2010*, Dallas, TX, USA (pp. 5150–5153).

Schuller, B., Wimmer, M., Arsic, D., Rigoll, G., & Radig, B. (2007). Audiovisual behavior modeling by combined feature spaces. In *Proc. ICASSP 2007* (Vol. II, pp. 733–736). New York: IEEE Press, Honolulu, Hawaii, USA.

Schuller, B., Müller, R., Eyben, F., Gast, J., Hörnler, B., Wöllmer, M., Rigoll, G., Höthker, A., & Konosu, H. (2009a). Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image and Vision Computing*, *27*(12), 1760–1774. Special issue on visual and multimodal analysis of human spontaneous behavior.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., & Wendemuth, A. (2009b). Acoustic emotion recognition: a benchmark comparison of performances. In *Proc. IEEE ASRU*, Merano, Italy (pp. 552–557).

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Communication*, *53*, 1062–1087.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2012, to appear). Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech and Language* (Special issue on paralinguistics in naturalistic speech and language) (2012) 39 p. http://dx.doi.org/10.1016/j.csl.2012.02.005.

Steidl, S., Polzehl, T., Bunnell, H. T., Dou, Y., Muthukumar, P. K., Perry, D., Prahallad, K., Vaughn, C., Black, A. W., & Metze, F. (2012). Emotion identification for evaluation of synthesized emotional speech. In *Proc. of speech prosody*.

Wu, W., Zheng, T. F., Xu, M. X., & Bao, H. J. (2006). Study on speaker verification on emotional speech. In *Proc. of INTERSPEECH*, Pittsburgh, PA (pp. 2102–2105).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. (2006). *The HTK book, for HTK version 3*, 4th ed. Cambridge: Cambridge University, Engineering Department.

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

Zhang, Z., Weninger, F., Wöllmer, M., & Schuller, B. (2011). Unsupervised learning in cross-corpus acoustic emotion recognition. In *Proc. IEEE automatic speech recognition and understanding workshop (ASRU)*, Big Island, HI, USA.