Estimation of Emotional States Enhanced by A Priori Knowledge

Branimir Dropuljić, Siniša Popović, Davor Petrinović, Krešimir Ćosić
Faculty of Electrical Engineering and Computing, University of Zagreb
Zagreb, Croatia
branimir.dropuljic@fer.hr, sinisa.popovic@fer.hr, davor.petrinovic@fer.hr, kresimir.cosic@fer.hr

***Abstract*—This paper presents an improvement of conventional supervised-learning emotional state estimation in the form of dimensional valence-arousal values. In the proposed approach, outputs of the conventional estimator are additionally adapted using *a priori* knowledge about valence-arousal relations, which is extracted from the estimator's training set. Different approaches to *a priori* knowledge modeling have been undertaken: (a) single integral model over valence-arousal space, and (b) integration of multiple models that represent different discrete emotions in the valence-arousal space, specifically happiness, sadness, fear, anger and neutral state. This emotion estimation approach has been applied to conventional valence-arousal estimation from acoustic speech features based on support vector machines, using data from Croatian emotional speech corpus. Improvement of the results has been demonstrated.**

Keywords—emotional state estimation; a priori knowledge; valence-arousal; acoustic speech features; support vector machines

I. INTRODUCTION

The merging between computer technologies and cognitive and emotional issues results in new multidisciplinary research areas like affective computing, human-computer interaction, cognitive infocommunications, etc. [1]. Automated emotion recognition systems are nowadays applied in many domains like safe driving system, e-learning, call centers, etc. in order to further enhance human-computer communication. Large challenges lie in emotionally based computing systems in high-risk operations like medical and psychology treatments. Application of methods for prevention and treatment of stress related disorders were presented in [2].

Typical way to estimate an emotional state is to extract features of a raw signal, recorded as a response of a person to emotionally rich stimulus and to perform an estimation using one of machine learning methods. Various modalities like physiology, acoustic and linguistic aspects of speech, facial expressions, etc. can be acquired, as emotions are inherently multimodal [3]. Different techniques like Gaussian mixture models (GMM), hidden Markov models (HMM), artificial neural networks (ANN), support vector machines (SVM), etc. have been applied for automatic emotion recognition. Current trend in affective computing society is continuous estimation of emotion dimensions, e.g. *valence* and *arousal*, what is foundation for real-time emotion recognition applications [4].

One of the contemporary approaches for estimation of emotion dimensions uses support vector regression (SVR) method [5]. Estimation was performed on emotion dimensions *valence*, *activation* (equivalent to *arousal*) and *dominance*. Furthermore, classification of discrete emotional states: anger, fear, happiness, neutral and sadness using SVMs obtained high classification accuracy in [6]. Research proposed in this paper is based on the same methods as in [5] and [6], with the focus on additional contributions that may improve the estimation results. Estimation has been performed using acoustic features, which were extracted from emotional speech utterances that were collected and annotated within the Croatian emotional speech corpus.

Enhanced method for estimation of valence and arousal is proposed in this paper. The general idea is similar to a *maximum a posteriori* (MAP) estimation, i.e. a regularization of the *maximum likelihood* (ML) estimation, using *a priori* distribution over valence-arousal space. Depending on the uncertainty of valence-arousal estimate, this estimate is shifted in valence-arousal space toward more probable valence-arousal values in the prior model. A MAP strategy has already been applied for emotion recognition tasks, for example within naive Bayes classifiers for discrete emotions [7]. On the contrary, this paper describes prior knowledge models in dimensional valence-arousal space, as well as their application to adapt the outputs of any type of valence-arousal estimator.

II. EMOTIONAL SPEECH CORPUS

Croatian emotional speech corpus (CrES) was collected and emotionally annotated from various prerecorded sources, like Internet or movies. A detailed description of building the initial version of the corpus was presented in [8]. In this paper, an upgraded version is used, which contains total of 1140 utterances from 341 different male and female speakers with the total duration of approximately 85 minutes. Utterances were initially categorized into five emotion categories: *happiness*, *sadness*, *anger*, *fear* and *neutral state*, based on subjective opinion of people involved in the collection phase. Distribution of collected utterances per emotions is presented in Table 1.

The annotation process was organized and performed in collaboration with Departments of Psychology and Phonetics, at University of Zagreb. A total of 109 undergraduate students,

978-1-4799-1546-0/13/$31.00 ©2013 IEEE

aged 18 to 24, took part in this experiment as annotators. Each annotator's task was to listen to one part of the corpus and to give assessment of emotional states of a speaker in each recording. Entire corpus was divided into five equal parts; each consisted of 228 utterances with 5 discrete emotions included in randomized order. Annotators were divided into two groups: one group for annotation of discrete emotions and the other one for annotation of emotion dimensions. In this way, at least 10 annotations were obtained for each utterance for both, discrete and dimensional representations of emotions.

In assessing discrete emotions, students had to annotate speaker's emotional state by choosing between one of five emotional classes. They also had to express the level of their certainty about the annotation. Furthermore, they had to designate whether their assessment was established predominantly from the acoustic speaker's emotional expression, from the semantic expression, or equally from both (in 5 steps). Students who graded emotion dimensions had to annotate valence and arousal using 9-level scales [9]. Information about the certainty and the acoustic/semantic expression was provided the same way as for discrete emotions.

After collecting all annotations, some utterances were removed. Discrete emotion annotations were analyzed according to the following relation:

$$I(e) = \frac{1}{A}\sum_{a=1}^{A} rel(a)\cdot sel(a,e), \quad e = 1,...,5, \tag{1}$$

where $I(e)$ represents intensity for each of 5 discrete emotions for each utterance, and is calculated as a weighted sum of A annotations. It should be noted that A varies through utterances, but at least 10 annotations are provided for each utterance ($A \geq 10$). For each annotation, $sel(a,e)$ is either 0 or 1, depending on a^{th} annotator's unique selection of one of the 5 emotional classes. Annotations are additionally weighted by relevance factor $rel(\text{a}) = cer(\text{a})\cdot ac(\text{a})$, which consists of the certainty level of the annotator $cer(\text{a})$ and annotator's opinion of the acoustic richness in the expression $ac(\text{a})$. Each of these tree factors has the value in the range of [0:1]. For each of the annotated utterances, the dominant emotion e_{max} is determined, that maximizes $I(e)$ for e = 1,..,5. Finally, utterances were filtered in accordance with two criteria based on the established dominant emotion. The first is the *agreement* criterion that is fulfilled if at least 50% of annotators choose exactly the established dominant emotion e_{max} with the relevance of at least ½. The second *prevalence* criterion checks whether the emotion with the second maximal value is at least 33% below the dominant emotion. Only the utterances fulfilling both criteria were kept for the further analysis. The final structure of the Croatian emotional speech corpus for discrete emotions can be seen in Table 1. Besides reduction of utterances during the filtering process, some utterances were classified differently during the annotation process. Most of these transitions between emotion classes occurred toward *anger*, as it can be seen in Table 1. Final *valence* and *arousal* labels of the corpus utterances will be described in the next section.

TABLE I. EMOTION ANNOTATION OF THE CORPUS UTTERANCES

Emotion	Number of Utterances per Emotion	
	Collection phase	***Annotation phase***
Happiness	249	179
Sadness	205	183
Fear	199	142
Anger	287	303
Neutral state	200	200
Total	1140	1007

III. A PRIORI KKOWLEDGE: APPLICATION TO EMOTION DIMENSIONS

As stated in [10], "*A priori* knowledge is knowledge that rests on *a priori* justification. *A priori* justification is a type of epistemic justification that is, in some sense, independent of experience." In the context of estimation process, *a priori* model can be applied to improve regular parameter estimation. For N parameters to be estimated, model proposed in this paper contains probabilities of all combinations of parameter values in N-dimensional space. Specifically, *a priori* probabilistic model that is based on GMM is described. Probability density function (PDF) of the model is calculated as a normalized sum of S single Gaussians:

$$P(o) = \frac{1}{S}\sum_{s=1}^{S} g(o \mid \mu_s, \Sigma_s), \tag{2}$$

where o is observation vector, i.e. N-dimensional vector of input values. Single Gaussians $g(o|\mu_s,\Sigma_s)$, for each sample $s = 1,\ldots,S$, are defined by model parameters μ_s and Σ_s, where μ_s represents N-dimensional centroid vector of the Gaussian density and Σ_s represents full-rank $N\times N$ covariance matrix.

Emotional state estimation in the form of emotion dimensions *valence* and *arousal* can be viewed as 2-parameter estimation problem (N = 2). *A priori* model given in (2) was built from annotations of S = 1007 emotionally annotated samples, i.e. corpus utterances. Centroid vector μ_s and covariance matrix Σ_s of each Gaussian were determined by iterative *expectation-maximization* (EM) algorithm based on synthetic samples, which were generated from $A \geq 10$ different *valence* and *arousal* annotations of each utterance. Synthetic sample generation process, which is illustrated in Fig. 1., was created for each annotation of each utterance in order to address the problem of limited amount of annotations per utterance and potential problem of model under-training, as well as problems related to discrete (integer) nature of input annotations. Therefore, for each individual annotation a= 1,…,A, ten synthetic samples (B = 10) were generated in valence-arousal plane that were randomly spread within the circle, centered at valence-arousal values $\mu_{val}(a)$ and $\mu_{ar}(a)$ of annotation a, with radius $R(a)$ defined as:

$$R(a) = R_{\min} + (R_{\max} - R_{\min})\cdot(1 - cer(a)), \tag{3}$$

where *cer*(*a*) is certainty described in section II. Minimal and maximal possible radii R_{min} and R_{max} were experimentally set to 0.25 and 1, which reflect maximal (100%) and minimal (0%) possible annotator's certainty, respectively. In this way, annotations with lower certainty will have more spread out synthetic samples in the valence-arousal space and will therefore contribute less when building a Gaussian of a particular utterance than the higher certainty annotations. Each utterance sample *s* was represented as ordered pair (**VAL**, **AR**) where **VAL** and **AR** are $A \times B$ matrices, with $a = 1,\ldots,A$ and $b = 1,\ldots,B$:

$$\mathbf{VAL}(a,b) = \mu_{val}(a) + r(a,b)\cdot\cos\left(\varphi(a,b)\right), \quad (4)$$

$$\mathbf{AR}(a,b) = \mu_{ar}(a) + r(a,b)\cdot\sin\left(\varphi(a,b)\right). \quad (5)$$

The radius $r(a,b)$ of the synthetic sample was randomly selected, with normal distribution assumption, from the range of $[0{:}R(a)]$, while angle $\varphi(a,b)$ of the synthetic sample was randomly selected from the range of $[0{:}2\pi]$, with uniform distribution assumption. Parameters μ_s and Σ_s of the Gaussian of each utterance sample *s* were determined from (**VAL**, **AR**), i.e. from $A \times B$ synthetic samples, performing EM algorithm. The total PDF that represents *a priori* knowledge model of the entire corpus (Fig. 2.) is constructed according to (2), as a normalized sum of *S* single Gaussians, where each Gaussian corresponds to individual utterance.

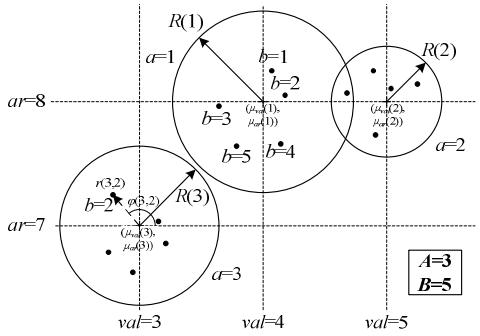


Fig. 1. Illustration of a process for synthetic sample generation. Example for $A = 3$ and $B = 5$, where *val* = valence and *ar* = arousal.

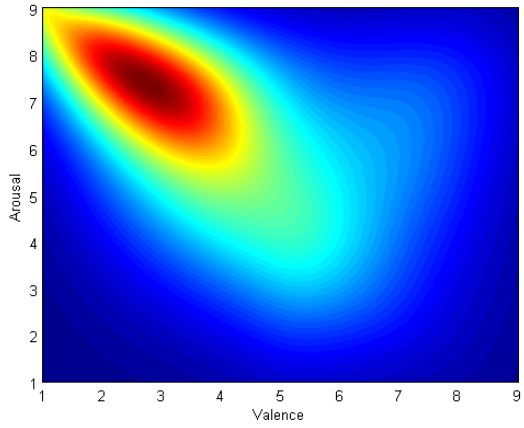


Fig. 2. The PDF of *a priori* valence-arousal model.

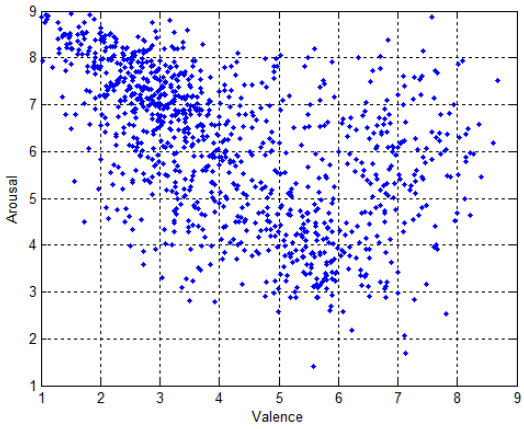


Fig. 3. Distribution of 1007 utterance samples from the corpus.

Each sample *s*, i.e. utterance of the corpus, is emotionally represented as the centroid μ_s of the defined single Gaussian. The distribution of the utterance samples in valence-arousal space is presented in Fig. 3.

IV. Emotional State Estimation Methodology

Estimation of emotional states was performed with conventional methodology first. Support vector regression (SVR) was used to estimate emotion dimensions. Estimations were based on a vocal emotional expression, i.e. an acoustic feature set was extracted from the corpus utterances.

A. Acoustic Feature Set

For each utterance, one feature vector is calculated. For emotional state estimation, a total of 472 acoustic features were considered. Relevant acoustic cues from emotionally rich speech expressions were taken from phonation and articulation speech phases. Features were extracted mostly from speech prosody information, i.e. pitch, energy and duration, and from spectral domain parameters like formants and mel-frequency cepstral coefficients (MFCC). Following groups were used for feature categorization: *raw speech*, *speech rate*, *zero-crossing rate*, *short term energy*, *fundamental frequency*, *spectrum*, *formant* and *voice harmonic*. Similar categorizations were also used in [11] – [13]

A large amount of features was calculated from statistical measures of speech parameters, like mean value, minimum, maximum, median, 25th and 75th percentiles, standard deviation, skewness, kurtosis, etc. Speech rate features were calculated separately from voiced, unvoiced and silence intervals in an utterance, while several features were calculated from interval relations. Beside whole utterance features, some short term energy (STE) and fundamental frequency (F_0) features were calculated from specific rising and falling parts of the utterance, as well as minimum and maximum plateau. Furthermore, F_0 was used for measuring of typical period-to-period fluctuations (jitter) and period-to-period variability of the amplitude value (shimmer). Spectral domain features like spectral flux, energy of spectral banks, center of gravity, spectral roll-off-point, etc. were calculated

from: short-term spectrogram, long-term spectrum of the whole utterance and individually averaged short-term spectra of voiced and unvoiced parts of the utterances. Also, statistical features from 13 MFC coefficients (12 plus 0^{th} order coefficient) were calculated. Formant features were taken from central frequencies and bandwidths of the first four formants. Finally, analysis of the harmonic excitation of the human voice was performed, i.e. features were calculated from voice quality and harmonic-to-noise ratio (HNR) measures.

B. Estimation of Emotional States

Estimation of emotion dimensions was performed using LIBSVM implementation of SVR. Following parameters were applied: 10-fold cross-validation (CV) process (k = 10) was selected for each task (*valence*, *arousal*); radial basis function (RBF) was used as a kernel function with γ set to $1/F$ (number of features, F = 472); the cost parameter C was set to 1; and threshold ε was set to 0.001. Furthermore, the sequential floating forward selection (SFFS) algorithm was used to select 50 most relevant features for each task, with tolerance set to 2 features. The mean squared errors (MSE) were set as a criterion function. The reference values, i.e. utterance labels, for 10-fold CV were defined as centroids μ_s of the Gaussians, presented in Fig. 3. Minimal MSE for estimation of *valence* was 2.2497, achieved with 44 selected features, while minimal MSE for *arousal* was 1.8147, achieved with 51 features. It should be noted that emotion labels of each utterance are continuous variables from intervals of [1:9] for *valence* and *arousal*.

Classification of discrete emotional states: *happiness*, *sadness*, *anger*, *fear* and *neutral state* was also performed, using LIBSVM implementation of the SVM, with the same parameters as for SVR. Classification accuracy was used as a criterion function and the referent knowledge for the 10-fold CV was defined as described in the last column in Table 1. Maximal obtained accuracy for 5 discrete emotions classification was 69.41%, with 40 features selected. Confusion matrix is presented in Table 2.

TABLE II. CONFUSION MATRIX FOR 5 DISCRETE EMOTIONS

		Predicted Emotion					
		H	***S***	***F***	***A***	***N***	**Total**
Actual emotion	***H***	**103**	8	11	44	13	179
	S	17	**100**	16	34	16	183
	F	7	12	**90**	24	9	142
	A	18	11	12	**241**	21	303
	N	4	7	2	22	**165**	200
Total		149	138	131	365	224	1007

Note: *H = happiness*; *S = sadness*; *F = fear*; *A = anger*; *N = neutral state.*

V. METHODS FOR EMOTIONAL STATE ESTIMATION BASED ON A PRIORI KNOWLEDGE

Two methods are proposed in this paper that can be applied as an enhancement of conventional estimation of emotion dimensions *valence* and *arousal*, described in the previous section. The first one is a basic method in which adaptation is preformed using single integral *a priori* model, and the second method is based on integration of multiple *a priori* models that represent different discrete emotions.

A. Emotional State Estimation Using Single A Priori Model

This approach used *a priori* model of relations between emotion dimensions *valence* and *arousal*. Similar approach is performed in MAP estimation procedure where ML estimation is regularized using *a priori* distribution over the estimation parameters. However, it is possible to combine propose *a priori* model with conventional emotional state estimation approaches, even if they do not exactly correspond to ML estimation.

As described in section III, *a priori* model $P(\theta)$ was built by substituting $\theta = (\theta val, \theta ar)$ for observation vector o in (2), where θ represents 2-dimensional emotion parameters *valence* and *arousal*, i.e. $\theta \in [1:9]\times[1:9]$. Furthermore, as 10-fold CV (k = 10) was performed for emotional state estimation, *a priori* model was generated from the training set of the corpus $Tr \subseteq COR$ in each validating iteration k, where COR is a total corpus set, which consist of S samples. According to 10-fold CV paradigm, training set Tr and validation set $Va \subseteq COR$ are defined in the 90/10 ratio, where $Tr \bigcup Va = COR$. Such *a priori* model is applied on the conventional estimation process in order to shift estimation results in valence-arousal space toward more probable values, according to the level of the estimation uncertainty.

The output $\hat{\theta}_s^{(1)}$ from the estimator for sample s is used to create 2D Gaussian model $P(\theta|\hat{\theta}_s^{(1)})$ that represents uncertainty of this estimator in the point $\hat{\theta}_s^{(1)}$ and is constructed from the Tr, according to 10-fold CV. The centroid of the uncertainty model is defined as $\hat{\theta}_s^{(1)}$, while the diagonal covariance matrix is calculated using the kernel density estimation (KDE), where variances depend on the accuracy of estimations of Tr samples, and are calculated separately for each emotion dimension. For example, variance of *arousal* is calculated by:

$$\sigma_{ar}^2(x) = \frac{\sum_{i\in Tr}\left(\hat{\theta}ar_i^{(1)} - \theta ar_i\right)^2 g\left(x \mid \theta ar_i, 0.1\right)}{\sum_{i\in Tr} g\left(x \mid \theta ar_i, 0.1\right)}, \quad (6)$$

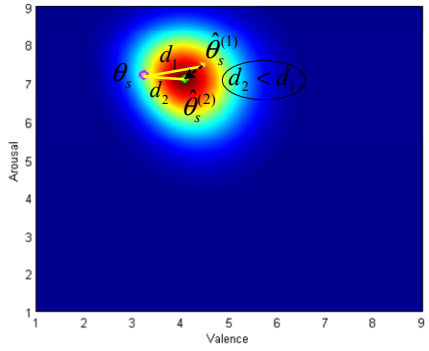
where *arousal* estimate $\hat{\theta}ar_s^{(1)}$ is substituted for $x \in [1:9]$. Variance of 1D Gaussian kernel $g(x|\theta ar_i,0.1)$ is experimentally set to a constant value 0.1, θar_i are *arousal* references and $\hat{\theta}ar_i^{(1)}$ are arousal estimates for $i \in Tr$.

Results were calculated from the product of the uncertainty model $P(\theta|\hat{\theta}^{(1)})$ and *a priori* model $P(\theta)$. Therefore, final estimate of the sample s was calculated as the location, i.e. index, of a maximum value of the product in 2D valence-arousal space:

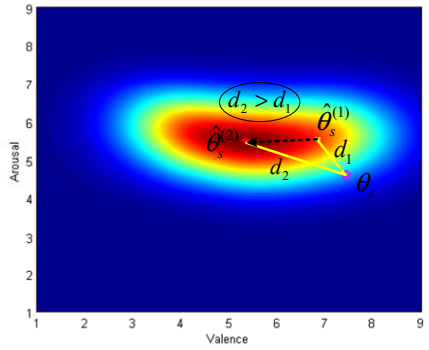
$$\hat{\theta}_s^{(2)} = \arg\max_{\theta}\left\{P\left(\theta \mid \hat{\theta}_s^{(1)}\right)\cdot P(\theta)\right\}. \quad (7)$$

Two illustrations of estimation adaptations based on *a priori* model are presented in Fig. 4. It must be noted that in hypothetical case, where estimation is 100% accurate ($\hat{\theta}_s^{(1)} = \theta_s$, for s = 1,..,S), where θ_s are referent values, adapted value will be equal to the original value ($\hat{\theta}_s^{(2)} = \hat{\theta}_s^{(1)}$ for each s). It is because variance of the uncertainty model will be zero. Structure of the adaptation process is given in Fig. 5.

The adaptation method with single *a priori* model did not improve accuracy of emotional states estimation, either for *valence* or *arousal*. Increased MSE of 2.3688 was achieved for *valence*, while for *arousal* it increased to 1.8230. Only 48.76% of the samples were shifted in the right direction, i.e. the 2D Euclidean distance d from the reference point θ_s to an estimate was decreased.



(a) *example of positive shifting*



(b) *example of negative shifting*

Fig. 4. Adaptation of valence-arousal estimates using *a priori* model.

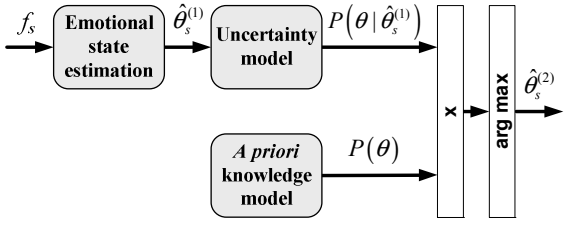


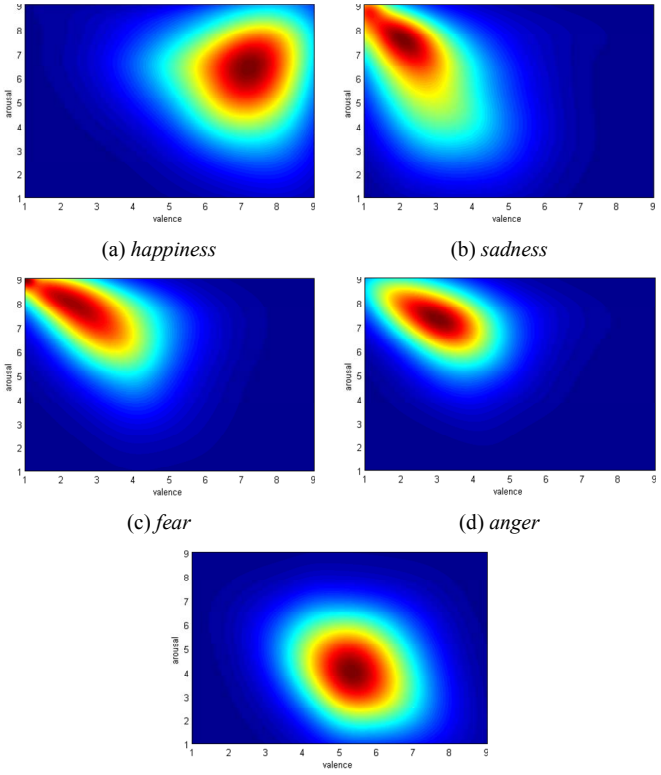
Fig. 5. Structure of the emotional state estimation process using *a priori* knowledge model, where f_s is acoustic speech feature vector, $\hat{\theta}_s^{(1)}$ is valence-arousal estimate calculated from the conventional estimator, and $\hat{\theta}_s^{(2)}$ is final estimate of the adapted estimation process.

B. Emotional State Estimation Using Multiple A Priori Models

The second method for adaptation of conventional emotional state estimation procedure is based on integration of multiple *a priori* models, one for each discrete emotion. This approach is less sensitive to predominance of negative, highly arousing emotions, as well as different proportions of specific discrete emotions in the corpus. The theoretical foundations of relationship between emotion dimensions and discrete emotions were introduced in [14], and since then implemented in several research paradigms, like in [15] and [16].

Adapted valence-arousal estimates were calculated in collaboration with system for discrete emotional state classification. Therefore, multiple *a priori* models $P(\theta|e)$, with e = 1,..,5, were developed, one for each discrete emotion. Estimation results for emotion dimensions of each sample were processed in the identical way as in (7), now using $P(\theta|e)$, i.e. *a priori* model of the recognized emotion e for sample s (Algorithm 1.).

Estimation results were improved. For *valence*, MSE was decreased to 1.9601 and for *arousal* it was decreased to 1.5832. A total of 71.5% of the samples where shifted in the right direction. In hypothetical case, if classification of discrete emotions is 100% accurate, then the MSE decreases to 1.5184 for *valence*, and to 1.5803 for *arousal*. There are 82.13% positive shifts in this case. *A priori* models of discrete emotions are presented in Fig. 6., while summary of the estimation results is presented in Table 3.



(a) *happiness* (b) *sadness*

(c) *fear* (d) *anger*

(e) *neutral state*

Fig. 6. *A priori* models for 5 discrete emotions.

Algorithm 1. Estimation of emotional states using multiple *a priori* models

i) Calculate the discrete emotion e_s = 1,…,5 from the acoustic feature vector f_s of the sample s;

ii) Estimate *valence* and *arousal* ($\hat{\theta}_s^{(1)}$) from f_s;

iii) Create the uncertainty model $P(\theta|\hat{\theta}_s^{(1)})$ using $\hat{\theta}_s^{(1)}$;

iv) Adapt the estimate $\hat{\theta}_s^{(1)}$ using *a priori* model $P(\theta|e_s)$ of discrete emotion e_s, calculated in (*i*): $\hat{\theta}_s^{(2)} = \arg\max_{\theta}\left\{P\left(\theta|\hat{\theta}_s^{(1)}\right)\cdot P\left(\theta|e_s\right)\right\}$.

TABLE III. ESTIMATION STATISTICS

	CE	**AE1**	**AE2**	**REF AE2**
Valence (MSE)	2.25	2.37	1.96	1.52
Arousal (MSE)	1.81	1.82	1.58	1.58
Positive shifts (%)	-	48.76	71.5	82.13

Note: *CE = conventional estimation*; *AE1 = adaptation of the estimation using single a priori model*; *AE2 = adaptation using multiple a priori models*; *REF AE2 = reference for AE2.*

VI. CONCLUSION

Two methods of *a priori* knowledge modeling were presented in the paper: single model over the valence-arousal space, which did not increase estimation accuracy; and integration of multiple models that represent different discrete emotions in the valence-arousal space, which resulted in higher accuracy of estimation. Superiority of the second method is due to its robustness against various sources of bias that exist in the current version of Croatian emotional speech corpus, like predominance of negative, highly arousing emotions and unbalanced representation of different discrete emotions in valence-arousal space. On the contrary, single model reflects more the process of selecting emotional speech utterances for inclusion in the corpus, than the underlying emotion-theoretic relationships between valence and arousal, which the model should try to capture.

One direction of future work may be related to refining the proposed uncertainty model for the conventional emotional state estimator used in this paper. Current uncertainty model represents separately marginal uncertainties for estimated valence and arousal, which could be replaced by the joint valence-arousal uncertainty model.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under project: "Adaptive control of scenarios in virtual reality therapy of posttraumatic stress disorder (PTSD)" (#036-0000000-2029).

Authors would like to thank Departments of Psychology and Phonetics, from the Faculty of Philosophy at University of Zagreb and Department of Electronic Systems and Information Processing, from the Faculty of Electrical Engineering and Computing at University of Zagreb for support in organization and participation in building process of the Croatian emotional speech corpus. We also thank the anonymous reviewer for valuable comments.

REFERENCES

[1] P. Baranyi and A. Csapo, "Definition and Synergies of Cognitive Infocommunications," Acta Polytechnica Hungarica, vol. 9, pp. 67–83, 2012.

[2] K. Ćosić, S. Popović, D. Kukolja, M. Horvat and B. Dropuljić, "Physiology-driven adaptive virtual reality stimulation for prevention and treatment of stress related disorders," CyberPsychology, Behavior, and Social Networking, vol. 13(1), 73-78, 2010.

[3] K.R. Scherer, "Appraisal considered as a process of multi-level sequential checking," in Appraisal processes in emotion: theory, methods, research, New York and Oxford: Oxford university press, pp. 92-120, 2001.

[4] F. Eyben et al., "On-line Emotion Recognition in a 3-D Activation-Valence-Time Continuum using Acoustic and Linguistic Cues," Journal on Multimodal User Interfaces, 2010.

[5] M. Grimm, K. Kroschel and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," ICAASP '07, 2007.

[6] H. Hu et al., "GMM supervector based SVM with spectral features for speech emotion recognition," ICAASP '07, 2007.

[7] N. Sebe et al., "Emotion recognition using a cauchy naive bayes classifier, " Pattern Recognition, vol. 1., 2002.

[8] B. Dropuljić et al., "Emotional Speech Corpus of Croatian Language," International Symposium on Image and Signal Processing and Analysis, ISPA '11, pp. 95-100, 2011.

[9] P.J. Lang, M.M. Bradley and B.N. Cuthbert, "International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual," Technical Report A-6. University of Florida, Gainesville, FL, 2005.

[10] B. Russell, "A Priori Justification and Knowledge," The Stanford Encyclopedia of Philosophy, 2013.

[11] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," J. Personality Social Psych., 70(3), pp. 614-636, 1996.

[12] B. Schuller et al., "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," IEEE Proc. ICASSP '04, pp. 577-580, 2004.

[13] M. Lugger and B. Yang, "Psychological motivated multi-stage emotion classification exploiting voice quality features," Speech Recognition, In-Tech, 2008.

[14] J.A. Russell, "A circumplex model of affect," Journal of Personality and Social Psychology, Vol. 39, pp. 1161-1178, 1980.

[15] J.A. Mikels et al., "Emotional category data on images from the International Affective Picture System," Behavior Research Methods, vol. 37 (4), pp. 626-630, 2005.

[16] K. Sun, J. Yu, Y. Huang and X. Hu, "An improved valence-arousal emotion space for video affective content representation and recognition," in ICME '09, pp. 566-569, 2009.