Emotional Speech Corpus of Croatian Language

Branimir Dropuljić*, Miłosz Thomasz Chmura, Antonio Kolak, Davor Petrinović
Faculty of electrical engineering and computing
Zagreb, Croatia
*branimir.dropuljic@fer.hr

Abstract—**As a first step in developing an emotion recognition system from human voice, it is necessary to collect relevant set of emotionally rich utterances that will be used for system training. Thus, a first emotional speech corpus of Croatian language (KEG) was built and annotated. The collection and annotation process together with some interesting statistical properties of the designed corpus are described in this paper. Utterances were collected from both male and female speakers, from child age to adults, verbally expressing their emotions. Materials were taken from Internet and other public media sources, with the total duration of approximately 40 minutes. Emotion classification used for annotation has been based on 5 discrete emotional states: happiness, sadness, fear, anger and neutral state. For each of the non-neutral emotional states, the perceived intensity was also annotated in 10 steps. Preliminary KEG evaluation was performed by building and testing an emotion recognition system based on this specific corpus. Initial results are presented in this paper.**

I. INTRODUCTION

Emotion recognition issues gain more and more popularity in various scientific fields lately. Growing interest of human-computer interaction community (HCI) in the study of emotions surely gives significant contributions to this field. Modeling of advanced interfaces for HCI should, at least in part, be based on accurate identification of emotional states. Some of successfully built emotion recognition models were presented in [1] – [7]. Currently, the most popular application domains are safe driving systems, e-learning, telemedicine, improvement of the remote home care, etc.

Emotional states can be manifested via various modalities, like psychophysiology, speech, facial expressions, gestures, etc. This paper describes early development of emotion recognition system that is based on acoustic and linguistic features of speech signals. Considering both the advantages and disadvantages of all modalities, we have sorted out speech as a great potential, being the most noninvasive recording sensors that gives the clue of the subject's emotional state. The naturalness of expressing emotions through speech without hesitation, greatly simplifies the acquisition of emotion expression materials.

Measurable relation between emotions and speech was scientifically discovered in 1930-ies. In 1936, Cowan made the first analysis of acoustic features of a human voice recorded during public speeches [8]. Several years later, Fairbanks and Pronovost went a step further by analyzing speeches recorded in the expression of a wider spectrum of emotions [9]. After revealing its potential, this interdisciplinary topic began expanding circles of interest. Psychologists and linguists were joined by neurologists and more recently, by computer experts, who contribute greatly to this field by developing computer systems for automatic emotion recognition, as well as for analysis and selection of appropriate voice features using statistical methods. Some of the most significant scientific breakthroughs in this field were following works: Frick [10], Scherer [11] – [13], Murray [14], Russell [15], Lee [3], Schuller [16], Wu [17] and Rong [18].

Emotional information from human speech can be extracted from two main features: acoustic and linguistic. Emotions can be expressed non-verbally, through prosodic structure of an utterance, i.e. acoustic information, but also verbally, by directly expressing thoughts and feelings using words, i.e. by linguistic information. Most of the authors put emphasis on voice acoustics, because it's more vulnerable to the emotional impact (emotions are harder to conceal or control) [12], [13], [18]. The most relevant acoustic features can be extracted from the vocal cords oscillation period, energy of the voice, speech rate and the voice spectrum distribution [12]. Some authors focused their research on extracting emotions from linguistic information [17]. Valuable information can be extracted from simple n-gram language models describing statistical probability of emotional keywords and phrases. Emotion recognition can be done on three levels: lexical, syntactic and semantic level. As acoustic and linguistic features are two important aspects of an affective speech that are not necessarily correlated, modality fusion can provide integral speech information for emotion recognition that exceeds performance of each individual source [19], [3], [20], [16].

Starting point, when building emotion recognition system from human voice is collection of emotionally rich speakers' utterances, i.e. affective speech corpus. Some researchers prefer real-life emotions for this purpose, while others prefer acted emotions. According to Sherer [12] it's hard to separate this two modalities: "Although natural expressions are partly staged, acted expressions are also partly natural.". Quality of the corpus depends on different factors, like variety of speakers (age differences, gender, speaking styles and different dialects), the quantity of collected materials, but also the quality and good balance of targeted emotional expressions. Other factors that also must be considered when building a corpus are quality of recording and data storage systems, intelligibility of utterances and avoidance of double-talk, noise, music or other sounds that are not related to the emotional speech.

Although re-using existing corpora is generally desirable, the only currently available corpus of Croatian language is

VEPRAD [21], [22]. It consists of news and weather reports from radio speakers, and is mainly constructed from neutral speech utterances, that are insufficient for our application. Reuse of available English emotional corpuses ([12], [23], [24], etc.) or Serbian emotional corpus ([25]) was not an option since target applications are related to Croatian language, primarily in Stress Inoculation Training (SIT), Virtual Reality Exposure Therapy (VRET), and other similar stress reduction and therapy methods at Croatian psychotherapy institutions [26] – [28].

Hence, a first emotional speech corpus of Croatian language (KEG - *Korpus emocionalnog govora*) was built in two phases: collection phase and annotation phase, which are described in section 2. The concept of emotion recognition system from human speech based on KEG, with preliminary results, is described in section 3 and conclusions in section 4.

II. Collection and Annotation of Emotional Speech Corpus

Current version of KEG consists of approximately 40 minutes of affective speech segments. Utterances were annotated with both discrete and dimensional emotional representation [29]. Besides emotions, utterances were annotated with speaker ID, gender and age group, together with emotional expression modality (acted or a real-life emotion). Additionally, word level transcription was included for each utterance in the corpus. Besides regular Croatian words a few special non-speech acoustic events were added to the vocabulary, that are common for emotional speech.

A. Collection Phase

There are three main approaches to collect an emotional affected speech corpus. One way is to organize audition sessions with professional actors who would act a specific scenario eliciting a specific emotion. Second one is to organize some kind of controlled environment situation, e.g. computer game playing session, where the participants would not know that they are being recorded, as suggested in [30]. By controlling such environment, the participants are stimulated to elicit different emotions. The third option is to collect emotional speech utterances from various prerecorded sources like Internet or movies. Each of the described approaches has its cons and pros. It is usually hard to acquire real emotions with controllable sound quality from prerecorded materials. On the other hand, studio recorded emotions are usually not naturally elicited.

The most popular method when collecting corpus is the first approach. Emotion expressions are recorded in the studio by professional actors based on prepared scenarios for eliciting a particular emotion, like in [12], [20] and [18]. In our case, due to limited project funding, the third method was chosen for building KEG. For the initial version, Internet and movie clips were collected and filtered with intention to extract short audio excerpts with pure emotion expression. The first part called “real-life emotions” was collected from Internet, mostly from Croatian reality shows and from different documentaries. The second part called “acted emotions” was collected from Croatian movies, TV Shows and Books-Aloud programs. Such initial classification enables further research on differences between acted and real-life emotions.

There are three main precautions which were taken into account while collecting utterances: 1) there had to be only one person expressing only one emotion, according to our initial judgment; 2) other sounds like environment sounds, other voices or recording noise had to be limited to the minimum; and 3) quality of the voice must be high enough such that a human listener is able to clearly recognize conveyed information from the speech together with potential emotional expression. Such stringent requirements typically yield only 2 minutes of usable emotion material from one hour of a soundtrack. The main problem while extracting prerecorded acted emotions is to find parts with no background sounds. On the other hand, problem with real-life emotion utterances is much lower audio quality, and abundance of doubletalk.

The main result of the project was creation of a database that enables building of an automatic emotion recognition system for Croatian language. The system is based on acoustical properties of emotional speech in Croatian language, but also on Croatian linguistic properties for expression of different emotions. The issue of proper coverage is very important when building a corpus. All dialects of the language must be covered within the corpus. Additionally, our KEG corpus includes a small percentage of utterances from speakers with Bosnian and Herzegovinian, as well as Serbian origin. Such inhomogeneity is justified with the fact that Croatian population is significantly mixed with these national minorities.

Collected utterances were normalized and stored in "wav" format with 11025Hz sampling frequency, 16 bits per sample, monaural. At the collection phase, emotionally interesting utterances were extracted and initially classified in one of the five discrete emotion categories: happiness, sadness, anger, fear and neutral state.

A total of 714 utterances were collected with duration of 56:22 minutes. After retrospective filtering process that was performed by a neutral listener who was not involved in the collection phase, 40 utterances were removed and the resulting KEG collection statistics is presented in Table 1.

There are 674 utterances in total with duration of 46 minutes and 55 seconds. It is interesting that utterances expressing happiness and anger have the shortest average duration of 3s, probably due to the highest speaking rate. On the other hand, neutral utterances have the largest average duration of 10.39s. The number of different male and female speakers in KEG is almost identical (104/100). Utterances and speakers were also classified into 3 age categories: children, adolescent and adults. Most utterances belong to the adult class, since our targeted applications involve mainly adults. Based on the source of the extracted audio clip, each utterance was classified as either a real-life emotion or an acted one. As can be observed from the table, both classes are almost equally represented in KEG (331/343).

TABLE I. KEG STATISTIC AT THE COLLECTION PHASE

			Emotions					Total
			Happiness	*Sadness*	*Anger*	*Fear*	*Neutral*	
Utterance statistic	*Number*		145	105	287	72	65	**674**
	Average duration (s)		3.04	5.51	3	3.6	10.39	**4.18**
	Duration (mm:ss)		07:21	09:39	14:21	04:19	11:15	**46:55**
	Gender	*Male*	53	28	157	19	38	**295**
		Female	92	77	130	53	27	**379**
	Age category	*Child*	9	2	3	22	3	**39**
		Adolescent	1	0	2	0	0	**3**
		Adoult	135	103	282	50	62	**632**
	Emotional expression	*Real-life*	77	50	145	13	46	**331**
		Acting	68	55	142	59	19	**343**
Speaker statistic	*Number*		49	30	61	28	36	**204**
	Gender	*Male*	27	11	31	13	22	**104**
		Female	22	19	30	15	14	**100**
	Age category	*Child*	3	1	1	4	2	**11**
		Adolescent	1	0	1	0	0	**2**
		Adoult	45	29	59	24	34	**191**

Besides just described annotated metadata, each utterance was also transcribed manually using word-level transcription in Croatian. This transcription was enriched by inclusion of several special acoustic events like crying, laughing, breathing, hesitation, yelling, coughing, a few types of exclamations, etc. Such transcription is very important for training of an automatic speech recognition system (ASR). Its output is then used to train the classifier of the emotional state from the enriched linguistic information.

B. Annotation Phase

The initial emotional classification was potentially biased due to the fact that most of the audio excerpts were extracted from video clips. Since the automatic system for emotion recognition will be designed to use only the audio part, we have performed additional round of annotations based on subjective evaluation of extracted audio utterances. This annotation process was devised in collaboration with colleagues from the Department of Psychology, at University of Zagreb. Five undergraduate students aged 22 to 24, took part in this experiment as annotators. Three of them have had a formal education in psychology. Their task was to listen to the entire corpus and give their assessment of emotional states for each recording. Average duration of the annotation was approximately 15 hours for each student.

All students have undergone a rigorous training in the devised evaluation procedure. In order to motivate the students further, a double bonus was promised to the student that produced the best annotation. For the purposes of annotation, application for recordings grading was developed in Visual Basic. Screenshot can be seen in Fig. 1. Each annotator was given his version of randomly permutated recordings to prevent result comparison with others and to minimize convergence/saturation of results at the end of the corpus. Annotators were also instructed to annotate at most 100 recordings in a row, and to take pauses of at least 3 hours between annotation sessions.

Specifically, both discrete and dimensional emotions were annotated, i.e. two different grades for the same emotional state were attributed. Non-neutral discrete emotions (happiness, sadness, fear and anger) were assessed on a scale of 0 to 10, where zero denotes the absence of emotion. If all of them were rated by low intensity grades (0 to 3), then it was considered to be neutral emotional state. It was not necessary to judge one recording by only one discrete emotion. Annotators were rather instructed to give the grade distribution along all four emotions. That means that the same recording could be present in all emotional classes with varying intensity.

The same application was used to annotate the so-called dimensional emotions of recordings, described by the scale of valence and arousal [29]. We have defined 9 levels for valence and arousal, as suggested in [31] and [32]. Valence levels ranged from 1 that was the maximum unpleasant state up to 9 that was the maximum pleasant state. Similar grading scale was used for the arousal levels that ranged from level 1 that represents sleepiness up to level 9, representing the maximum arousal. Level 5 is considered neutral on both scales.

Beside described emotion annotation, annotators were instructed to add special remarks in the corresponding textual field of the application, for recordings for which emotions of the speakers were not fully describable with regular grading method. Such cases were rare because ambiguous recordings were avoided during the collection stage. Four characteristic remarks were suggested to the annotators as follows: 1) Some other discrete emotions beside the basic four appear in the recording. They had to list these emotions along with their intensity; 2) It was not possible to determine the discrete or dimensional emotional state in the recording due to various reasons, e.g. loud noise, conflicting sounds, bad voice quality, etc.; 3) Inhomogeneity of the emotional state, if either discrete or dimensional emotional states vary throughout the utterance by more then 3 grades; 4) Two or more speakers, of nearly equal dominance are noticed in the same utterance.

After collecting all annotations, the results were analyzed and statistically processed. The first step of this procedure was removal of utterances for which majority of annotators added remarks 2), 3) or 4) thus marking them as unusable. This was the case for only four questionable utterances. The second step was to verify agreement between annotators. For each utterance, the dominant (the most prominent) annotated discrete emotion was compared between annotators' responses. If majority agreed on the same emotion, the utterance was retained; otherwise it was removed from the corpus.

Although all additional emotions reported in the first remark were also taken into consideration, none was selected as representative, since minority of annotators reported such anomalies for each annotated utterance. The final results presented in Table 2 show that the number of utterances was reduced from 674 to 496. When comparing new utterance annotations agreed upon by the majority of annotators with those from the collecting phase, the same emotion label was given to 462 of 496 retained utterances. Gender, age and emotional expression ratios of utterances changed minimally. Although the corpus was reduced for almost 200 removed utterances, the total time duration decreased by only 5 minutes. Obviously, it was for short utterances, that annotators were not able to agree upon the same emotion. Thus, the average duration of the retained recordings increased by almost a second.

Fig. 2 shows statistical results that were first averaged across all annotators for each utterance per all four non-neutral emotions (happiness, sadness, fear and anger). These averages were then analyzed across all utterances of the corpus. Utterances were classified according to the emotion annotation of the utterance and plotted in 5 graphs using "box and whiskers" plot method. The whiskers extend to the most extreme data points (not considering the outliers). Outliers are plotted individually, using red crosses.

It can be seen that utterances annotated as happiness involve insignificant presence of other emotions, while on the other side, sadness, anger and fear interact with each other more frequently. Such behavior is justified and logical. According to the Russell's Circumplex model [29], emotions like fear and anger are much closer to each other than each of them to the happiness.

III. Initial Evaluation of Emotional Speech Corpus

With approximately forty minutes of annotated affective speech material, KEG represents a good starting point for training of a speech based emotion recognition system. As it also contains a word-level transcription of each utterance, both acoustic and linguistic features can be used for data-driven design of an automatic emotion recognition system.

Emotion classification using acoustic features of the speech is performed by extracting relevant statistical features from short-time estimates of speech energy and power (in decibels), pitch contour (vocal cords oscillation period), zero-cross rate (ZCR) and Mel frequency cepstral coefficients (MFCC) [33]. Given the feature set, emotions are modeled by two selected statistical methods: Gaussian mixture models (GMM) [34] and Hidden Markov models (HMM) [35]. GMM model captures only the static probability distribution of feature vectors for each emotion. Only one feature vector is computed for the whole utterance and the most probable model is selected given the observation. With HMM model, the time dynamics is also included in the emotional model (i.e. changes across the utterance). Recognition is performed by evaluating the probability of a time series of observed feature vectors given the set of designed HMM models, one for each emotion. Each vector is calculated from a subsequence of utterance extracted with a time window of an appropriate duration and time shift. The recognition accuracy of 40% was achieved for our initial HMM models with 5 discrete emotions, what leaves ample room for future improvements [36].

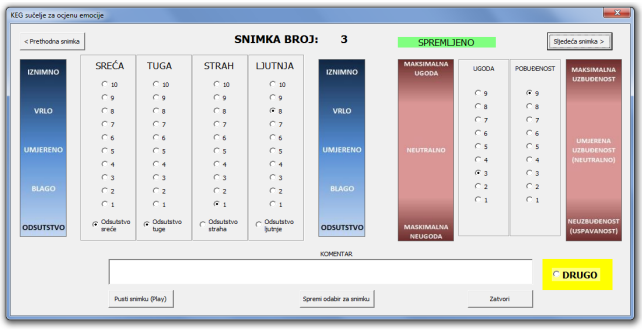

Figure 1. Application for emotion annotation

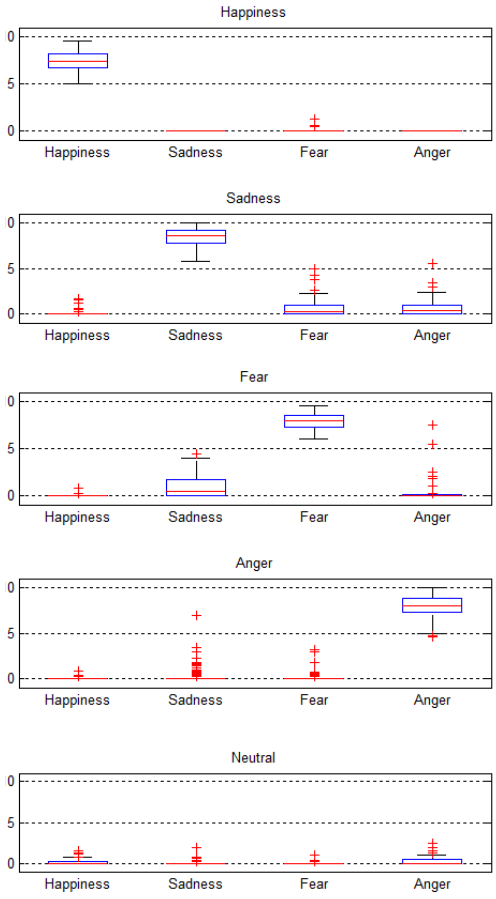

Figure 2. Results per each emotion for all utterances

TABLE II. KEG STATISTIC AT THE ANNOTATION PHASE

			Emotions					Total
			Happiness	*Sadness*	*Anger*	*Fear*	*Neutral*	
Utterance statistic	*Number*		77	83	231	33	72	**496**
	Average duration (s)		4.25	6.47	3.49	4.42	9.15	**4.99**
	Duration (mm:ss)		05:27	08:57	13:27	02:26	10:59	**41:16**
	Gender	*Male*	25	16	133	7	43	**224**
		Female	52	67	98	26	29	**272**
	Age category	*Child*	10	1	2	14	1	**28**
		Adolescent	1	0	2	0	0	**3**
		Adoult	66	82	227	19	71	**465**
	Emotional expression	*Real-life*	34	44	105	3	61	**247**
		Acting	43	39	126	30	11	**249**
Speaker statistic	*Number*		34	30	60	16	50	**146**
	Gender	*Male*	15	7	31	6	30	**69**
		Female	19	23	29	10	20	**77**
	Age category	*Child*	3	1	1	3	1	**7**
		Adolescent	1	0	1	0	0	**2**
		Adoult	30	29	58	13	49	**137**

As for the linguistic features, current emotion recognition system is based on a unigram language model for each discrete emotion, where words are presented as a model input. Recognition accuracy of such linguistic based emotion estimator is 47% [37], what is slightly better than the acoustic counterpart. Currently, the input to this linguistic based recognition system was the manual word-level utterance transcription of the corpus, rather then the automatic speech recognition system output (ASR). The adapted version of ASR for Croatian language presented in [38] and [39] that was designed using the collected KEG corpus did not result with sufficient recognition accuracy.

IV. CONCLUSION AND FUTURE WORK

This paper describes a building and evaluation process of the first emotional speech corpus of Croatian language (KEG). KEG currently consists of 496 emotionally annotated utterances making a total duration of 41:16 minutes. Utterances where collected and filtered mostly from Internet and movie clips with intention to extract audio excerpts with speakers' pure emotion expression. Annotation process was devised in collaboration with colleagues from the Department of Psychology. Besides emotions, KEG content was also classified based on gender, age category and emotion expression (real-life or acted emotions).

Initial results of automatic emotion classification show satisfactory accuracy and thus provide encouragement for further improvements of training procedures for statistical models used for emotion recognition. Future work also includes the analysis of annotated dimensional emotions and development of statistical models for estimation of valence and arousal levels from the speech. Expansion of the corpus is also planed in the future in order to achieve better balance between emotions and more reliable models especially for fear that is least represented in the current corpus.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia under project: "Adaptive control of virtual reality scenarios in therapy of posttraumatic stress disorder (PTSD)" (#036-0000000-2029). Authors would like to thank colleagues from the Department of Psychology, at University of Zagreb for useful discussion and comments during preparation of the annotation experiment.

REFERENCES

[1] R. W. Piccard, "Affective computing," Technical Report No. 321, Media Laboratory Perceptual Computing Section, M.I.T, 1995.

[2] R. Fernandez and R.W. Picard, "Modeling drivers' speech under stress," Speech Comm., vol. 40, pp. 145-159, 2003.

[3] C.M. Lee and S.S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans. Speech and Audio Processing, vol. 13, no. 2, pp. 293-303, Mar. 2005.

[4] C. Lisetti and F. Nasoz, "Affective intelligent car interfaces with emotion recognition," In Proceedings of 11th International Conference on Human Computer Interaction, Las Vegas, NV, USA, July 2005.

[5] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," Proc. Ninth Int'l Conf. Spoken Language Processing, 2006.

[6] F. Nagel, R. Kopiez, O. Grewe and E. Altenmüller, "EMuJoy: Software for continuous measurement of perceived emotions in music," Behavior Research Methods, 39 (2) , pp. 283-290, 2007.

[7] D. Datcu and L.J.M. Rothkrantz, "Multimodal recognition of emotions in car environments," D|C|I&I Prague, 2009.

[8] M. Cowan, "Pitch and intensity characteristics of stage speech", Arch. Speech, 1936.

[9] G. Fairbanks and W. Pronovost, "An experimental study of the pitch characteristics of the voice during the expression of emotion," Speech monograph, vol 6, pp. 87-104, 1939.

[10] R. Frick, "Communicating emotion: The role of prosodic features," Psychol. Bull., vol. 97, pp. 412-429, 1985.

[11] K. Scherer, "Vocal affect expression: A review and a model for future research," Psychological Bulletin, vol. 99, pp. 143-165, 1986.

[12] R. Banse and K. Scherer, “Acoustic profiles in vocal emotion expression,” J. Personality Social Psych., vol. 70, no 3, pp. 614-636, 1996.

[13] P.N. Juslin and K.R. Scherer, “Vocal expression of affect,” The New Handbook of Methods in Nonverbal Behavior Research, Oxford Univ. Press, 2005.

[14] I.R. Murray and J.L.Arnott, “Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion,” J. Acoust. Soc. Amer., vol. 93, pp. 1097-1108, 1993.

[15] J.A. Russell, J.A. Bachorowski, and J.M. Fernandez-Dols, “Facial and vocal expressions of emotion,” Ann. Rev. of Psychology, vol. 54, pp. 329-349, 2003.

[16] B. Schuller, R.J. Villar, G. Rigoll, and M. Lang, “Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition,” Proc. IEEE Int’l Conf. Acoustics, Speech, and Signal Processing, 2005.

[17] C.-H. Wu, Z.-J. Chuang and Y.-C. Lin, “Emotion recognition from text using semantic labels and separable mixture models,” ACM Transactions on Asian Language Information Processing, vol. 5, no. 2, pp. 165-182, Jun. 2006.

[18] J .Rong, G. Li, Y.P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” Information processing and management, vol. 45, pp. 315-328, 2009.

[19] T.S. Polzin, “Verbal and non-verbal cues in the communication of emotions,” in Proc. ICASSP ‘00, pp. 2429-2432, 2000.

[20] B. Schuller, G. Rigoll and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture”, in Proc. ICASSP ‘04, pp. 577-580, 2004.

[21] S. Martinčić-Ipšić, “Croatian speech recognition and synthesis based on context-dependent Hidden Markov model,” PhD Thesis, in Croatian, University of Zagreb, Croatia, Nov. 2007.

[22] S. Martinčić-Ipšić, S. Ribarić and I. Ipšić, “Acoustic modelling for Croatian speech recognition and synthesis,” Informatica, vol. 19, pp. 227-254, 2008.

[23] S. Burger, V. MacLaren, and H. Yu, “The ISL meeting corpus: The impact of meeting type on speech style,” Proc. Eighth Int’l Conf. Spoken Language Processing (ICSLP), 2002.

[24] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, “Emotional speech: Towards a new generation of database,” Speech Comm., vol. 40, nos. 1/2, pp. 33-60, 2003.

[25] S.T. Jovičić, Z. Kašić, M. Đorđević and M. Rajković, “Serbian emotional speech database: design, processing and evaluation,” in 9th Speech and Computer Conference, Sep. 2004.

[26] D. Kukolja, S. Popović, B. Dropuljić, M. Horvat and K. Ćosić, “Real-time emotional state estimator for adaptive virtual reality stimulation,” Lecture Notes in Computer Science, vol. 5638, pp. 175-184, Jul. 2009.

[27] 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, 13, pp. 1; 73-78, 2010.

[28] K. Ćosić et al., “Virtual reality adaptive stimulation in stress resistance training,” Proceedings RTO-MP-HFM-205 on "Mental Health and Well-Being across the Military Spectrum", 2011.

[29] J.A. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, Vol. 39, pp. 1161-1178, Dec. 1980.

[30] B.J. Vaughan, C. Cullen, S. Kousidis and J. McAuley, "Emotional speech corpus construction, annotation and distribution," The 6th edition of the Language Resources and Evaluation Conference Marrakech (Morocco), 2008.

[31] 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.

[32] M.M. Bradley and P.J. Lang, “The International Affective Digitized Sounds, 2nd edn. (IADS-2): Affective ratings of sounds and instruction manual,” Technical report B-3. University of Florida, Gainesville, FL, 2007.

[33] S.B. Davis, and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 28, pp. 357-366, Aug. 1980.

[34] A.P. Dempster; N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B (Methodological) 39, pp. 1–38, 1977.

[35] L.E. Baum and T. Petrie, “Statistical interface for probabilistic functions of finite state Markov chains,” The Annals of Mathematical Statistics, vol. 37, pp. 1554-1563, 1966.

[36] M.T. Chmura, “Automatic classification of discrete emotional states based on acoustic speech properties,” Diploma Thesis, in Croatian, University of Zagreb, Croatia, Jun. 2011.

[37] A. Kolak, “Automatic classification of discrete emotional states based on linguistic speech properties,” Diploma Thesis, in Croatian, University of Zagreb, Croatia, Jun. 2011.

[38] B. Dropuljić, “Development of acoustic and lexical model for automatic speech recognition for Croatian language”, Diploma Thesis, in Croatian, University of Zagreb, Croatia, Jan. 2008.

[39] B. Dropuljić and D. Petrinović, “Development of acoustic model for Croatian language using HTK”, Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, vol. 51, pp. 79-88, 2010.