Original scientific paper / Izvorni znanstveni rad
Manuscript received: 2016-12-15
Revised: 2016-12-20
Accepted: 2016-12-22
Pages: 85 - 96

Croatian Emotional Speech Analyses on a Basis of Acoustic and Linguistic Features

Branimir Dropuljić
IN2data
Zagreb, Hrvatska
branimir.dropuljic@in2data.eu

Sandro Skansi
IN2data
Zagreb, Hrvatska
sandro.skansi@in2data.eu

Robert Kopal
IN2data
Zagreb, Hrvatska
robert.kopal@in2data.eu

Abstract: Acoustic and linguistic speech features are used for emotional state estimation of utterances collected within the Croatian emotional speech corpus. Analyses are performed for the classification of 5 discrete emotions, i.e. happiness, sadness, fear, anger and neutral state, as well as for the estimation of two emotional dimensions: valence and arousal. Acoustic and linguistic cues of emotional speech are analyzed separately, and are also combined in two types of fusion: a feature level fusion and a decision level fusion. The Random Forest method is used for all analyses, with the combination of Info Gain feature selection method for classification tasks and Univariate Linear Regression method for regression tasks. The main hypothesis is confirmed, i.e. an increase of classification accuracy is achieved in the cases of fusion analyses (compared with separate acoustic or linguistic feature sets usages), as well as a decrease of root mean squared error when estimating emotional dimensions. Most of other hypothesis are also confirmed, which suggest that acoustic and linguistic cues of Croatian language are showing similar behavior as other languages in the context of emotional impact on speech.

Keywords: emotional state estimation, acoustic and linguistic speech features, feature fusion, Croatian emotional speech

INTRODUCTION

Emotional speech features can be basically divided into two groups:

- linguistic features – measures of the verbal expression; extracted from the content of speech expression in the format of keywords, keyphrases or higher level indicators that are inherent to certain emotional states of the speaker;
- acoustic (or paralinguistic) features – measures of the variation of verbal expression, i.e. the modality of speech; often related, but not limited, to speech prosody, i.e. speech intonation, tone, stress and rhythm.

Acoustic features are more often used in literature to identify emotions. It is claimed that acoustic features are more universal, i.e. that they depend to a lesser extent on the language and are less susceptible to voluntary control. Authors mainly achieve higher accuracies with acoustic features when classifying emotions, but some papers have also reported that the estimation accuracy can be improved when combining acoustic and linguistic cues from the spoken expression [8], [9], [10], [12], [13]. As these two speech components are not necessarily correlated, their fusion can provide an integral information about the impact of emotions on speech.

A fusion can be generally done on two levels:

- on a feature level (as done in [8] and [13]), by constructing a large feature vector with acoustic and linguistic features combined – the problem with such approach is the potential of having to face the curse of dimensionality due to the increase in the feature dimension [7];
- on a decision level, where a simple fusion ([9], [10], [12]) or a discriminative approach ([12]) can be applied.

In the case of a simple fusion, output probabilities of individual acoustic and linguistic classifiers are averaged for each emotion, followed by an adjacent maximum likelihood decision (as described in [9] in details). This approach, however, neglects the fact that for each emotion the prior confidences in acoustical and language-based estimations differ. A discriminative approach on the other hand helps to integrate the knowledge of all accessible emotion confidences in one decision process, e.g. with the use of another machine learning model, as suggested in [12].

Furthermore, the evidences indicate that acoustic cues of speech are more related to the archetypal or primary emotions (i.e. happiness, sadness, fear, anger, surprise and disgust) and that linguistically based cues are more related to non-archetypal or secondary emotions (i.e. jealousness, pride, etc.) [1]. One possible explanation by Cowie et al. is that this may be because “archetypal emotions are signaled paralinguistically and others by linguistic signs.” Given that the secondary emotions are formed in parallel with the development of culture, and therefore the language [2], these results have a foothold.

Two general conclusions, related to estimation of valence and arousal levels (two emotional dimensions) based on speech features, are evident in [8]: a more accurate result

is achieved when estimating arousal than valence; and acoustic features are better for estimation of those two dimensions than linguistic ones. It can additionally be observed that the superiority of acoustic features decreases in the case of valence estimation, which suggests that linguistic features are relatively important for the estimation of this emotional dimension.

The core of this paper is the fusion analysis of acoustic and linguistic features in the case of utterances from Croatian Emotional Speech (CrES) Corpus. Fusion is performed on a feature level, and also on a decision level (where a simple fusion is implemented). Several hypotheses are tested in the context of classifying five discrete emotions (happiness, sadness, fear, anger and neutral state), as well as in the context of estimating valence and arousal using both, acoustic and linguistic speech features. Hypotheses are defined as follows:

1. Classification accuracy of five discrete emotions is higher when using acoustic features, compared with linguistic ones.
2. A fusion of acoustic and linguistic approaches improves classification accuracy, compared with individual acoustic or linguistic approach.
3. Root mean squared error (RMSE) of valence and arousal estimation is smaller when using acoustic features, compared with linguistic ones.
4. RMSE of arousal estimation is smaller than RMSE of valence estimation for both, acoustic and linguistic features.
5. A relative RMSE difference between linguistic and acoustic features ($RMSE_{LIN}$ – $RMSE_{ACO}$) is smaller in the case of valence estimation, compared with arousal estimation.
6. A fusion of acoustic and linguistic approaches decreases RMSE for both, valence and arousal estimation.

The Random Forest method is used for the classification task and Random Forest regression method for the estimation of valence and arousal. Collection and annotation processes of CrES corpus are described in the next chapter. The following chapters describe acoustic and linguistic feature sets and analyses results.

CROATIAN EMOTIONAL SPEECH CORPUS

The Croatian emotional speech (CrES) corpus was collected and emotionally annotated from various prerecorded sources. 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. A detailed description of building the initial version of the corpus is presented in [6], and an upgraded version, which will be used in this paper, is presented in [4]. This upgraded version contains total of 1140 utterances from 341 different male and female speakers with the total duration of approximately 85

minutes. The average duration of utterances is thus around 4.5 s. Audio data, which will be used for acoustic analysis, was stored in wav format with 11025Hz sampling frequency, 16 bits per sample, monaural. Manually written transcriptions are affiliated to all the utterances and will be used for linguistic analysis.

The utterances were initially categorized during the collection process into five emotion categories: happiness, sadness, fear, anger and neutral state (the first column in Table 1). They were then labeled with final categorization and valence and arousal level, and also filtered on a basis of subjective opinion of 109 annotators. At least 10 annotations were obtained for each utterance for both, discrete and dimensional representations of emotions.

The annotation process is described in [4] in details, but annotations used in this paper are slightly modified. In previous paper, the relevance factor of each discrete emotion annotation include the annotators' opinions about the acoustic richness in the expression, while in this paper, this information is excluded as emotion recognition is performed on a basis of both acoustic and linguistic features. Furthermore, in this paper, it is not important that the same utterances are used for discrete emotion classification and for valence and arousal estimation, so the separate filtering criteria are used.

Discrete emotion annotations were thus 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 the 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 $rel(a)$, which is in this paper defined as a certainty level of the annotator about his/her annotation. Each of these two factors (*sel* and *rel*) 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 further analysis. In this way, a total of 937 utterance remain for further analysis of discrete emotions, which can be seen in the second column in Table 1.

Table 1: Distribution of CrES corpus utterances per discrete emotions

Emotion	Number of Utterances per Emotion	
	Collection phase	***Annotation phase***
Happiness	249	166
Sadness	205	173
Fear	199	123
Anger	287	292
Neutral state	200	183
Total	**1140**	**937**

Valence and arousal labels of the corpus utterances were defined as centroids of 2D Gaussians, as described in [4] in details. The *agreement* filtering criterion is also applied here in a way that samples are not removed if at least 50% of annotators annotate valence and arousal levels with the relevance of at least ½. All 1140 utterances remained this way for valence and arousal analysis.

Additionally, 9 utterances were removed from the dataset as it was not possible to calculate acoustic features. Three of them were already removed within the filtering process for discrete emotion analysis and six of them are from the 'neutral state' subset. A total of 931 utterances are thus used as a final dataset for the discrete emotion analysis (166 + 173 + 123 + 292 + 177), and a total of 1131 utterances are used for valence and arousal estimation.

ACOUSTIC FEATURES

Acoustic features are extracted by using the open-source Emotion and Affect Recognition (openEAR) toolkit's feature extracting backend openSMILE [14]. A total of 1941 features were thus calculated for each utterance, by using the config file created for the 1st International Audio/Visual Emotion Challenge and Workshop (AVEC 2011) [15]. The Config file is slightly modified in order to extract one feature vector for the whole utterance, while within the original file, feature vectors were computed per user-uttered word, analogous to the periods for which audio labels are computed.

The acoustic feature set is composed of 25 energy and spectral related low-level descriptors (LLD) x 42 functionals, 6 voicing related LLD x 32 functionals, 25 delta coefficients of the energy/spectral LLD x 23 functionals, 6 delta coefficients of the voicing related LLD x 19 functionals, and 10 voiced/unvoiced durational features. Details for the LLD and functionals are given in Table 2 and Table 3 respectively. A standard range of commonly used acoustic features in emotional speech recognition is thus used [11]. It must be noted that LLD are actually contours, or time series, calculated from the speech signal, which is processed frame-by-frame in overlapping intervals. In this case, 60 ms and 25 ms frames are used, with the framerate of 100 fps (frames-per-second). Functionals (features) are, on the other hand, calculated from the contours of the whole utterance, i.e. only once per each utterance.

Table 2: Low-level descriptors (adapted from [16])

Energy & Spectral
loudness (auditory model based)
zero crossing rate
energy in bands from 250 – 650 Hz, 1 kHz – 4 kHz
25%, 50%, 75% and 90% spectral roll-off points
spectral flux, entropy
spectral variance, skewness, kurtosis
psychoacousitc sharpness, harmonicity
mel-frequency cepstral coefficients (MFCC) 1-10
Voicing related
fundamental frequency (F_0): subharmonic-sampling (SHS) and Viterbi smoothing
voicing probability
jitter, shimmer (local)
jitter (delta: "jitter of jitter")
log. harmonics-to-noise ratio (HNR)

Table 3: Functionals (adapted from [16])

Statistical Functionals
arithmetic mean, root quadratic mean
standard deviation, flatness
skewness, kurtosis
quartile 1-3
inter-quartile ranges
1%, 99% percentile
percentile range 1%–99%
%-age of frames contour is above min. + 25, 50, and 90 % of range
%-age of frames contour is rising
max, mean, min segment length
standard deviation of segment length
Regression functionals
linear regression slope, linear regression approximation error
quad. reg. c1, and lin. app. err.
Local min/max related functionals
mean and standard deviation of rising and falling slopes (min to peak)
mean and standard deviation of inter peak distances
amplitude mean of peaks
amplitude mean of minima
amplitude range of peaks
Linear Predictive Coefficients (LPC) functionals
LP gain, LPC 1-5

LINGUISTIC FEATURES

A standard Bag of Words (BOW) model is used with 3870 words for the transcription and is augmented with several additional handcrafted features. The simplest features we used are the percentage of unique words per utterance, the vowel-to-consonant ratio and the average number of characters per word for each utterance. We also used the number of most frequent vowel divided by (i) the number of total characters and (ii) the total number of vowels. Lastly, we included a feature which counts the number of special vocalizations in utterance. These special vocalizations are given in Table 4.

Table 4: Special vocalizations

snuffler	"SMRC"
weeping	"CMIZDR"
prolongations	"β"
inhaling	"UDH"
exhaling	"IDH"
cough	"KMH"
shouts	"HA", "HM","MHM", "EA","CC", "OO", "OOO", "HO", "HEJ", "EJ","OJ"
laughter	"HAHA", "HEHE", "HOHO"
growl	"GRRR"

The vocalizations themselves have already been parsed and treated as words by the BOW model, so we did not include any feature that counts them as such, but only this feature which returns the cumulative count. The reasoning behind this feature is that the corpus has special annotations such as "HAHA" and "HEHE" which are treated as different words, but utterances showing multiple special vocalizations are emotionally imbued so ignoring this fact might result in missing some emotional aspects of the transcription.

ANALYSES

We have analyzed emotional states from the CrES utterances in a form of the classification of 5 discrete emotions (happiness, sadness, fear, anger and neutral state), as well as the estimation of valence and arousal level ranging from 1 to 9. Previous analyses of CrES utterances were reported in [3], [4], [5] and [6]. CART based Random Forests are used in this paper with 100 trees and no splitting of subsets with less than 100 instances. The analysis is conducted using a 10-fold stratified cross-validation.

An analysis of the acoustic features is conducted, as well as an analysis of the linguistic features, both as full sets and with only the top 100 most relevant features for each case (selected with Info Gain method for classification and Univariate Linear Regression method for regression). Finally, two fusion analyses are made, one on a feature level,

with the top 100 acoustic and top 100 linguistic features, and one as a simple decision level fusion, described in the Introduction.

Results are presented in the following tables. Classification results in a form of an accuracy and also F_1 score, precision and recall averaged over classes, are presented in Table 5. Tables 6 to 9 present confusion matrices for classifying five discrete emotions by using 100 most relevant acoustic features, 100 most relevant linguistic features, a fusion of 100 most relevant acoustic and 100 most relevant linguistic features, and a simple decision level fusion, respectively. Table 10 presents results of valence and arousal estimation in a form of mean squared error, root mean squared error, mean averaged error and R^2.

Table 5: Results of discrete emotions classification

		Classification accuracy [%]	F_1 score [%]	Precision [%]	Recall [%]
Acoustic features	Full set	58.9	58.0	59.4	58.9
	100 features (Info Gain)	58.0	57.2	58.0	58.0
Linguistic features	Full set	52.3	48.5	62.3	52.3
	100 features (Info Gain)	53.2	50.4	53.6	53.2
Fusion	Feature level fusion	59.9	59.2	60.5	59.9
	Simple decision level fusion	65.4	64.1	67.9	65.4

Table 6: Confusion matrix for discrete emotions classification using 100 most relevant acoustic features

		Predicted Emotion					**Total**
		H	***S***	***F***	***A***	***N***	
Actual Emotion	***H***	**64**	13	4	76	9	**166**
	S	20	**78**	19	42	14	**173**
	F	9	22	**59**	24	9	**123**
	A	28	13	18	**209**	24	**292**
	N	2	6	4	35	**130**	**177**
Total		**123**	**132**	**104**	**386**	**186**	**931**

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

Table 7: Confusion matrix for discrete emotions classification using 100 most relevant linguistic features

		Predicted Emotion					Total
		H	***S***	***F***	***A***	***N***	
Actual Emotion	***H***	**78**	17	1	55	15	**166**
	S	14	**66**	5	65	23	**173**
	F	3	24	**10**	75	11	**123**
	A	8	20	11	**224**	29	**292**
	N	2	7	0	51	**117**	**177**
Total		**105**	**134**	**27**	**470**	**195**	**931**

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

Table 8: Confusion matrix for discrete emotions classification using 100 most relevant acoustic features and 100 most relevant linguistic features (feature level fusion)

		Predicted Emotion					Total
		H	***S***	***F***	***A***	***N***	
Actual Emotion	***H***	**65**	10	5	78	8	**166**
	S	15	**81**	22	41	14	**173**
	F	6	20	**64**	27	6	**123**
	A	25	9	19	**215**	24	**292**
	N	1	7	2	34	**133**	**177**
Total		**112**	**127**	**112**	**395**	**185**	**931**

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

Table 9: Confusion matrix for discrete emotions classification based on decision level fusion of acoustic and linguistic classifiers' outputs

		Predicted Emotion					Total
		H	***S***	***F***	***A***	***N***	
Actual Emotion	***H***	**79**	10	2	64	11	**166**
	S	10	**92**	9	52	10	**173**
	F	4	19	**42**	51	7	**123**
	A	14	4	8	**252**	14	**292**
	N	0	3	0	30	**144**	**177**
Total		**107**	**128**	**61**	**449**	**186**	**931**

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

Table 10: Results of valence and arousal estimation

			Mean Squared Error (MSE)	Root Mean Squared Error (RMSE)	Mean Absolute Error (MAE)	R^2
Valence	Acoustic features	Full set	2.415	1.554	1.275	0.193
		100 features (Univar. Lin. Reg.)	2.356	1.535	1.236	0.213
	Linguistic features	Full set	2.306	1.519	1.191	0.229
		100 features (Univar. Lin. Reg.)	2.257	1.502	1.193	0.246
	Feature level fusion		2.123	1.457	1.172	0.291
Arousal	Acoustic features	Full set	1.076	1.037	0.836	0.561
		100 features (Univar. Lin. Reg.)	1.093	1.045	0.834	0.554
	Linguistic features	Full set	1.970	1.403	1.154	0.195
		100 features (Univar. Lin. Reg.)	2.002	1.415	1.167	0.182
	Feature level fusion		1.081	1.040	0.830	0.558

CONCLUSION

Five of six hypotheses are confirmed and one (concretely, the third one) is partially confirmed in this paper. This suggest that acoustic and linguistic cues of Croatian language are showing similar behavior as other languages in the context of emotional impact on speech. More precisely:

1. Classification accuracy of five discrete emotions is higher when using acoustic features, compared with linguistic ones, i.e. 100 most relevant acoustic features results with 58% accuracy, while the set of 100 linguistic features results with 53.2% accuracy.
2. A fusion of acoustic and linguistic approaches improves classification accuracy, compared with individual acoustic or linguistic approach, i.e. the accuracy when using feature level fusion method is 59.9% and when using simple decision level fusion is 65.4%, which are both better results than individual accuracies presented above.
3. RMSE of arousal estimation is smaller when using acoustic features, compared with linguistic ones ($RMSE_{ACO}$ for 100 most relevant features is 1.045 and $RMSE_{LIN}$ for 100 most relevant features is 1.415), but RMSE of valence estimation turns

to be slightly higher when using acoustic features, compared with linguistic ones ($RMSE_{ACO}$ for 100 features is 1.535 and $RMSE_{LIN}$ for 100 features is 1.502).

4. RMSE of arousal estimation is smaller than RMSE of valence estimation for both, acoustic and linguistic features, i.e. the RMSEs for acoustic and linguistic sets are 1.535 and 1.502 respectively when estimating valence, and the RMSEs for those two sets are 1.045 and 1.415 when estimating arousal.
5. A relative RMSE difference between linguistic and acoustic features is smaller in the case of valence estimation, compared with arousal estimation, i.e. $RMSE_{LIN} - RMSE_{ACO} = -0.033$ for valence estimation, while $RMSE_{LIN} - RMSE_{ACO} = 0.37$ for arousal estimation.
6. A fusion of acoustic and linguistic approaches decreases RMSE for both, valence and arousal estimation, i.e. RMSEs of valence and arousal estimations using feature fusion sets are 1.457 and 1.040 respectively, which is slightly lower than RMSEs of estimations with individual acoustic and linguistic sets presented above (1.535 and 1.502 for valence and 1.045 and 1.415 for arousal).

REFERENCES

[1] Cowie, R. et al. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32-80.

[2] Damasio, A.R. (1999). *The feeling of what happens: body and emotion in the making of consciousness*. New York: Harcourt Brace & Company.

[3] Dropuljić, B. (2014). *Emotional state estimation based on data mining of acoustic speech features*, PhD Thesis (in Croatian language), University of Zagreb.

[4] Dropuljić, B., Popović, S, Petrinović, D. and Ćosić, K. (2013). Estimation of Emotional States Enhanced by A Priori Knowledge. *4th IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2013)*, pp. 481-486.

[5] Dropuljić, B., Skansi, S. and Kopal, R. (2016). Analyzing Affective States using Acoustic and Linguistic Features. *Proceedings of Central European Conference on Information and Intelligent Systems (CECIIS)*, pp. 201-206.

[6] Dropuljić, B., Thomasz Chmura, M., Kolak, A. and Petrinović, D. (2011). Emotional Speech Corpus of Croatian Language. *IEEE International Symposium on Image and Signal Processing and Analysis (ISPA 2011)*, pp. 95-100.

[7] Duda, R., Hart, P. and Stork, D. (2001). *Pattern Classification*, 2nd ed. New York: Wiley.

[8] Eyben, F., Wollmer, M., Graves, A., Schuller, B., Douglas-Cowie, E. and Cowie, R. (2010). Online Emotion Recognition in a 3-D Activation-Valence-Time Continuum using Acoustic and Linguistic Cues. *Journal on Multimodal User Interfaces*.

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

[10] Lee, C.M. and Pieraccini, R. (2002). Combining acoustic and language information for emotion recognition. *Proceedings of the ICSLP 2002*, Denver, CO, USA, 2002.

[11] Schuller, B. et al. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, vol. 53, no. 9, pp. 1062-1087.

[12] Schuller, B., Rigoll, G. and Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *Proceedings of the ICASSP '04*, pp. 577-580.

[13] Schuller, B., Villar, R.J., Rigoll, G. and Lang, M. (2005). Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. *Proceedings of the ICASSP '05*, pp. 325-328.

[14] (2016-12-04) http://audeering.com/technology/opensmile/

[15] (2016-12-04) http://sspnet.eu/avec2011/

[16] (2016-12-04) http://sspnet.eu/avec2011/challenge-data/