25th International Congress on Sound and Vibration
8-12 July 2018 HIROSHIMA CALLING
ICSV25

WEAKNESSES OF VOICE BIOMETRICS – SENSITIVITY OF SPEAKER VERIFICATION TO EMOTIONAL AROUSAL

Milan Rusko, Marian Trnka, Sakhia Darjaa,

Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia

email: milan.rusko@savba.sk

Tim Stelkens-Kobsch and Michael Finke

Institute of Flight Guidance, German Aerospace Center (DLR), Braunschweig, Germany

In our series of experiments we study weaknesses of the voice biometric systems and try to find solutions to improve their robustness. The acoustical features that represent human voices in the current automatic speaker verification systems change significantly when the person's emotional arousal deviates from the neutral state. Speech templates of a given speaker used for enrollment are generally recorded in a neutral emotional state using "normal" speech effort. Therefore speaking with higher or lower voice tension causes a mismatch between training and testing resulting in a higher number of verification errors. The acoustical cues of increased emotional arousal in speech are highly non-specific. They are similar to those of Lombard speech, warning and insisting voice, emergency voice, extreme acute stress, shouting, and emotions like anger, fear, hate, and many others. As the available spontaneous emotional speech databases do not cover the full range of the emotional arousal for individual voices, and do not have enough utterances per speaker, we decided to use our CRISIS acted database containing speech utterances at six levels of tense emotional arousal per speaker. Sensitivity of the state of the art i-vector based speaker recognizer with PLDA scoring to arousal mismatch was validated. The speaker verification system was successfully implemented in the online "Speaker authorization" module developed in the frame of the European project Global ATM Security Management (GAMMA). It has been observed that at extreme arousal levels the reliability of the verification decreases. Mixed enrollments with various levels of arousal were used to create more robust models and have shown a promising improvement in the verification reliability compared to the baseline.
Keywords: biometry, speaker verification, stress, arousal

1. Introduction

The voice biometric systems seem to be sensitive to certain phenomena. We decided to identify and evaluate these weak points and find possible solutions improving the robustness of voice biometry. In our previous work [1] we studied vulnerability of the automatic speaker verification (ASV) systems to the attacks using speech synthesis. Attention of the scientists has recently been focused on the impact of emotions on biometric systems [2]. Here we present our work studying the robustness of ASV against changes in the speaker's emotional state, namely in the level of tense emotional arousal reflected in voice.

When analyzing vocal behaviour as a marker of affect we focus on the nonverbal aspects of speech. The basic assumption is that there is a set of objectively measurable voice parameters that reflect the affective state a person is currently experiencing [3], as most affective states involve physiological reactions, which in turn modify the voice production process. [4].

As we have formulated in [5], we use in our work a two dimensional model of emotional space, conceptualizing human emotions by their position in the dimensions of valence and arousal.

Thayer [6] distinguishes two dimensions of subjective arousal; energetic arousal and tense arousal. Energetic arousal is associated with readiness for vigorous and muscular-skeletal activation. Tense arousal represents a preparatory-emergency system, activated by some real or imagined danger.

The affective phenomenon, which has influence on speaker verification we study in this work, is the "tense emotional arousal" or "emergency-preparatory activation", which is one of the dimensions of the emotional space model and we will denote it in this work as arousal.

It is known that symptoms of increased emotional arousal in speech are non-specific. For instance with higher activation the speaker tends to speak louder; the acoustical cues can be similar to those of Lombard speech [7], warning and insisting voice, emergency voice, extreme acute stress, shouting, and emotions like anger, fear, hate, and many others.

Owren and Bachorowski [8] give an example of the acoustical cues of high-arousal speech: "participants who experienced high arousal … tend to show increased speech rate and a number of changes in the source energy of the sounds. For example, mean F_0 and F_0 range both increase, a larger number of the high-arousal pitch contours are found to be generally rising." They add that the number of inflection points increases, vocal perturbation goes up, which is reflected in higher jitter (cycle-to-cycle variation of fundamental frequency), shimmer (variability of the peak-to-peak amplitude) and lower harmonic-to-noise ratios. The vocal-fold vibration becomes noisier, causing an upward frequency shift both in short-term spectral moments and spectral tilt and in long-term spectral outcomes. They also observed that high arousal is associated with higher vocal amplitude. They have also pointed out that this cannot be quantified without a having common reference or calibration of the recording equipment.

2. Speaker verification system and its validation

The ASV system we used in this study was a standard state-of-the-art text-independent speaker verification system, exactly the same as we used in our former work of this series [1]. It was created using KALDI research toolkit [9]; the i-vector approach [10] was used with PLDA scoring [11]. LibriSpeech corpus [12] was used for UBM training (2500 English speakers, 3 minutes of speech per each).

A part of the VoxForge database [13] was used as a test set. 400 speakers were chosen as target speakers, and other 1500 speakers as "impostors" (non-target). One minute of speech per speaker was used for enrolment. The utterances were taken from different recording sessions when available. The total number of 80 000 test utterances was used with the length of 2 to 10 seconds.

In the ASV system validation test each of the enrolled speakers has been tested against his own utterances (target) and against utterances of each of all the other speakers (non-target). The verification worked well on the VoxForge data, with Equal Error Rate about 3%.

Similarly to [1] the Universal Background Model (UBM) was trained on English speakers, but according to the relatively language-independent nature of ASV task and the aim to evaluate the influence of the mismatch in arousal level between enrolment and test utterances (which both were in Slovak) it was possible to use in the experiments presented in Section 5 a cross-language approach with English UBM and Slovak expressive database for enrolment and testing.

2.1 ASV implementation in the GAMMA prototype and real-time simulation experiments

The research presented in this work was realized partly in a frame of the European project Global ATM Security Management (GAMMA). The goal of this was to develop solutions to emerging air traffic management vulnerabilities backed up by practical proposals for the implementation of

these solutions. The project also had to consider the new scenarios created by the Single European Sky program.

2.2 The GAMMA prototype validation

From 2016 to 2017 the ASV system described in this paper was integrated and tested in a set of air traffic control simulations conducted at the German Aerospace Center (DLR) in Braunschweig, Germany. Speaker verification from voice was successfully applied to typical air-ground voice communication which was simulated using a VOIP communication channel and involving one air traffic controller and several so called pseudo pilots (i.e. simulation pilots). The voice communication between pilots and air traffic controllers is standardized worldwide [14] and uses short and easily understandable phrases which have been designed to avoid any misunderstandings even with poor audio quality. On the other hand, air-ground voice communication may be contaminated with background noise, disturbances or low voices from other people in the cockpit or in the control room. In a busy traffic environment, air-ground voice communication can be very frantic with lots of speakers making very short transmissions quickly one after the other.

During the mentioned simulation campaign, active air traffic controllers from the German air navigation service provider DFS and from the Romanian air navigation service provider ROMATSA were recruited to take part in the simulation runs as radar approach controllers. They were confronted with a medium to high traffic load modeled from real flight plan data for Düsseldorf Airport. Pilots were simulated by DLR employees who have been comprehensively trained. The pseudo pilots hold an aeronautical radio telephony certificate which allows the most realistic simulation of real air-ground voice communication. There were always at least 4 different speakers taking part in the simulation in all simulation runs. As a result, the speaker verification from voice biometrics was able to determine the identity of every known speaker. In addition to that, this simulated air-ground voice communication was intruded from time to time by an unauthorized speaker, which was detected by the system with a very high reliability.

2.3 Problems arise with extreme levels of arousal

The original assumption was that the attack of the false air traffic controller (ATCo) and the subsequent issuing of incorrect instructions will cause an increase in stress for both pilots ATCos, who in some geographical and flight conditions may not hear the false ATCo's communication with pilots. The problem is that pilots and ATCos are trained not to show their stress and cope with critical situations in a calm way as far as possible. Therefore, any significant signs of elevated stress were not recorded in their voice during the tests.

Anyway this was only a simulation and so their stress and emotions were not fully blown.

However, it can be assumed that extremely dangerous situations could occur in future in which the increased emotional activation in the voice would be manifested. For ethical reasons, it was impossible to expose the test subjects to real dangers and therefore we chose to use an acted speech. We asked subjects not to limit their voices, and try to play what a really emotional reaction might look like. At extremely high and low arousal levels, we immediately noticed an increase in the number of verification errors. It was obvious that more attention has to be paid to this issue to assess a potential degradation of ASV reliability at extreme arousal levels.

3. Expressive speech data – a new use for an existing database

For the evaluation of the influence of arousal on ASV a speech database was needed, that would fulfill two conditions. First it should have enough speakers to evaluate the error rate of the speaker verification. And second it should have enough utterances at all the levels of arousal per speaker. As the available spontaneous emotional speech databases do not cover the full range of the emotional arousal for individual voices, nor do they have enough expressive utterances per speaker, we decided to use our CRISIS acted database containing speech utterances portraying six levels of tense

emotional arousal. This database was originally designed at our department for research in acoustical cues of emotional arousal in speech and for the development of expressive speech synthesizers. It is still being enriched with further recordings, but the description of its structure and details on recording conditions can be found in [5].

The experiments presented in this work were done on the recordings of 22 speakers (12 male and 10 female). Each speaker recorded in two sessions two sets of utterances – first he recorded a set with increasing arousal (levels 1, 2, 3; 150 utterances per level) and in the second session he recorded a set with decreasing arousal (levels -1, -2, -3; 150 utterances per level). For higher tense arousal the texts of warning messages were used with lengths ranging from one word to four sentences. For lower tense arousal the sentences had soothing texts.

As the voice talent is often unable to keep the level of portrayed emotion consistent for a longer time we designed a three step method of recording of the expressive database [5]. In this method the speaker varies the emotional load three times with every utterance, producing triplets of lexically identical utterances trying to keep same steps in tense arousal levels. The speaker was instructed to utter the message once in a neutral manner ("level 1" - reference level for higher tense arousal triplet), then with higher imperativeness, like a serious command or directive ("level 2"), and finally like an extremely urgent warning, command or statement being declared in a situation when human lives are directly endangered ("level 3"). The parts of the database with corresponding levels of arousal are marked as a1, a2, and a3.

When recording the "lower tense arousal" triplets of utterances the speaker was instructed to utter the prompted message once in a neutral way, which was natural and comfortable for him. We assume that this level reflects the neutral state of the speaker at that particular recording session ("level -1" - reference level for low arousal). Then the speaker then has to imagine that he has to announce to a bigger group of adult people that the emergency situation has passed, and they can calm down and stay at ease ("level -2" - lower activation/emotional arousal). The speaker should then imagine that he is speaking to scared small children, or to a seriously ill or wounded person. His speech should not be motherese, or whispered, but has to be very peaceful ("level -3" - extremely low arousal). The parts of the database with corresponding levels of arousal are marked as b1, b2, and b3.

This acted database was chosen for our experiment, because it contains relatively well defined "discrete" levels of arousal. The second reason is that according to the instructions the subjects had to portray full-blown emotions and therefore utilize the full range of their expressive speech voice profile. To illustrate the acoustic cues of arousal in the speech signal we present in Fig. 1 histograms of the fundamental frequency F_0 (Hz) measured on one speaker's utterances with decreasing and increasing tense arousal adopted from [5]. F_0 was measured on 25 ms frames. Y axis (count) represents number of frames with a given F_0 in the database.

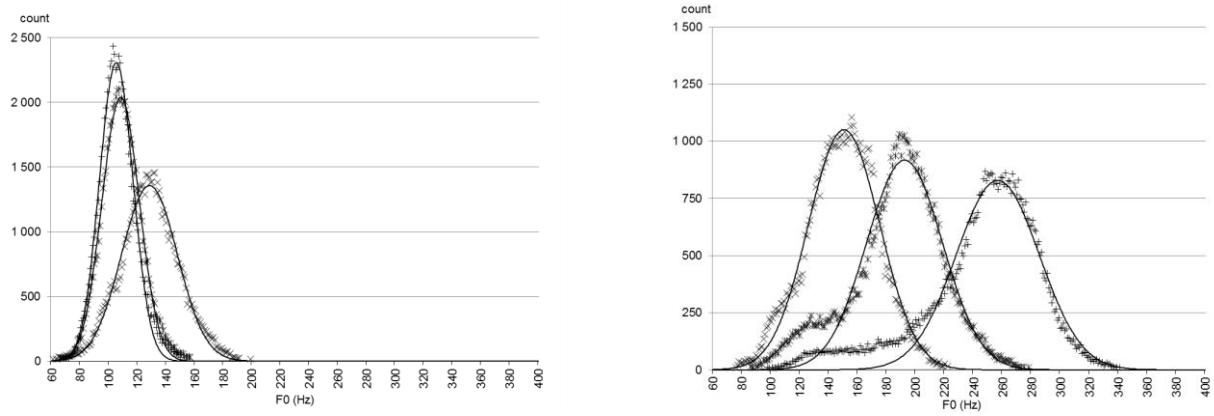

Figure 1: left - Histograms of F_0 from the utterances with decreasing arousal (from left: level -3, level -2, level -1); right: - Histograms of F_0 obtained from the measurements on the utterances with increasing arousal (from left: level 1, level 2, level 3); speaker *mr* (adopted from [5]).

Similar results were obtained for sound intensity represented with values of Sound Pressure Level (SPL) in dB SPL, i.e. dB relative to $2 \cdot 10^{-5}$ Pascal. 1000 Hz calibration signal was recorded at the beginning of every record. In Fig. 2 we present histograms of the SPL (dB) on the utterances with decreasing and increasing tense arousal. SPL was measured on 25 ms frames. Y axis (count) represents number of frames with a given SPL in the database. It can be seen, that:

a) there is a significant inter-session variability in between the increasing and decreasing arousal sessions
b) the range of both F_0 and SPL is much bigger for high levels of arousal, than for the low levels
c) differences between the two lowest arousal levels (-2 and -3) are very small in comparison to those between the higher arousal levels (2 and 3, or 2 and 1). As it was shown in [5], these trends can be observed in all speakers.

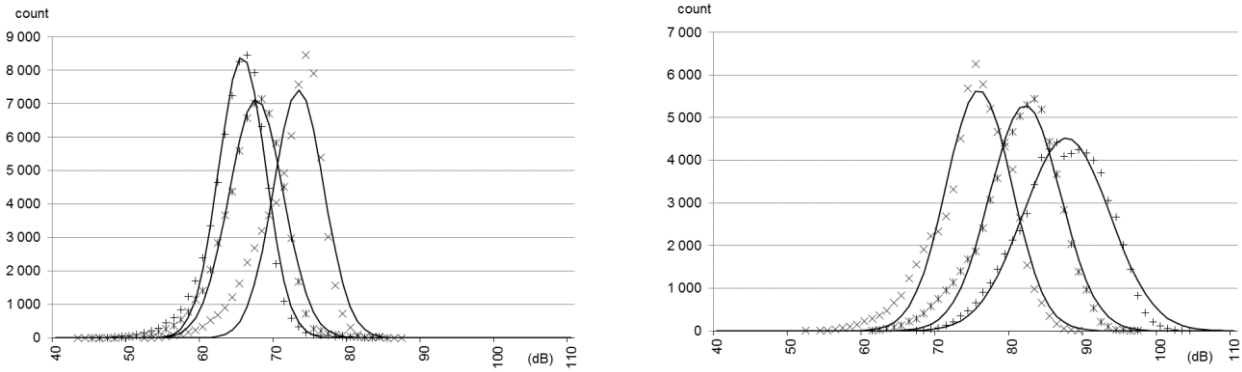

Figure 2: left: Histograms of SPL (dB) from the three databases with decreasing arousal (from left: level -3, level -2, level -1); right: Histograms of SPL obtained from the measurements on the three databases with increasing arousal (from left: arousal level 1, level 2, level 3); speaker mr. (adopted from [5])

4. Experiments

The speaker verification has two phases, enrolment and verification. During enrolment, the speaker's voice is recorded, representative acoustic features are extracted to form a voice template, or model. In the verification phase, a model is derived from the test utterance, which is then compared against the template created in enrolment. In general, the enrolment is made from emotionally neutral speech recorded in non-stressing conditions.

In our experiments, we tested the impact of a mismatch between the emotional activation level in speech utterances used for enrolment and for testing. In the second step, we tried to find out whether introducing non-neutral utterances in the enrolment will increase the robustness of the automatic speaker verification system against variability of arousal in the testing utterances. We tried to find out which mixture of arousal levels would be optimal for the enrolment.

Sixty-five utterances were used to represent one speaker at one arousal level. Twenty-two speakers were enrolled at each of the combinations of the six arousal levels. For instance in the enrolment mixture (combination of utterances) marked as “b1a123” the enrolment for each speaker consisted of 65 of his utterances at arousal level -1, together with the same number of utterances at levels 1, 2, and 3. Let us remind that the database parts containing arousal levels -3, -2, -1, 1, 2, 3, are labelled as b3, b2, b1, a1, a2, and a3 (for better readability of tables and graphs). All of them contain segregate training and testing sets of utterances.

The i-vector representation [15] is a fixed-length representation of speech utterance, which means that enrolments of different lengths result in models of the same size. Our approach used a frame-based fusion of the models by i-vector averaging. For every enrolment mixture an i-vector was computed for each speaker. This speaker template was then compared to the i-vectors representing test utterances of all speakers at all six arousal levels. Each decision of the ASV was evalu-

ated as correct or incorrect and the average equal error rate was computed for all tested arousal levels across all the speakers. (The test signals did not contain mixtures of arousal levels.) An average length of one utterance in the database was approximately 6 seconds. The most informative results of the experiments are presented in Tab. 1. where the error rates are reported (in percentage value) for each combination of enrolment/test set conditions. All the different enrolments (single arousals or their combinations) are reported in column while the test sets are reported in the first 6 rows. In rows 7 to 9, the average errors are reported for full test set, test set with both extreme arousal levels - highest (a3) and lowest (b3) excluded, and test set with only highest arousal level a3 excluded.

The results of the experiments indicated that mismatch in arousal between the enrolment and test utterances had a considerable impact on the reliability of the ASV.

Enrolment with neutral speech (a1 set) works pretty well except for the extreme arousal levels (test sets b3 and a3). When testing with b1, b2, a1, and a2 sets only, this model reached the smallest average error rate - 8.12% (see Average ERR b12a12 test in Tab. 1). It should be noted that the extreme values of arousal are rare in standard speech communication, and in most of the applications a system enrolled with neutral speech will function reliably.

For the full range of arousal we can see that the biggest part of errors is caused by the highest arousal – level 3 (test set a3). To decrease the error rate for this level, utterances with the highest arousal level have to be added to the enrolment. Enrolling only the utterances with the highest arousal is of course not sufficient as they exhibit the strongest mismatch to the other levels (see enrolment a3 in Tab. 1 or in Fig. 3). As expected, the lowest average verification error for the full range of arousal, 10.18%, was obtained with the complete mixture of all the arousal levels (see Average ERR full range test in Tab. 1).

The lowest level of arousal (b3) generally does not cause any extreme problems to the ASV, as its acoustical characteristics do not differ too much from those of the utterances in b2 and b1 sets. As the range of emotional activation displayed in these three sets seems to be relatively small, it has to be considered if they are not given too high weight in comparison to other levels, and if they are not overrepresented by introducing similar information several times in the enrolment. Moreover "b1,b2,b3" have also a lower variability of acoustical characteristics than "a1,a2,a3".

Table 1: Dependence of the error rates of the ASV for the test utterances of all six arousal levels for various enrolment mixtures.

Test set	**ENROLMENT**														
	b3	b2	b1	a1	a2	a3	b1a1	b2a2	b3a3	b123	a123	b123 a123	b1a123	b2a123	b3a123
Level -3 (b3)	5.18	7.06	11.89	12.13	15.39	21.4	8.12	9.65	8.02	7.53	15.31	9.76	14.92	12	10.94
Level -2 (b2)	**4.24**	**5.06**	9.54	9.62	14.51	20.98	**6.24**	7.41	**6.74**	**5.53**	13.66	7.64	12.39	**9.33**	9.36
Level -1 (b1)	10.01	10	**5.65**	7.13	11.85	20.06	7.18	**7.09**	7.29	6.47	10.1	**6.95**	**9.29**	12.4	**9.31**
Level 1 (a1)	16.02	14.07	12.51	**5.21**	11.24	21.92	10.73	13.83	20.73	13.88	9.21	12.5	12.24	12.02	12.21
Level 2 (a2)	17.38	15.45	14.19	10.52	**5.92**	14.04	13.6	10.93	14.67	15.26	**5.31**	12.03	10.46	10.67	10.9
Level 3 (a3)	23.04	25.8	25.8	22.51	22.03	**8.78**	27.34	23.34	15.92	23.43	12.18	12.18	18.58	17.85	17.6
Average ERR full range test	*12.65*	*12.91*	*13.26*	*11.19*	*13.49*	*17.86*	*12.2*	*12.04*	*12.23*	*12.02*	*10.96*	***10.18***	*12.98*	*12.38*	*11.72*
Average ERR b12a12 test	*11.91*	*11.15*	*10.47*	***8.12***	*10.88*	*19.25*	*9.44*	*9.82*	*12.36*	*10.29*	*9.57*	*9.78*	*12.59*	*12.45*	*10.45*
Average ERR b12a123 test	*14.14*	*14.08*	*13.54*	*11*	*13.11*	*17.16*	*13.02*	*12.52*	*13.07*	*12.91*	***10.09***	*10.26*	*12.65*	*12.44*	*11.88*

Assuming that "b1,b2,b3" reflect a significantly narrower range of tense arousal, we also made experiments omitting the b3 set from testing. We simply limited the low arousal utterances to b1 and b2 sets. When testing with utterances from "b2,b1,a1,a2, and a3" test sets only, the winner with the lowest mean error, 10.09%, is the enrolment using the mixture of a1, a2 and a3 levels (see row Average ERR b12a123 test in Tab. 1). Please note, that no utterance from the sets with lower arousal needed to be included in the enrolment in this case. The speaker characteristics of levels -1

and -2 were probably represented enough with arousal level 1. For easy comparison selected results are shown in a graphical form in Fig. 3.

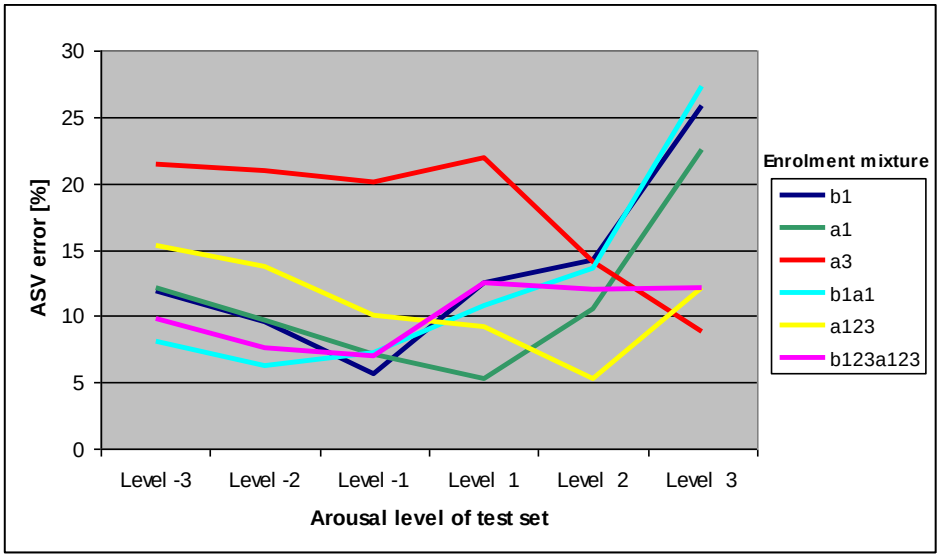

Figure 3.: Dependence of the error rates of the ASV on the composition of the enrolment mixture for six arousal levels

5. Discussion and conclusions

The effect of arousal on speech is very diverse and depends on conditions, situation and personality. We are aware of the fact that speech variability according to the arousal level is a continuous phenomenon in contrast to discrete classification used in the CRISIS database. It is also obvious that the results obtained on an acted database should not be generalized. However, until there is a sufficiently large, well annotated and representative database of spontaneous emotional vocal expressions available, we can at least formulate our first remarks.

One of the most surprising observations was that the results for b1 and a1 sets that were considered as neutral (arousal levels -1 and 1 respectively) were not approximately the same. These sets were assumed to have statistically very similar properties, as the instructions given to the recording subjects were the same for these two levels (normal, neutral, relaxed speech). The results of cross tests a1 versus b1 however show considerable differences. We assume that this is caused by inter-session variability between the session aimed at recording the increasing levels of arousal and the one recording the decreasing levels. The second reason may be different psychological setting of the speaker when he is getting ready to record the warning messages and calming messages. The results show that it can be problematic to create an enrolment representative enough for different sessions even for the speech that was uttered with an effort to speak in an emotionally neutral way.

The second important observation is that the acoustic characteristics of speech with the levels -1, -2, and -3 of arousal are similar, the ranges of F_0 and SPL are narrower and the negative influence of their variability on the ASV is smaller than that of levels 1, 2, and 3. We can hypothesize that this phenomenon may have its roots in the process of evolution. The high levels of arousal are used in the situations and emotional states, when both delivering the communication content of the message (e.g. warning others) and its highly expressive presentation (e.g. threatening the enemy) are very important, and can even save the life. The range of arousal levels is much wider for the vocal manifestations needed for survival, than those used in a calm state. We also present in the graph the results for the a3 enrolment. As expected, it gives acceptable results only for a3 test set as the acoustic characteristics of this highly aroused speech are very different from those of other levels.

The third observation to point out is that for current systems there is no universal instruction for assembling a mixture of speech utterances for enrolment. This depends on the application. The em-

phasis on speech with extreme arousal levels, which are relatively rare in real-world situations, can significantly deteriorate ASV for the neutral and slightly elevated or decreased arousal levels, that are frequent. On the other hand in the applications for extreme situations (e.g. help in acute danger) it may be needed to give higher weight to the extreme arousal levels.

We can conclude that the reliability of the ASV is strongly influenced by the mismatch in tense arousal between the enrolment and test speech utterances. The combination of speech with different levels of arousal in the enrolment is far from being the only way to compensate this effect, but is probably the easiest and most straightforward, and can help the ASV to work more consistently through a wider range of arousal level in expressive speech.

6. ACKNOWLEDGEMENT

The research leading to the results presented in this paper has received funding from the European Union FP7 under grant agreement n° 312382. More information can be found under http://www.gamma-project.eu/. The research of Slovak authors was also supported by the Slovak Agency for Science and Research, grant n° APVV-15-0517, and Slovak Scientific Grant Agency VEGA, grant n° 2/0197/15.

REFERENCES

1 Rusko, M., Trnka, M., Darjaa, S., Ritomsky, M., Weaknesses of voice biometrics - speaker verification spoofing using speech synthesis. *Proceedings of the 24th International Congress on Sound and Vibration*, IIAV, London, 8 p. (2017).
2 Pleva, M, Bours, P., Hladek, D., Juhar, J., Using current biometrics technologies for authentication in e-learning assessment, *Proceedings of 14th International Conference on Emerging eLearning Technologies and Applications ICETA*, Košice, (2016).
3 Patrik N. Juslin and Klaus R. Scherer : Speech emotion analysis. *Scholarpedia*, **3**(10):4240, (2008).
4 Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, **99**, 143-165.
5 Rusko, M., Trnka, M., Darjaa, S., Sabo, R., Ritomsky, M. Expressive speech synthesis for critical situations. *In Computing and informatics*, 2014, vol. **33**, no. 6, p. 1312-1332. (2014).
6 Thayer, R. E.. *The Biopsychology of Mood and Arousal.* New York: Oxford University Press, (1989) Appendix 1.
7 Simko, J., Benus, S., Vainio M. Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue, *The Journal of the Acoustical Society of America* **139**(1):151-162, (2016).
8 Owren, M. J., & Bachorowski, J.-A.: Measuring emotion-related vocal acoustics. *J. A. Coan & J. J. B. Allen (Eds.), Series in affective science. Handbook of emotion elicitation and assessment,* New York: Oxford University Press, pp. 239-266, (2007).
9 *Kaldi* [Online.] available: http://kaldi-asr.org/, January 2016
10 Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., Frontend factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process., vol. 19, Issue: 4, 2010.
11 Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S. PLDA based Speaker Recognition on Short Utterances. *Proc.Oddysey Speaker and Language Recognition Workshop.* (2012)
12 *LibriSpeech*, [Online.] available: http://www.openslr.org/12/
13 *VoxForge*, [Online.] available: http://www.voxforge.org/
14 International Civil Aviation Organization, *Annex 10 to the Convention on International Civil Aviation, Aeronautical Telecommunications*, **Vol. II**, 6th Edition, July 2001.
15 Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P. Front-end factor analysis for speaker verification, *IEEE Trans. Audio Speech Lang. Process*. **19** (4) (2011) 788–798.