Audio Engineering Society

Convention Paper 8615

Presented at the 132nd Convention
2012 April 26–29 Budapest, Hungary

Emergency Voice/Stress-level Combined Recognition for Intelligent House Applications

Konstantinos Drossos[1], Andreas Floros[1], Kyriakos Agavanakis[2], Nicolas - Alexander Tatlas[3], and Nikolaos - Grigorios Kanellopoulos[1]

[1] *Ionian University, Department of Audiovisual Arts*

[2] *BlueDev Ltd.*

[3] *Technological Educational Institute Of Piraeus, Department of Electronics*

Correspondence should be addressed to Andreas Floros (floros@ionio.gr)

ABSTRACT

Legacy technologies for word recognition can benefit from emerging affective voice retrieval, potentially leading to intelligent applications for smart houses enhanced with new features. In this work we introduce the implementation of a system, capable to react to common spoken words, taking into account the estimated vocal stress level, thus allowing the realization of a prioritized, affective aural interaction path. Upon the successful word recognition and the corresponding stress level estimation, the system triggers particular affective-prioritized actions, defined within the application scope of an intelligent home environment. Application results show that the established affective interaction path significantly improves the ambient intelligence provided by an affective vocal sensor that can be easily integrated with any sensor-based home monitoring system.

1. INTRODUCTION

Current intelligent house technologies can offer a variety of features for home automation, such as automatic monitoring of bedrooms [1], monitoring of human comfort with wireless sensors [2], control and monitoring of energy consumption and ambience intelligence [3], as well as speech-based interactive control [4] to name a few. In all the above cases, a direct communication channel between the user and the corresponding system interface exists, usually based on visual (i.e. gestural) or aural (i.e. speech) information received, processed and recognized. A

typical example is demonstrated in [5], where both audio and visual channels are employed for realizing the necessary human-machine interaction path.

Focusing on the audio channel as a means for physical interaction, speech is suitable for immediate, time-critical interaction. Legacy techniques for keyword recognition, allow the implementation of voice-driven functionalities for home automation, for example for monitoring and detecting emergency conditions [6]. In the literature, a wide range of approaches exist dealing with the problem of voice-command recognition [7]. Focusing on techniques for restricted sets of commands and for any speaker, which is the most typical case met in interactive home automation environments, efficient performance in terms of correct word recognition can be achieved, frequently at the expense of high computational loads and complexity.

A significant aspect that may enhance the supported user - system interaction is the ability to detect the users emotional state. This information can be combined with the audiovisual channel interaction, aiming to ensure the correctness of the decisions that will be further taken. Currently, there are numerous researches regarding affective computing, user emotion modeling and retrieval and, particularly, emotion recognition from audio (music and speech) data. Focusing exclusively on speech emotion recognition, there are recent researches investigating various emotions modeling cases, like fear, stress, anger etc [8, 9]. On the other hand, the major limitation of existing speech-emotion retrieval techniques is the absence of unified models for defining the exact relations between the audio signal characteristics and the raised affective condition.

In this paper we present the implementation of a a system for recognizing specific spoken words, enhanced with the ability to categorize the recognized words based on the speaker stress level. Although the exclusive focus on stress as the only recognizable emotion can be considered restricted, it provides a robust and efficient implementation framework in terms of affective modeling and emotion retrieval algorithmic complexity, which is fully adapted to the particular requirements imposed by the targeted home-automation applications. Hence, the proposed approach allows the consideration of every-day life spoken words, which are mapped to specific monitoring decisions based on the respective stress-level. Moreover, the proposed affective/priority mapping mechanism allows for fine-tuning of the monitoring mechanism and the significant reduction of false alarm cases that can be raised due to the employment of commonly used verbal commands. It should be also noted that the basic design aim of this work was the ability to integrate the system implementation within any (digital) microphone capsule, targeting to the production of an integrated, intelligent affective vocal sensor that can be combined with alternative sensor mechanisms existing in an in-house monitoring environment.

The rest of the paper is organized as follows: In Section 2, a brief overview of existing works on both voice keyword recognition and voice emotion recognition is presented. Section 3 analyses the architecture of the proposed system and provides a further description of the major implementation issues. Next, Section 4 includes the description of a sequence of tests performed in order to assess the performance of the system and provides a summary of the results obtained. Finally, Section 5 concludes the work, defining specific issues that can be considered in the future.

2. SPEECH RECOGNITION AND VOICE EMOTION RETRIEVAL

In this section, we attempt to present the state-of-the-art in the areas of speech/word recognition and voice emotion assessment, aiming to establish the fundamental concepts of both topics for presenting the proposed system implementation details.

2.1. Voice activity detection

Voice activity detection (VAD) is a task that takes place whenever it is required to trim the input signal to those time regions containing voice activity. It is usually used as a pre-processing module in complete speech recognition systems and it can be realized as a real-time or non-real time task. For real-time processing, existing implementations delay the input buffers of the recording device and compare their content to the data contained in the consecutive ones. There are also many implementations of VAD systems which make use of more sophisticated algorithms, regarding the decision for voiced and unvoiced regions, like [10] that is based on wavelets and

[11], which uses wavelets and support vector machine.

The aforementioned algorithms achieve significant performance under the expense of introducing heavy computational load (especially for real-time applications supported by embedded hardware with small processing power). However, in [12], a robust and relatively simple algorithm for VAD is described. This algorithm makes use of short term features of the input signal and achieves a detection accuracy in the area of 97%. The short term features used are the instantaneous energy, the spectral flatness and the dominant frequency of the input signal. Each of these features is calculated in successive frames and upon the comparison with specific feature-dependent thresholds, a decision whether or not the input frame indicates voice activity is made.

2.2. Voice keywords recognition

In general, legacy voice recognition techniques for independent speakers extract technical features from the voice signal and compare them to the corresponding features obtained from a specific data corpus, which is used as the ground truth data [13]. Standard methods for isolated voice keywords recognition include the extraction of the frequency components of the voice signal, processing of the feature components and comparison of the processing outcome against the corresponding features of the data set used for validation [14].

Among the most commonly used features are the Mel Frequency Cepstrum Coefficients (MFCC) [13, 14, 15], whereas the Dynamic Programming (DP) algorithm used for comparison is the Dynamic Time Warping (DTW) and the k-th Nearest Neighbours (k-NN) method is employed for categorization. The incorporation of MFCC allows direct frequency mapping to the Mel frequency space, while the usage of DTW allows the comparison of signals with different time lengths.

The aforementioned recognition strategies are well-known for their robustness and, although many alternative methods for feature extraction have been proposed, like the Perceptual Linear Prediction, no significant efficiency variations are reported [16]. In addition, regarding the pattern matching process, Hidden Markov Models (HMM) are also widely used. Nevertheless, even if weaknesses for DTW are mentioned regarding speakers' dependencies [17], DTW is considered to achieve better efficiency with smaller data sets against HMM [18].

2.3. Voice emotion recognition

Emotion recognition from voice channel is an emerging field of research with applications varying from plain emotion recognition to human-to-machine speech communication [26]. Moreover, speech emotion recognition has been used in applications for call centers, regarding the management of incoming calls according to the emotional state of the caller [27].

In general, speech/voice emotion recognition can be considered as a pattern matching problem [26]. In that sense, the process concerns the extraction of voice-signal characteristics and their comparison to specific thresholds. More specifically, a number of acoustic cue values is extracted from the voice channel and is further compared against ground truth data using categorization algorithms. In order to produce the above ground truth data, an affective model must be considered. Many relative works have pointed out that emotions are short and intense reactions to an external stimuli originating from the subject's environment [19, 20]. However, a relatively high variety of models for emotion recognition exist. A general categorization of these models includes a) discrete and b) dimensional models [21]. In the existing literature an ongoing debate is observed about which of the two is most appropriate to use [22, 23]. Nevertheless, in these two categories lay the most frequently used models in the field of audio emotion recognition, i.e.: a) basic emotions, b) list of adjectives, and c) valence - arousal or valence - dominance - arousal space [24, 25], with the later being a dimensional model.

The main acoustic cue usually considered is the fundamental frequency (pitch) of voice [26, 28]. Additional acoustic cues may include speech rate (tempo) [29, 30], the instantaneous voice energy [30], as well as the combined variability of the above cues [29, 30, 31]. Different ranges of the aforementioned acoustic cues values are related to different emotions [29, 30]. The above relations between the acoustic cue values and the recognized emotions are termed as emotions' profiles. For example, the acoustic profiles for anger, fear, sadness and happiness, compared to normal speech, are shown in Table 1, following the emotions' profiles proposed in [26, 29, 30, 32].

Finally, the above acoustic cue values serve as input to the categorization algorithm employed for deriving the desired affective classification, taking into account the emotional profile considered. For example, if an increment in the fundamental frequency is observed and both the speech rate and signal energy are decreased, then the identified emotion is happiness.

Emotion	Acoustic Cue		
	F_0	S_R	E
Anger	Increase	Increase	Increase
Happiness	Increase	Decrease	Increase
Sadness	Decrease	Decrease	Decrease
Fear	Increase	Increase	Increase

Table 1: Emotion profiles for anger, fear, sadness and happiness compared to normal speech. F_0 is the fundamental frequency, S_R is the speech rate and E is the energy of the voice signal

3. THE PROPOSED SYSTEM

The proposed system recognizes specific spoken words, belonging into a pre-defined verbal set, enhanced with the ability to categorize them based on the speaker stress level. This categorization is equivalent to an action prioritization scheme based on the assessed stress level. The employed command sets are defined using every-day life spoken words, which are however used in emergency situations, e.g. "Fire".

The general architecture of realized system includes three subsystems, namely the Voice Activity Detector (VAD), the Voice Keyword Recognizer (VKR) and the Voice Stress Classifier (VSC). Figure 1 outlines the overall system architecture, which is further analyzed in the next Sections.

3.1. Voice Activity Detector Subsystem

The VAD algorithm used in this work was originally introduced in [12]. The input voice signal is divided in frames of length equal to N=160 samples, denoted here as S_i. For each frame, the total energy (E), the spectral flatness (Sf) and the dominant frequency (F_D) are calculated. E is calculated as:

$$E = \frac{1}{N}\sum_{i=1}^{N} {S_i}^2 \qquad (1)$$

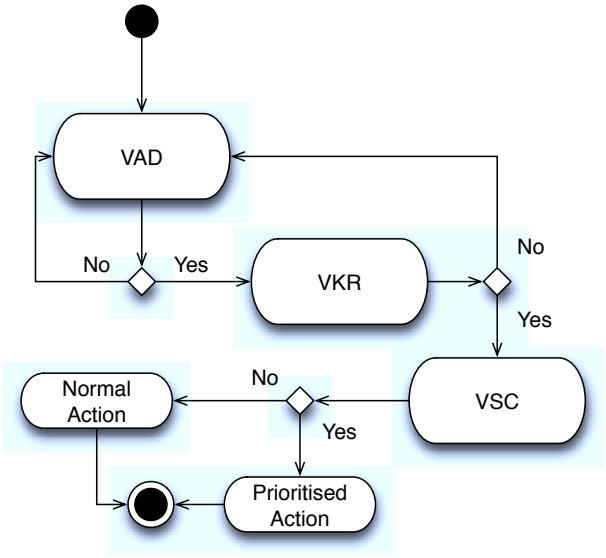

Fig. 1: The emergency voice/stress-level combined recognition system architecture

Accordingly, Sf is calculated as

$$Sf = 10\log_{10}\left(G_m \,/\, A_m\right) \qquad (2)$$

where A_m and G_m are the arithmetic and geometric means of the signal's spectrum, respectively. Finally, F_D is calculated using the equation:

$$F_D = \max\left(|S_f|\right) \qquad (3)$$

where $|S_f|$ denotes the signal's spectrum magnitude.

For a complete algorithm analysis, the reader is encouraged to take a look in [12]. In brief, during consecutive signal's blocks processing, if at least two of the E, Sf and F_D measures are found to exceed the corresponding defined thresholds, then the segment is considered voice-active. The algorithm ignores less than 10 successive segments marked as voice inactive and less than 5 successive segments marked as voice-active. The VAD subsystem architectural layout is illustrated in Figure 2. The primary threshold values used in the proposed system are ilustrated in Table 2.

The above VAD algorithm was chosen in respect with the induced computational load, since for the

Energy value	Frequency bin	Sf value
10^{-4}	2	5

Table 2: Primary threshold values used in the VAD subsystem

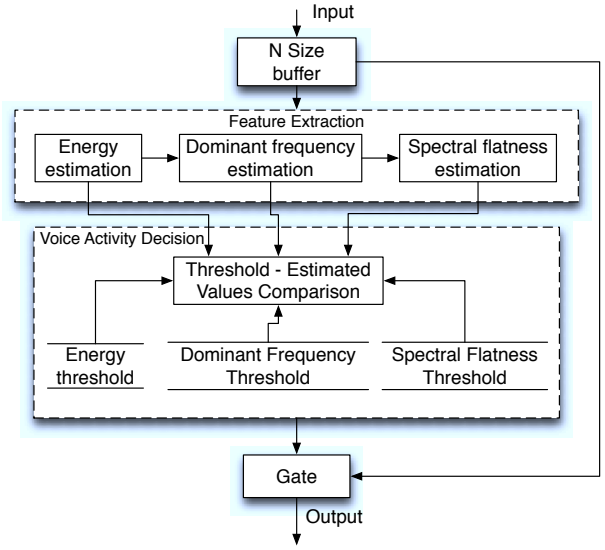

Fig. 2: VAD subsystem architecture

purposes of the current work, it was realized on embedded hardware. Towards this aim, all calculations were carried out using fixed point arithmetic, including the Fast Fourier Transformation, for which the Ooura's library [33] was used.

3.2. The voice Keyword Recognizer

The VKR subsystem makes use of the Mel Frequency (MF) scale and the Discrete Cosine Transformation (DCT) for signal feature extraction. Moreover, DTW and k-NN are employed for pattern matching and final decision making. The overall layout of the VKR subsystem is presented in Figure 3.

The Hamming windowing function used is applied on input signal's segments equal to $M = 512$ samples. Next, the Fast Fourier Transform (FFT) for each segment is calculated, followed by an MF mapping and a DCT on the obtained MF's. The result of this procedure is an $L \times M$ matrix, where L equals to the total number of the recorded signal segments within the specific voice activity region. This matrix is compared to the training data set using the DTW algorithm, thus producing comparison scores. The most dominant score is determined by the k-NN algorithm, indicating the recognized keyword.

The VKR subsystem was trained using a small data set consisting of every-day life words indicating emergency conditions (see below for the selected complete set of these words). Hence, DTW and k-NN were chosen as DP and categorization algorithms, since it is known that they provide better results for such data sets [18]. Moreover, MFCC were chosen against LPC, due to its much simpler implementation [16]. During training, the results provided were stored in the system internal storage (flash memory). The small size of the available memory in embedded systems represents a major limitation for the selection of the training data set volume. For this reason, in order to allow the increment of the recognized words number, only normal (unstressed) words were used as training data.

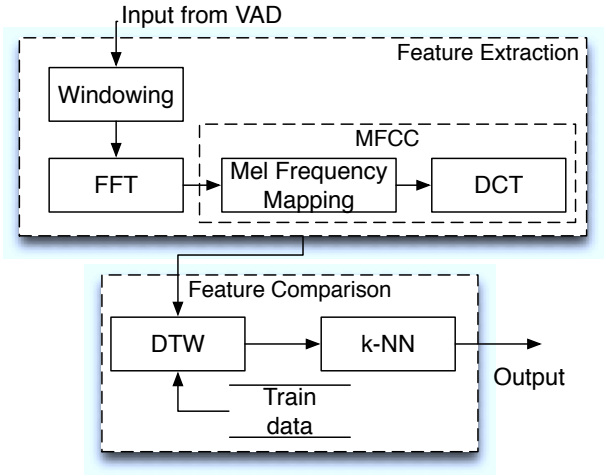

Fig. 3: VKR subsystem architecture

3.3. The voice stress classifier

The voice stress level estimation is based on the Arousal - Valence affective model that was originally proposed in [34]. Based on this model, fear and anger can be expressed in terms of low valence and high arousal values and they are mapped into the same quadrant of the Arousal - Valence space (see the corresponding emotional profile in Table 1).

It should be also noted here that these two emotions are considered to belong to the same family of emotions, even when following the alternative discrete emotions model [30]. However, since stress is not directly included in the discrete emotion set, we hereby consider the same acoustic profile for stress level recognition, since stress is a common component of both aforementioned emotions, especially under emergency situations particularly considered in this work.

Using the above valence-arousal stress representation, the computational load required for realizing the voice stress classifier is obviously reduced, compared to sophisticated methods for emotion recognition that typically employ Support Vector Machines (SVM), HMMs and neural networks [31, 35, 36]. Under the proposed approach, we only need to calculate simple signal characteristics and parameters, such as energy, speech rate and fundamental frequency.

The architectural layout of the VSC module is illustrated in Figure 4. It has two inputs. The first one is the raw voice signal derived from the VAD. The second is the recognized word from the VKR subsystem fed to the VSC as an 8-bit unsigned integer index. Regarding the former input, as mentioned previously, three acoustic cues are calculated: a) the instantaneous energy value (E), b) the speech rate variability (S_{r_V}), and c) the fundamental frequency variability (F_{0_V}).

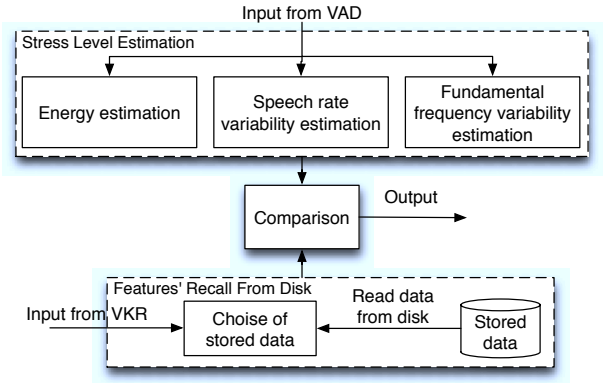

Fig. 4: VSC subsystem architecture

The values of each of the aforementioned acoustic cues are compared to the corresponding thresholds stored in the system internal memory. These thresholds are defined during the training period of the VSC module. Further details on this issue are provided in Section 4.1. If all the measured values are exceeding the above thresholds, then the output of the VSC subsystem simply indicates wether the speaker experiences stress. Hence, the priorities used here are "High" and "Low", corresponding to the presence and non-presence of stress respectively. This indication may serve as input to the action-taking system, being part of the intelligent house monitoring application environment.

4. RESULTS

System training and testing was performed using different training and testing speaker volunteers. They were all equally selected in terms of genre. All participants were students from the department of audiovisual arts, having basic skills as actors and performers. This was a basic requirement, since they had to train and evaluate the system performance under controlled stress and no-stress conditions.

Due to the absence of an emotion annotated speech corpus in Greek, all of the considered words had to be recorded in both stress conditions needed. In order to include all possible usage cases, these words were chosen as following: Nearly half of them indicate emergency situations, while the rest have irrelevant semantic content. The usage of the latter emergency-irrelevant words was chosen in order to decorrelate the semantic content of the words with the acoustic cues' differentiation in different stress conditions. The application vocabulary finally included five spoken greek words, with their direct translation to english being:

1. Thief
2. Earthquake
3. Fine
4. Fire
5. Good morning

4.1. System training

During system's training, ten participants were given instructions regarding the training procedure. They were instructed to enunciate all the

words under stress and normal (no-stress) conditions. Their speech was digitally recorded. Ten different training participants were then employed to cross-validate the emotional state of the recorded words. Only those recordings that obtained confusion scores lower than 20% were finally incorporated in the training data set.

For the words included in the training data set, the mean signal energy, the mean fundamental frequency and the mean speech rate for both stress cases were calculated. These values were used to estimate for each acoustic cue the stress-related threshold per word. More specifically, the energy threshold was defined as the difference of the mean voice signal energy under stress and no-stress conditions. The same approach was employed for defining the speech ratio variation threshold. Finally, the fundamental frequency threshold was set equal to the difference of the mean fundamental frequency minus the 80% ratio of the mean variation of the fundamental frequency in the different emotional states. The results obtained during the above procedure, including the derived threshold values, are summarized in Figures 5, 6, and 7.

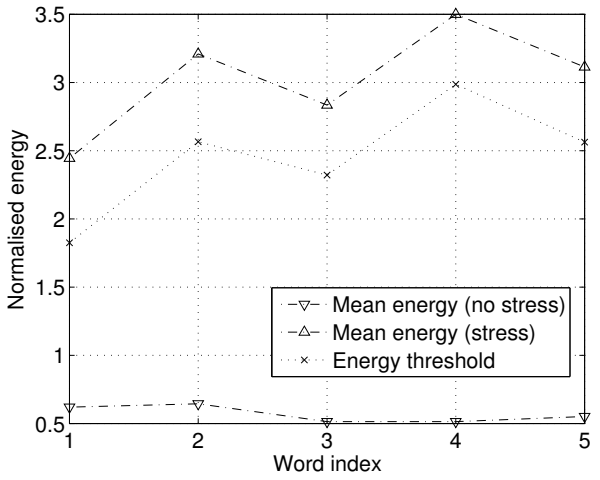

Fig. 5: Mean energy values and thresholds (per word)

With respect to the speech rate acoustic cue, it must to be noted that, although an increase is mentioned in the literature with the anger/fear intensity

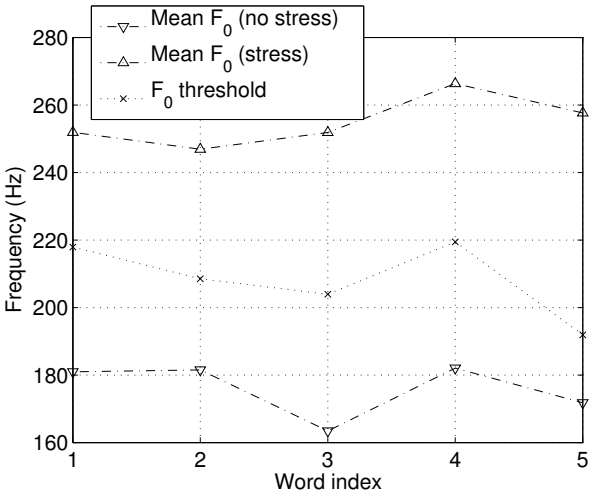

Fig. 6: Mean fundamental frequency values and thresholds (per word)

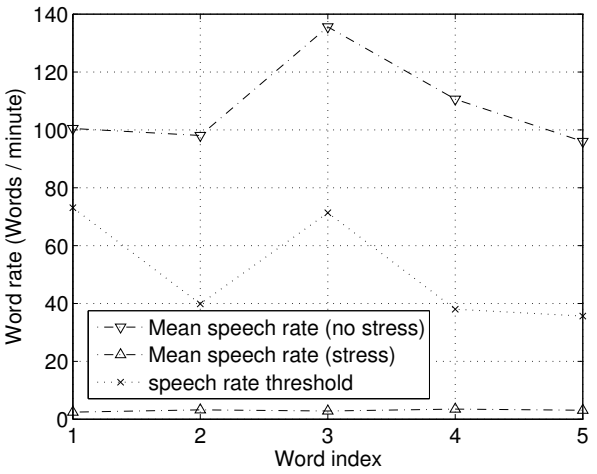

Fig. 7: Mean speech rate values and thresholds (per word)

[29, 30, 31], in Figure 7 a decrease with stress is observed. This trend can be explained taking into account that we consider single words (and not phrases of words), as well as their semantic content. In particular, when a speaker under stress yells for example the single word "thief", it is likely that he/she prolongs the word. Hence, the encountered speech rate

is reduced under stress conditions.

4.2. Subjective performance evaluation

During the system's evaluation period, each participant was asked to perform each of the five words under both stress conditions in a random order. A total of sixteen (16) human subjects participated in this sequence of tests. The obtained confusion matrices for all the words recorded under normal and stress conditions are shown in Tables 3 and 4 respectively.

From Table 3, it is clear that the successful stress recognition rate among all words reaches up to 91%, independently from the stress condition. Moreover, the false alarm cases occurred due to stress miss-recognition is nearly 25%. On the other hand, in the stress case (see Table 4), the overall recognition rate independently from the emotional condition is close to 88%. The non-triggered alarm cases are nearly 25%. Moreover, one can observe nearly absolute correct stress level identification in the case that the words are performed under stress. However, in this case, a small percentage of false word recognition occurs. Table 5 also summarizes the maximum and mean accuracy achieved for all the words - members of the considered corpus and for both stress and non-stress conditions. It must be noted that both VKR and VSC subsystems are speaker independent. Moreover, as mentioned in the previous Section, the training data set employed was obtained from non professional actors and performers. Thus, the authors believe that the above small probability of misjudging the speaker stress condition can be significantly decreased by realizing more accurate data sets.

Recognised as		**No Stress Test Words**				
		1	2	3	4	5
1	Normal	61%	0%	0%	0%	0%
	Stress	26%	0%	6%	0%	0%
2	Normal	0%	72%	0%	0%	0%
	Stress	0%	22%	0%	0%	0%
3	Normal	0%	0%	62%	6%	10%
	Stress	0%	0%	27%	0%	0%
4	Normal	0%	0%	0%	76%	6%
	Stress	0%	0%	0%	18%	6%
5	Normal	0%	0%	6%	0%	63%
	Stress	0%	0%	0%	0%	25%
Miss recognition		13%	6%	0%	0%	0%

Table 3: Confusion matrix for the test words performed without stress

Recognised as		**Stress Test Words**				
		1	2	3	4	5
1	Normal	18%	0%	0%	0%	0%
	Stress	64%	0%	0%	0%	0%
2	Normal	0%	25%	0%	0%	0%
	Stress	0%	63%	0%	0%	0%
3	Normal	0%	6%	18%	0%	0%
	Stress	12%	0%	76%	12%	0%
4	Normal	0%	0%	0%	18%	6%
	Stress	0%	0%	0%	70%	12%
5	Normal	0%	0%	6%	0%	25%
	Stress	0%	0%	0%	0%	57%
Miss recognition		6%	6%	0%	0%	0%

Table 4: Confusion matrix for the test words performed under stress

	No stress words		**Stress words**	
	Max	Mean	Max	Mean
VKR	94%	90%	94%	87%
VSC	76%	67%	76%	66%

Table 5: Mean accuracy results achieved by the VKR and VSC subsystems

5. CONCLUSIONS

New technologies for remote control of many in-house apparatus are emerging, focusing on multimodal interaction paths, typically including gesture/event and voice-signal recognition. Considering the latter case, in this work a system for recognizing specific spoken words is proposed, enhanced with the ability to categorize and prioritized them based on the speaker stress level. This approach allows the consideration of every-day life spoken words, which are mapped to specific monitoring decisions based on the respective stress-level. Moreover, in an in-house monitoring environment, it also offers the ability to prioritize the actions that should be further performed.

The proposed system incorporates typical voice/speech recognition tasks and a stress-

level assessment method. Voice recognition is based on acoustic cue extraction from short-term spectral features (i.e. Mel frequencies cepstral coefficients). Features like Dynamic Time Warping render it signal-length independent, while the k-th Nearest Neighbors algorithm is responsible for the final recognition decision. Voice activity detection is also used for determining instances of silence.

The stress-level detector used employs a combination of criteria such as the short-time energy, the instantaneous speech rate and the variability of the fundamental frequency, which are directly associated to physiological conditions commonly observed under fear and anger. These emotions belong in the same quadrant in the Arousal - Valence space used for modeling human emotions. We hereby consider that stress is a common affective factor for these emotions, hence its detection can be performed based on the same acoustic emotional profiles.

A sequence of tests has shown that the mean accuracy for the word recognition task is nearly 90% for both stress and no-stress cases. Stress estimation accuracy is in the range of 70%, resulting into non-recognized alarms and false alarm percentage in the range of 25%. Due to the high correlation observed between the results and the training data set, it is expected that with word recordings obtained from professional actors, the overall recognition and prioritization accuracy of the system can be improved. This is a task that will be considered in the near future by the authors, together with modifications on the particular signal processing algorithms for minimizing the implementation requirements raised in terms of computational load. The latter fact is fundamental for optimizing the real-time performance of the system in embedded platforms, as a part of an intelligent sonic sensor for in-house monitoring applications.

6. ACKNOWLEDGEMENTS

The research activities that led to these results, were co-financed by Hellenic Funds and by the European Regional Development Fund (ERDF) under the Hellenic National Strategic Reference Framework (ESPA) 2007-2013, according to Contract no. MIKRO2-40/E-II-A.

7. REFERENCES

[1] Y. Huishan, et al., "The Designs of Intelligent Bedroom Network Monitor System," Procedia Engineering, vol. 15, pp. 644-648, 2011.

[2] M. I. M. Rawi and A. Al-Anbuky, "Development of Intelligent Wireless Sensor Networks for Human Comfort Index Measurement," Procedia Computer Science, vol. 5, pp. 232-239, 2011.

[3] D. J. Cook, et al., "Ambient intelligence: Technologies, applications, and opportunities," Pervasive and Mobile Computing, vol. 5, pp. 277-298, 2009.

[4] J.-i. Takahashi, et al., "Interactive voice technology development for telecommunications applications," Speech Communication, vol. 17, pp. 287-301, 1995.

[5] M. Ben Ammar, et al., "The Affective Tutoring System," Expert Systems with Applications, vol. 37, pp. 3013-3023, 2010.

[6] M. Hamill, et al., "Development of an automated speech recognition interface for personal emergency response systems," Journal of NeuroEngineering and Rehabilitation, vol. 6, 8 July 2009.

[7] G. Poli, et al., "Voice Command Recognition with Dynamic Time Warping (DTW) using Graphics Processing Units (GPU) with Compute Unified Device Architecture (CUDA)," presented at the 19th International Symposium on Computer Architecture and High Performance Computing, 2007. SBAC- PAD 2007, Rio Grande do Sul, 2007.

[8] P. N. Juslin and K. R. Scherer, "Vocal expression of affect," in The New Handbook of Methods in Nonverbal Behavior Research, J. A. Harrigan, et al., Eds., ed Oxford, Great Britan: Oxford Unversity Press, 2005, pp. 65-135.

[9] T. Johnstone, et al., "The voice of emotion: an FMRI study of neural responses to angry and happy vocal expressions," Social Cognitive and Affective Neuroscience, vol. 1, pp. 242-249, December 1 2006.

[10] M. Eshaghi and M. R. Karami Mollaei, "Voice activity detection based on using wavelet packet," Digital Signal Processing, vol. 20, pp. 1102-1115, 2010.

[11] S.-H. Chen, et al., "Improved voice activity detection algorithm using wavelet and support vector machine," Computer Speech & Language, vol. 24, pp. 531-543, 2010.

[12] M. H. Moattar and M. M. Homayounpour, "A Simple But Efficient Real-Time Voice Activity Detection Algorithm," presented at the 17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland, August 24-28, 2009

[13] W. Huang, et al., "A neural net approach to speech recognition," presented at the International Conference on Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988, New York, U.S.A., 1988.

[14] T. Nomura and R. Nakatsu, "Speaker-independent isolated word recognition for telephone voice using phoneme-like templates," presented at the IEEE International Conference on ICASSP '86 Acoustics, Speech, and Signal Processing, Tokyo, Japan, 1986.

[15] M. Benzeghiba, et al., "Automatic speech recognition and speech variability: A review," Speech Communication, vol. 49, pp. 763-786, 2007.

[16] I. Mporas, et al., "Comparison of Speech Features on the Speech Recognition Task," Journal of Computer Science, vol. 3, pp. 608-616, Aug 2007.

[17] K. Chanwoo and S. Kwang-deok, "Robust DTW-based recognition algorithm for hand-held consumer devices," IEEE Transactions on Consumer Electronics, vol. 51, pp. 699-709, 2005.

[18] T. F. Furtuna, "Dynamic Programming Algorithms in Speech Recognition," Revista Informatica Economica, vol. 12, pp. 94-99, 2008.

[19] P. N. Juslin and P. Laukka, "Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening," Journal of New Music Research, vol. 33, pp. 217 - 238, September 2004.

[20] P. N. Juslin and D. Vastfjall, "Emotional responses to music: The need to consider underlying mechanisms," Behavioral and Brain Sciences, vol. 31, pp. 559-575, 2008.

[21] S. Kai, et al., "An improved valence-arousal emotion space for video affective content representation and recognition," presented at the IEEE International Conference on Multimedia and Expo, 2009. ICME 2009, Cancun, Mexico, 2009.

[22] P. N. Juslin and P. Laukka, "Communication of emotions in vocal expression and music performance: different channels, same code?," Psychological Bulletin, vol. 129, pp. 770-814, September 2003.

[23] L. Lie, et al., "Automatic mood detection and tracking of music audio signals," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, pp. 5-18, January 2006.

[24] K. R. Scherer, "Which Emotions Can be Induced by Music? What Are the Underlying Mechanisms? And How Can We Measure Them?," Journal of New Music Research, vol. 33, pp. 239 - 251, 2004.

[25] C. Laurier, et al., "Exploring Relationships between Audio Features and Emotion in Music " presented at the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009) Jyvaskyla, Finland 2009

[26] S. Giripunje and N. Bawane, "ANFIS Based Emotions Recognision in Speech," in Knowledge-Based Intelligent Information and Engineering Systems, ed, 2009, pp. 77-84.

[27] A. V. Petrushin, "Emotion in Speech: Recognition and Application to Call Centers," presented at the Conference on Artificial Neural Networks in Engineering (ANNIE 99), St. Louis, Missouri, 1999.

[28] P. Chang-Hyun and S. Kwee-Bo, "Emotion recognition and acoustic analysis from speech

signal," in Neural Networks, 2003. Proceedings of the International Joint Conference on, 2003, pp. 2594-2598 vol.4.

[29] K. R. Scherer, "Vocal Affect Signaling: A Comparative Approach," in Advances in the Study of Behavior. vol. Volume 15, J. S. Rosenblatt, et al., Eds., ed New York, U.S.A.: Academic Press, 1985, pp. 189-244.

[30] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," Journal of Personality and Social Psychology, vol. 70, pp. 614-636, 1996.

[31] T. L. Nwe, et al., "Speech emotion recognition using hidden Markov models," Speech Communication, vol. 41, pp. 603-623, November 2003.

[32] J. Rong, et al., "Acoustic feature selection for automatic emotion recognition from speech," Information Processing & Management, vol. 45, pp. 315-328, 2009.

[33] T. Ooura. (2006, 11 Nov.). FFT Package 1-dim / 2-dim. Available: http://www.kurims.kyoto-u.ac.jp/ ooura/fft.html

[34] J. A. Russell, "A circumplex model of affect," Journal of Personality & Social Psychology, vol. 39, pp. 1161-1178, 1980

[35] B. Schuller, et al., "Acoustic emotion recognition: A benchmark comparison of performances," in Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on, 2009, pp. 552-557.

[36] W.-J. Yoon and K.-S. Park, "A Study of Emotion Recognition and Its Applications," presented at the Proceedings of the 4th international conference on Modeling Decisions for Artificial Intelligence, Kitakyushu, Japan, 2007.