

Evaluation of Speech Perturbation Features for Measuring Authenticity in Stress Expressions

Branimir Dropuljić, Leo Mršić^(✉), Robert Kopal, Sandro Skansi,
and Andrijana Brkić

IN2data Data Science Company Ltd., Zagreb, Croatia
{branimir.dropuljic,leo.mrsic,robert.kopal,
sandro.skansi,andrijana.brkic}@in2data.hr

Abstract. Expressions can vary by the authenticity level, i.e. the real amount of emotion present within the person when expressing it. They are often sincere, and thus authentic and natural; the person expresses what he/she feels. But play-acted expressions are also present in our lives in a form of deception, movies, theater, etc. It was shown in the literature that those two type of expressions are often hard to distinguish. While some studies concluded that play-acted expressions are more intense, exaggerated or stereotypical than the natural ones, other authors failed to detect such a behavior. The goal of our analysis is to investigate whether speech perturbation features, i.e. jitter, shimmer, variance and features of disturbances in laryngeal muscle coordination, can be used as a robust measure for the analysis of the stress expression authenticity. Two subsets of the SUSAS database (Speech Under Simulated and Actual Stress) – the Roller-coaster subset and the Talking Styles Domain – are used for this purpose. It was shown that perturbation features in general show statistically significant difference between realistic and acted expressions, only the jitter features generally failed to discriminate these two type of expressions. The rising trend of perturbation feature values is observed from acted- to real-stress expressions.

Keywords: Speech perturbation features · Authenticity analysis · Emotional stress · Speech under stress

1 Introduction

The purpose of this paper is to evaluate speech perturbation features in the context of stress (or emotional) expression authenticity; also to discover whether such features can be used for distinguishing authentic (realistic) expressions from those that are acted. Of course, acted expressions are not necessarily nonauthentic. “Although natural expressions are partly staged, acted expressions are also partly natural.” [1] Actors’ portrayals can be influenced by subjective feelings, especially when produced via techniques based on emotional imagination or memory [2,3]. It was therefore argued that authentic expressions and play-acted ones are sometimes very difficult to distinguish [3].

Only a few studies compare these two types of expressions [4–8] and most of them concluded that play-acted expressions are more intense, exaggerated or stereotypical [7, 9–12]. E.g. it was shown in [13] that speech fundamental frequency (F_0) contour varies more and is generally higher in play-acted expressions than in authentic ones. Additionally, differences between professional and non-professional actors' expressions were analyzed and compared to the authentic expressions. It was shown that in terms of the acoustic characteristics, vocal expressions delivered by professional actors were not more similar to authentic expressions than the ones by non-actors [14]. Furthermore, the results do not support the view that play-acted expressions are necessarily stereotyped caricatures of authentic expressions. The similar findings were reported in [5, 15, 16], in which authors failed to detect an exaggerating behavior in the case of acting expressions.

The perturbation features (i.e. jitter and shimmer) were included in the authentic vs. play-acted expression analysis in [17] and the results show that there is no statistically significant difference between these two type of expressions for jitter measure, and that there is a significant falling trend from acted to authentic expressions for shimmer measure in the case of vowel “e” (the vowel “a” did not result in any difference). The authors used acted utterances produced by professional actors under four emotional states: ‘fear’, ‘anger’, ‘joy’ and ‘sadness’, and the same emotions were extracted as authentic set from various emotional reportages about situations in the past. The authors expect higher arousal for acted expressions, which is related with the exaggerating behavior, and some of the “arousal features” show such a behavior, and some of them don't. E.g. shimmer and F_0 contour variability decrease from acted to authentic emotions, which is an indicator of arousal decrease, while several measures like formant bandwidths and peak frequencies indicate that arousal was increased.

There are many vocal components of emotional (arousal) expression. Some of them are under voluntary control and some of them are autonomic [18]. As real arousal is not present, or is minimally present, within the person in the case of acted expressions, we expect that the autonomic vocal features of arousal will have smaller values than in the case of authentic (realistic) expressions. On the other hand, features that are affected by the voluntary control (e.g. F_0 variations, energy/intensity of the utterance, etc.) can have smaller or larger values, depending on the style of the expression (e.g. larger when a person is exaggerating in the expression). As speech perturbations increase as the arousal (and thus emotional stress) increases [19, 20], perturbation features should show the rising trend from acted to real-stress expressions if include only of the autonomic component of perturbations, i.e. if they are not under voluntary control.

The goal of our analysis is to investigate how speech perturbation features behave in acted- and real-stress conditions, compared to neutral state. We also investigate whether the features can be used as a robust measure for the analysis of the expression authenticity. An expanded set of perturbation features compared to [17] is proposed for this purpose, as will be described in the next section. Results presented in [21], indicates that some perturbation features are

less affected by the voluntary control and, on the other hand, more related to autonomic disturbances in laryngeal muscle coordination. Such features should therefore not depend on the style of acted expression; they should be relatively good at distinguishing between acted and realistic expressions (show a rising trend) regardless of whether acted expressions are exaggerated or moderate.

2 Speech Perturbation Features

Four groups of speech perturbation features are used for the analyses purpose:

1. ***peSNS* features** – speech features of disturbances in laryngeal muscle coordination; 8 features proposed in [21]: ***var*** (variance of the perturbation contour), ***mean abs*** (mean absolute value of the perturbation contour), ***max abs*** (maximal absolute value of the perturbation contour), ***dur per sec (v)*** (total duration of the perturbation intervals per second of a voiced speech), ***dur per sec*** (total duration of the perturbation intervals per second), ***mean dur*** (mean duration of the perturbation intervals), ***max dur*** (maximal duration of the perturbation intervals) and ***quan*** (perturbation quantity: the product of ***mean abs*** and ***dur per sec (v)*** features)
2. ***jitter* features** – measures of period-to-period fluctuations in glottal-cycle durations; 4 features calculated using the Praat functions [22]: ***local***, ***local (abs)***, ***rap*** and ***ppq5***
3. ***shimmer* features** – measures of the period-to-period variability of the speech amplitude value; 5 features calculated using the Praat functions [22]: ***local***, ***local (dB)***, ***apq3***, ***apq5*** and ***apq11***
4. ***variation* features** – 2 features: ***var diff*** (F_0) (variance of the first differential of F_0 contour) and ***std diff*** (F_0) (standard deviation of the first differential of F_0 contour)

Jitter, shimmer and variation features are state-of-the-art speech perturbation features for emotion and stress recognition [23]. *peSNS* features are proposed in [21] by the research group of which the authors of this paper are members, and are derived from F_0 contour decomposition described in [24].

3 Analyses

For the purpose of our analyses, we combined two subsets of the SUSAS database – the Actual Speech Under Stress Domain (Roller-coaster subset) and the Talking Styles Domain [25].

The Roller-coaster subset of the database is defined as actual high level stress and contains utterances of seven speakers (four male and three female) pronouncing 35 keywords like “break”, “change”, “gain”, etc., typical for communication between pilots and air traffic controllers. They were recorded at the ground (neutral samples) as well as during the ride on the roller-coaster (stress samples). A total of 414 stress samples and 701 neutral samples are available from this

subset, not equally distributed between people. The Talking Styles Domain is defined as simulated stress. It consists of utterances of nine people (actors), all male, talking in seven different styles (acted samples): slow, fast, angry, question, soft, loud, clear, as well as talking normally thus producing a neutral speech (neutral samples). For each person and each style (plus neutral speech) there are 70 samples, which is in total $9 \times 8 \times 70 = 5040$ samples. The same 35 words as in the Roller-coaster subset are used.

3.1 1st Analysis: *Neutral – Acted – Stress*

We first combined the described data sets into one data set with three levels – *neutral* (composed of both neutral samples from the first and the second domain; t-test show that there is no statistically significant difference between these two sets at α level of 0.05), *acted* (a portion of the Talking Styles Domain) and *stress* (a portion of the stress samples from the Roller-coaster subset). The combined data set was equally sampled over different people and consists of 1242 utterances (414 neutral, 414 acted, 414 stress).

We assume that perturbations are higher in stressful situations than in neutral situations or while acting and not actually experiencing stress (or an emotion). It was shown in [21] that perturbation features (in general) show a rising trend from neutral to real-stress samples. The relative position of acted-stress samples compared to neutral and real-stress samples is the focus of this paper. A separate ‘neutral vs. acted’ and ‘acted vs. stress’ analyses were therefore performed. The aim is also to evaluate speech perturbation features in order to explore which of them are relevant for the authenticity analysis in stress expressions. All features were thus analyzed individually for all samples. The left- and right-tailed t-tests, as well as the support vector machines (SVM) method are used. We have chosen the SVM classifier because it generally outperforms other classifiers in emotional speech recognition tasks.

The trend is calculated separately for ‘neutral’ to ‘acted’ changes in perturbation level and ‘acted’ to ‘stress’ changes. A moderate rising trend (\nearrow) is declared if left-tailed t-test show statistically significant difference at α level of 0.05 and a moderate falling trend (\searrow) is declared if right-tailed t-test show statistically significant difference at α level of 0.05. Additionally, strong rising (\nearrow/\nearrow) and falling (\searrow/\searrow) trends are defined in cases where p values of left- and right-tailed t-tests, respectively, are smaller than 0.0001. If neither left- nor right-tailed t-tests show statistically significant difference, then no trend (=) is declared.

A two-class¹ SVM model is trained for each type of comparison (‘neutral vs. acted’; ‘neutral vs. stress’; and ‘acted vs. stress’) and for each feature separately, i.e. with the observation vector consisting of only one feature. LIBSVM implementation of the SVM is used [26]. The following parameters were applied:

¹ The reason we use three two-class classifications instead of a multiclass classification is that we want the classifier to perform as close to random as possible when discriminating the classes ‘neutral’ and ‘acted’, while giving a good discrimination of ‘stress’ at the same time.

a 10-fold cross-validation (CV) process ($k = 10$) was selected; the radial basis function (RBF) was used as a kernel function; γ and C were set to 1; the threshold ε was set to 0.001.

The results are presented in Table 1. As the main goal of the analysis is to investigate a feature potential for separating ‘acted’ and ‘stress’ classes, the best features from each perturbation category are marked gray in the table, defined on the basis of ‘acted vs. stress’ SVM accuracy. The best feature from the entire feature set (*peSNS*: the maximal duration of the perturbation intervals) is marked with a dark gray color in the table. The box plot of the feature is presented in Fig. 1.

Table 1. *Neutral-acted-stress* analysis ($N =$ neutral, $A =$ acted and $S =$ stress)

Feature	Trend		Accuracy		
	$N \rightarrow A$	$A \rightarrow S$	N vs. A [%]	N vs. S [%]	A vs. S [%]
<i>peSNS features</i>					
var	=	↗↗	53.98	78.20	74.97
mean abs	=	↗↗	51.20	76.24	73.24
max abs	↗	↗↗	51.20	80.42	76.83
dur per sec (v)	↗	↗↗	55.37	86.03	81.09
dur per sec	↗↗	↗↗	60.68	88.38	78.96
mean dur	↗	↗↗	52.34	84.60	80.03
max dur	↗	↗↗	53.48	87.47	82.42
quan	↗	↗↗	54.74	85.77	80.69
<i>jitter features</i>					
local	↗	=	50.95	52.61	51.66
local (abs)	=	↘	50.95	52.61	51.66
rap	↘	↗	49.30	52.61	51.66
ppq5	=	=	48.29	52.61	51.66
<i>shimmer features</i>					
local	=	↗↗	50.95	70.63	70.71
local (dB)	=	↗↗	55.88	69.06	68.18
apq3	↘	↗↗	50.95	61.88	68.18
apq5	=	↗↗	48.93	73.50	74.03
apq11	=	↗↗	50.70	61.10	57.26
<i>variation features</i>					
var diff (F_0)	↗	=	60.05	75.07	63.38
std diff (F_0)	↗↗	↗	53.86	72.45	65.91

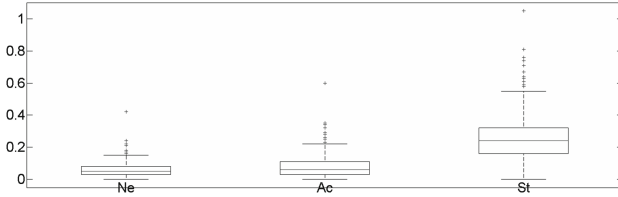


Fig. 1. Box plot of *peSNS max dur* feature distribution per three classes: *Ne* = neutral, *Ac* = acted and *St* = stress

It can be seen in the table that the SVM generally fails to discriminate jitter features. The deviation from the expected 50% is due to several NaN values in the features. Other feature groups are relatively good at separating acted-stress samples from the real-stress samples, while some of them also show significant rising trend from neutral to acted-stress samples. The possible explanation of such trend is that the talking style samples are partly realistic, but the possibility that perturbation features are partly affected by the voluntary control must not be disregarded. E.g. there is a possibility that some perturbations occur in voice when a person shout, even if there is no real anger behind. Results on a synthetic case, presented in [21], show that *peSNS* features have a potential for isolating autonomic perturbations from the voluntary components in speech, but further research must be done in this field.

One more parameter that must be taken into account is the level of stress, which a person tries to express during acting. The best way to compare the perturbations of acted- and real-stress samples is to ensure that the samples of these two data sets consist of the same level of stress expressions (not the real stress level as it is assumed that there is no real stress – or is minimal – in the case of acted expressions). The ‘loud’ talking style samples are thus chosen as acted-stress samples for the next analysis, as this style (according to our opinion) best suits the type and the level of stress within the roller-coaster samples.

3.2 2nd Analysis: *Neural – Loud – Stress*

Utterances from the ‘loud’ talking style were used as acted-stress samples for the purpose of this analysis. A total of 414 ‘loud’ utterances were randomly sampled from this subset, which form, together with 414 neutral samples and 414 real-stress (roller-coaster) samples, a new data set for the analysis. The ‘loud’ subset is used because the style and intensity of expression is very similar to the roller-coaster stress expression. It can be seen in Fig. 2 that the root-mean-square energy of a speech signal and a standard deviation of F_0 contour are rather similar for the samples in these two classes when compared with the neutral state.

The same SVM parameters and the same rules for trend calculation were used as for previous analysis. Results are presented in Table 2 and the best features from each feature group are again marked as gray. The ‘acted vs. stress’ SVM

accuracy is used as a measure for this. The best feature in this analysis (shimmer: the five-point amplitude perturbation quotient – apq5) is marked with a dark gray color in the table and the box plot is presented in Fig. 3.

It can be seen in the table that jitter features again results with the lowest accuracies of ‘acted vs. stress’ SVM classifies. Shimmer features performed better even with smaller perturbations in the ‘acted’ class compared with the ‘neutral’ class. Most of features show a statistically significant difference between ‘acted’ and ‘stress’ classes with the rising trend between these two classes. Again, the rising trend between the neutral and acted-stress samples (in this case the loud samples) is present and is even stronger in the case of *peSNS* features. A more detailed analysis per each speaking style is presented in the next section.

Table 2. *Neutral-loud-stress* analysis (N = neutral, A = acted and S = stress)

Feature	Trend		Accuracy		
	$N \rightarrow A$	$A \rightarrow S$	N vs. A [%]	N vs. S [%]	A vs. S [%]
<i>peSNS features</i>					
var	=	↗↗	60.69	78.20	72.46
mean abs	↗↗	↗↗	59.46	76.24	69.63
max abs	↗↗	↗↗	63.88	80.42	75.93
dur per sec (v)	↗↗	↗↗	65.85	86.03	74.39
dur per sec	↗↗	↗↗	72.60	88.38	68.08
mean dur	↗↗	↗↗	62.04	84.60	73.10
max dur	↗↗	↗↗	64.99	87.47	78.25
quan	↗↗	↗↗	62.53	85.77	76.96
<i>jitter features</i>					
local	=	=	48.53	52.61	53.28
local (abs)	↘	=	50.86	52.61	53.28
rap	↘	↗↗	50.86	52.61	53.28
ppq5	↘	↗↗	50.86	52.61	53.28
<i>shimmer features</i>					
local	↘	↗↗	60.44	70.63	79.28
local (dB)	↘	↗↗	67.08	69.06	79.28
apq3	↘	↗↗	47.91	61.88	76.58
apq5	↘	↗↗	48.89	73.50	81.60
apq11	↘	↗↗	54.42	61.10	73.49
<i>variation features</i>					
var diff (F_0)	=	↗	68.43	75.07	56.89
std diff (F_0)	↗	↗↗	70.15	72.45	61.13

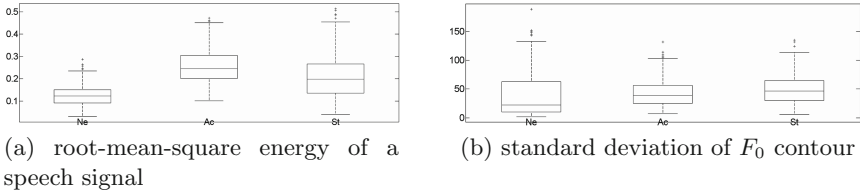


Fig. 2. Box plots of feature distributions per three classes: *Ne* = neutral, *Ac* = acted and *St* = stress

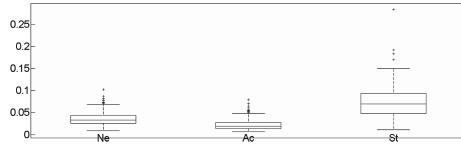


Fig. 3. Box plot of *shimmer apq5* feature distribution per three classes: *Ne* = neutral, *Ac* = acted and *St* = stress

3.3 3rd Analysis: Neural – 7 Speaking Styles – Stress

All seven speaking styles are compared with neutral and roller-coaster stress samples within this analysis. A total of 414 randomly chosen samples were selected for each speaking style class, i.e. for ‘angry’, ‘clear’, ‘fast’, ‘loud’, ‘question’, ‘slow’ and ‘soft’ type of stress expression. A data set is thus created with the total of $9 \times 414 = 3726$ samples.

It can be seen in Fig. 4 how samples are distributed for two best features from the previous analyses. As already mentioned, some perturbations exist in speaking styles and its level differs per styles and per features. The largest *peSNS max dur* perturbations are present for speaking style ‘angry’, while the largest *shimmer apq5* perturbations are in speaking style ‘soft’.

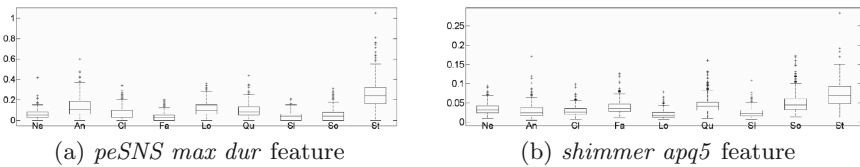


Fig. 4. Box plots of feature distributions per nine classes: *Ne* = neutral, *An* = angry, *Cl* = clear, *Fa* = fast, *Lo* = loud, *Qu* = question, *Sl* = slow, *So* = soft and *St* = stress

4 Conclusion

We can conclude that there is a statistically significant difference between various speaking style (acted) stress samples and realistic roller-coaster stress samples

for most of perturbation features. The rising trend of perturbations is achieved from acted to realistic samples. Only the jitter features generally failed to discriminate these two classes. When all talking-style samples were included in the same ‘acted’ class and compared with the real-stress samples, the maximal duration of *peSNS* perturbation intervals turns to be the most discriminative feature resulting with the highest ‘acted vs. stress’ SVM accuracy. The best feature when comparing these two classes (with only ‘loud’ speech samples included in ‘acted’ class) is the shimmer: five-point amplitude perturbation quotient (apq5). The main goal is thus achieved. We can say that *peSNS* and shimmer perturbation features can be used for distinguishing authentic (realistic) expressions from the acted ones. However, a significant rising trend in perturbation level is observed between neutral and acted samples, which varies per speaking style and per features. Such trend can appear as a result of two potential effects. First, maybe a significant amount of the expression authenticity was present within the speakers when pronouncing speaking style utterances and the perturbation features actually measure the authenticity level – the real stress behind the expressions. Second, a robustness of perturbation features for voluntary components in speech is only initially proved on a synthetic case in [21], so there is still a possibility that the features are partly affected by the voluntary control, i.e. by speaking loudly etc. In order to eliminate the voluntary component completely, a future research must be undertaken to analyze the specific type of perturbations from various neurological sources and also on the isolation of only autonomic, emotionally related, perturbations from *peSNS* intervals. In this way it will be possible to claim that the perturbation features measure pure authenticity of emotional or stressful expression.

References

1. Banse, R., Scherer, K.R.: Acoustic profiles in vocal emotion expression. *J. Personal. Soc. Psych.* **70**(3), 614–636 (1996)
2. Gosselin, P., Kirouac, G., Dore, F.Y.: Components and recognition of facial expression in the communication of emotion by actors. *J. Personal. Soc. Psych.* **68**, 83–96 (1995)
3. Scherer, K.R., Bänziger, T.: On the use of actor portrayals in research on the emotional expression. In: Scherer, K.R., Bänziger, T., Roesch, E. (eds.) *A Blueprint for an Affectively Competent Agent: Cross-Fertilization Between Emotion Psychology, Affective Neuroscience, and Affective Computing*, pp. 166–176. Oxford University Press, Oxford (2010)
4. Auberge, V., Audibert, N., Rilliard, A.: E-Wiz: a trapper protocol for hunting the expressive speech corpora in Lab. In: *Proceedings of the 4th LREC, Lisbon, Portugal* (2004)
5. Drolet, M., Schubotz, R.I., Fischer, J.: Authenticity affects the recognition of emotions in speech: behavioral and fMRI evidence. *Cogn. Affect. Behav. Neurosci.* **12**, 140–150 (2012)
6. Greasley, P., Sherrard, C., Waterman, M.: Emotion in language and speech: methodological issues in naturalistic approaches. *Lang. Speech* **43**, 355–375 (2000)

7. Laukka, P., Audibert, N., Auberge, V.: Exploring the determinants of the graded structure of vocal emotion expressions. *Cogn. Emot.* **26**, 710–719 (2012)
8. Williams, C.E., Stevens, K.N.: Emotions and speech: some acoustical correlates. *J. Acoust. Soc. Am.* **52**, 1238–1250 (1972)
9. Wilting, J., Krahmer, E., Swerts, M.: Real vs. acted emotional speech. In: INTER-SPEECH 2006, Pittsburgh PA, USA (2006)
10. Barrett, L.F.: Was Darwin wrong about emotional expressions? *Curr. Dir. Psychol. Sci.* **20**, 400–406 (2011)
11. Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E.: Desperately seeking emotions or: actors, wizards and human beings. In: ISCA Workshop on Speech and Emotion, Newcastle, Northern Ireland (2000)
12. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: towards a new generation of databases. *Speech Commun.* **40**, 33–60 (2003)
13. Drolet, M., Schubotz, R.I., Fischer, J.: Recognizing the authenticity of emotional expressions: F_0 contour matters when you need to know. *Front. Hum. Neurosci.* **8**, 1–11 (2014)
14. Jurgens, R., Grass, A., Drolet, M., Fischer, J.: Effect of acting experience on emotion expression and recognition in voice: non-actors provide better stimuli than expected. *J. Nonverbal Behav.* **39**, 195–214 (2015)
15. Jurgens, R., Drolet, M., Pirow, R., Scheiner, E., Fischer, J.: Encoding conditions affect recognition of vocally expressed emotions across cultures. *Front. Psychol.* **4** (2013)
16. Scherer, K.R.: Vocal markers of emotion: comparing induction and acting elicitation. *Comput. Speech Lang.* **27**, 40–58 (2013)
17. Jurgens, R., Hammerschmidt, K., Fischer, J.: Authentic and play-acted vocal emotion expressions reveal acoustic differences. *Front. Psychol.* **2**, 1–11 (2011)
18. Scherer, K.R.: Vocal correlates of emotional arousal and affective disturbance. In: Wagner, H., Manstead, A. (eds.) *Handbook of Psychophysiology: Emotion and Social Behavior*, pp. 165–197. Wiley, London (1989)
19. Scherer, K.R.: Nonlinguistic vocal indicators of emotion and psychopathology. In: Izard, C.E. (ed.) *Emotions in Personality and Psychopathology*, pp. 495–529. Plenum Press, New York (1979)
20. Scherer, K.R.: Vocal affect expression: a review and a model for future research. *Psychol. Bull.* **99**, 143–165 (1986)
21. Dropuljic, B., Petrinovic, D., Cosic, K.: Novel speech features of disturbances in laryngeal muscle coordination. In: 7th IEEE International Conference on Cognitive Infocommunications, pp. 175–180 (2016)
22. Praat. <http://www.fon.hum.uva.nl/praat/>
23. Schuller, B., et al.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Commun.* **53**(9), 1062–1087 (2011)
24. Dropuljic, B.: Emotional state estimation based on data mining of acoustic speech features. Ph.D. Thesis (in Croatian). University of Zagreb, Croatia (2014)
25. Hansen, J.H.L., et al.: The impact of speech under stress on military speech technology. *Nato Project 4* (2000)
26. LibSVM. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/>