Variability of voice fundamental frequency in speech under stress.

Grażyna Demenko [a] , *Magdalena Oleśkowicz-Popiel* [a],
Krzysztof Izdebski [b,c], *Yuling Yan* [c]

[a] Department of Phonetics, A. Mickiewicz University of Poznan, Poznan, (Poland)
[b, c] Pacific Voice and Speech Foundation, San Francisco, CA (USA)
[c] Santa Clara University, Santa Clara (USA)
lin@amu.edu.pl, mmj@amu.edu.pl
kizdebski@pvsf.org, yyan1@scu.edu

Abstract

By analyzing acoustic and phonetic structure of live recordings of 45 speakers from police 997 emergency call center in Poland, we demonstrated how stressful events are coded in the human voice. Statistical measurements of stressed and neutral speech samples showed relevance of the arousal dimension in stress processing. The MDVP analysis confirmed statistical significance of following parameters: fundamental frequency variation, noise-to-harmonic-ratio, sub-harmonics presence and voice quality irregularities. In highly stressful conditions a systematic over-one-octave shift in pitch was observed. Linear Discriminant Analysis based on nine acoustic features showed that it is possible to categorize speech samples into one of the following classes: male stressed or neutral, or female stressed or neutral.

Index Terms: call centers interfaces, detection of vocal stress, stress visualization, physiological correlates

1. Introduction

Recognition of whether a speaker is under stress is of crucial value in many civilian and military applications, hence automatic detections of vocal stress is becoming increasingly important. Applications of this technology can be found in multilingual communication, in security systems, in banking, in homeland security and in law enforcement [1, 2, 3, 4]. Automatic detection of voice under stress is specifically crucial in emergency call centers and in police departments, as all over the world these units of public safety are overloaded with different kinds of calls, only some of which represent a real danger and a need of an immediate response. Hence, to improve decision making process, response effectiveness, and to save lives, it is of pragmatic interest to detect automatically those speech signals that contain vocally mediated stress [2, 3, 4].

Several investigations [5, 6] showed direct evidence of emotion recognition to stress verification [5, 7, 8] by highlighting differences in acoustical features between the neutral and stressed speech signals brought by a variety of emotions [2, 9]. A number of these studies focused on the effects of emotions on stress because of a close relation between emotions and stress recognition, e.g. usage of similar acoustic features (F_0, intensity, speech unit duration) and arousal dimension [10, 11, 12]. These studies demonstrate that emotional speech correlates are dependent on physiological constraints and do correspond to broad classes of basic emotions, but disagree on the specific differences between the acoustic correlates of particular classes of emotions [11, 13].

Certain emotional states can be correlated with physiological states, which in turn have predictable effects on speech and on its prosodic features. For instance, when a person is in a state of anger, fear or joy, the sympathetic nervous system is aroused and speech becomes louder, faster and enunciated with stronger high-frequency energy. When one is bored or sad, the parasympathetic nervous system is also aroused, which results in a slow, low-pitched speech with little high-frequency energy [10]. Apart from these differences, other studies showed an increase in intensity and in fundamental frequency, a stronger concentration of energy above 500 Hz and an increase in speech rate in cases of stressed speech [10].

A number of studies have considered analysis of speech under both simulated and actual stress condition, though the interpretation of speech characteristics is not unambiguous [10]. Research frequently reports on conflicting results, due to differences in experimental design, categorization of actual or simulated stress, and/or interpretation of results [10]. Studies using actors, simulated stress or emotions have the advantage of a controlled environment, but their major disadvantage is, however, artificial nature of these signals that can result in producing highly exaggerated misrepresentations of emotions in speech [10].

Few studies focused on analysis of authentic recordings coming from actual stressful situations [10]. There is usually no doubt as to the presence of stress in these situations; however there is a problem with categorization of the homogeneous classes of vocally embedded stress. Our study therefore focuses on the analysis of voice stress produced in response to real live situations, eliminating variables present in simulated stress studies. The research aims at extraction of those acoustic features which produce stressed in vocalization in a relatively homogenous group of actual threat stressors.

While some progress has been made in the area of stress definition and assessment from the acoustic or visual signals [10, 7, 14], visual correlates of vocal folds or supraglottic larynx contribution of these affected signals are essentially non-existing [10, 15], hence, we analyzed only the third order of stressor --the psychological ones-- which have their effect at the highest level of speech production [2]. External stimuli such as a threat are subject to individual cognitive evaluations and the emotional states they may bring about (i.e. fear, anger, irritation) to affect speech production at its highest levels. We hypothesized that models using live speech samples from both, stressful and neutral environment, will provide better determinants of acoustic stress indicators, and will help answering questions which of the prosodic derivatives are most valuable vocal stress indicators and which can be used in automatic stress detection.

2. Speech corpus construction and annotation

The 997 - Emergency Calls Database is a collection of spontaneous speech recordings that comprises crime/offence notifications and police intervention requests. All recordings are automatically grouped into sessions according to the phone number from which the call was made. In all over 8 000 sessions were available.

From this corpus, a six-levels preliminary manual phonetic annotation was performed: (1) background acoustics, (2) types of dialog, (3) suprasegmental features such as: (3.1) speech rate (fast, slow, rising, decreasing), (3.2) loudness (low voice or whisper, loud voice, decreasing or increasing voice loudness), (3.3) intonation (rising, falling or sudden break of melody and unusually flat intonation), (4) context (threat, complaint and/or depression), (5) time (passed, immediate and potential), (6) emotional coloring (up to three categorical labels and values for three dimensions: potency, valency, arousal; where potency is the level of control that a person has over the situation causing the emotion, valency states whether the emotion is positive or negative and arousal refers to the level of intensity of an emotion [10, 8, 16].

The annotation allowed for choosing a fairly uniform group of 45 speakers, both males and females, and voice stress detection was performed only on those speakers who manifested different arousal level in two or more dialogs.

3. Pitch characteristics of stress

3.1. Pitch register

A key issue of stress detection by machine is defining utterance segmentation that would result in clear units with respect to perceptual and acoustic homogeneity. Vocal registering, which divides voice region ranges into registers, is an important perceptual category. There are many approaches to define vocal register. For example vocal registration is perceptually a distinct region of vocal quality that can be maintained over some ranges of pitch and loudness over consecutive voice frequencies without a break [17]. However, vocal register definition and register classification terminology is one of the most controversial problems and that physiological register correlates are not defined [15, 18]. Therefore, as a solution to that problem, three cases have been presupposed: (1) different pitch position, same pitch range, (2) different pitch position different pitch range, (3) same pitch position, different pitch range, were pitch range is the difference between F_{max} and $F_{min.}$ (values for F_{max} and Fmin averaged in the region were maximum and minimum was detected, in order to avoid pitch detection errors).

3.2. Pitch ranges

3.2.1. Different pitch position same pitch range

Based on statistical analysis three pitch position settings were observed in the studies: (1) relative constant pitch position within the utterance and dynamic pitch position changes within the utterance: (2) pitch position shifted upward, (3) pitch position shifting up and down. These are shown in the following Figures 1-5

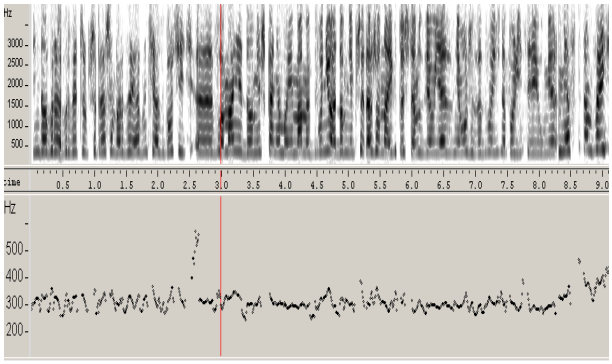

Figure 1a: *F_0 contour of constant stress in the utterance: "Please, come over, there's a house-breaking. She's scared to death" (F_{min} =240 Hz, F_{max}=352 Hz).*

1) Relative constant pitch position within the phrase.
Figure 1a is an acoustic representation of an utterance informing about a burglary and a life threat, whereas Figure 1b illustrates an utterance from the same person calling off the intervention (informing that the burglar has left the apartment), recorded one hour after the first call.

The follow-up call shows a downward F_0 shift in pitch position by approximately 40 Hz (Figure 1b), as compared to F_0 contour in utterance from Figure 1a.

The utterances in Figure 1a and 1b have similar pitch ranges but different pitch positions (we assume these were caused by stress).

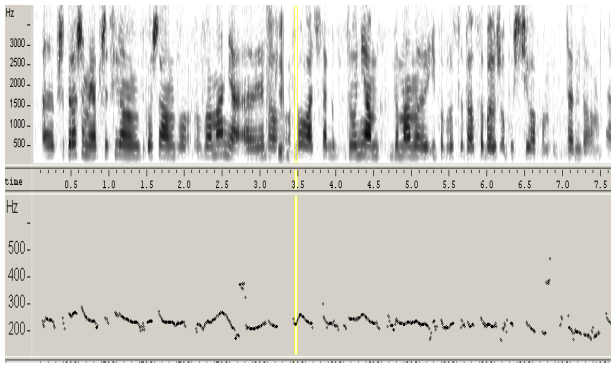

Figure 1b: *F_0 contour of neutral speech in the utterance: "I called one hour ago, I want to call off the intervention" (F_{min}=167 Hz, F_{max}=264 Hz).*

2) Dynamic change of pitch position within the utterance.
Pitch position shifted upward.

In cases of high stress levels, F_0 can reach extreme values. For example female voices may be elevated up to 700 Hz. Figure 2a illustrates an utterance of a female speaking with extreme stress as she reports to the police, "*a masked person has entered my apartment*". Vocal stress decreases only slightly at the end of the recording, after hearing a dispatcher prompt asking her to calm down. As the stress of the speaker increases the following is noted: 1) an upward shift in the voice pitch, 2) as well as a prominence of the higher

frequencies in the spectrum, 3) an increase in the signal's energy and 4) rate changes.

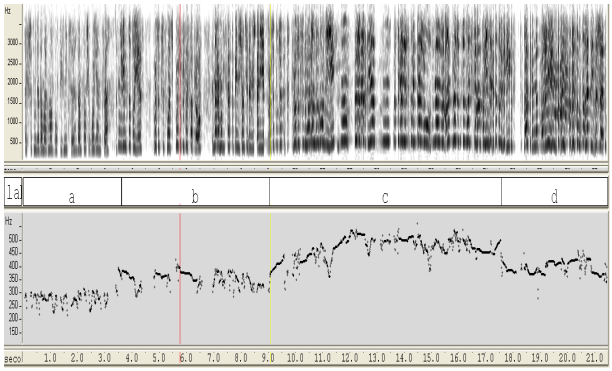

Figure 2a*: A gradual stress increase in the utterances: a) "Someone is entering the apartment" (F_{min} =220Hz), b) "He's masked" (F_{min} =260 Hz), c) "he is somewhere [here]"-- direct threat (F_{min} =320 Hz), d) "Please come to Kwiatowa Street"-- the answer after being asked by a police officer to calm down and tell him the address (F_{min} =280 Hz).*

In cases of high levels of stress F_0 values can rich extreme values (even up to 750 Hz). Figure 2b illustrates an utterance marked by extreme stress increase that ended with a scream and an exceeding lengthening of some syllables. In this case F_0 changes are located in the range of 220 Hz - 750 Hz. As the stress of the speaker increases the following is observed: 1) an upward shift in the voice pitch, 2) as well as a prominence of the higher frequencies in the spectrum, 3) an increase in the signal's energy and 4) rate changes.

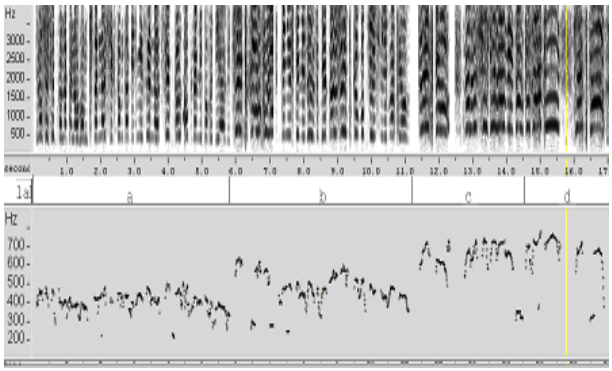

Figure 2b*: A gradual increase in stress in the utterances: (a) "Please, [come] quickly to Kanałowa [street] 18, they want to kill my son, they've broken the window" (F_{min} =289 Hz), (b) "M E (name of the caller withheld), quickly, the mobsters have come" (F_{min} =345 Hz), (c) "Quickly. It's happening, they want to kill him" (F_{min}=495 Hz), (d) "Quickly, S… is killing him (scream)" (F_{min} =495 Hz, F_{max}=748 Hz).*

3) Dynamic change of pitch position within the utterance. Pitch register shifted upward and downward.

The shaded part in the Figure 3 shows an utterance by male voice characterized by a significant, over 50Hz, upward shift of F_0 position.

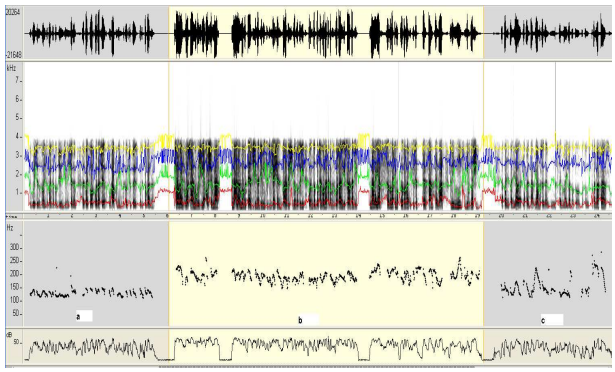

Figure 3*: a) "I keep trying to get through…" (F_{min} =121Hz), b) "I've reported it so many times already…"-- clearly audible irritation (F_{min} =173 Hz) c) "… so I don't know anything anymore…"-- the answer after being asked by a police officer to calm down (F_{min} =115 Hz).*

3.2.2. *Different pitch position different pitch range*

In cases of anger and mixed emotions significant changes of both pitch position and pitch range were observed. Figure 4 illustrates F_0 contour for an utterance in a female voice, where first part (the end of which has been marked by the cursor) has been classified as the voice of indignation. The speaker can easily control her emotional state so that her message is clearly perceived by the listener. Each syllable that is lexically permissible is clearly stressed.

By comparison, in the final part of the recording (beginning of which has been marked by the cursor), as a result of the discourse, the female speaker softens and calms down her manner of speaking, so the recording has a different F_{min} and pitch range width than its first part.

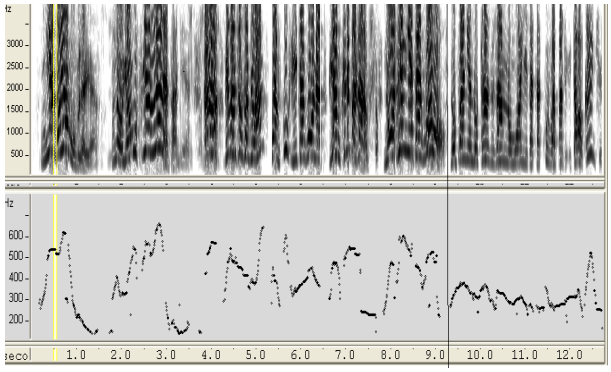

Figure 4*: F_0 contour for an expressive utterance (indignation): "I've got here such a drunkard, he's maltreating me, I am going to trash him…" (F_{max}=675Hz, F_{min} =139 Hz, first part of the utterance), "But what can I do…" (F_{max}=275Hz, F_{min} =206 Hz, second part of the utterance).*

3.2.3. *Same pitch position different pitch range*

Figure 5a and 5b illustrate utterances of the same male speaker, in neutral state and in anger respectively. Both utterances have similar F_{min}, their ranges of F_0 fluctuations however differ significantly.

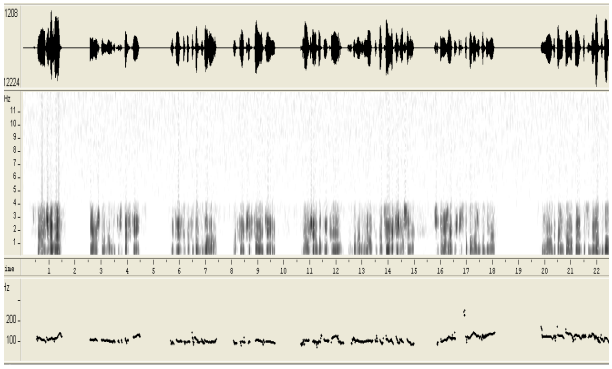

Figure 5a*: F_0 contour for a neutral utterance: "Hi, I live on XXX street…" (F_{max}=137Hz, F_{min} =92Hz).*

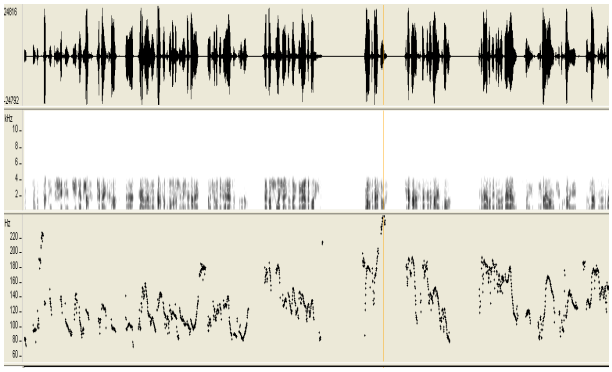

Figure 5b*: F_0 contour for an expressive utterance of indignation: "I hear some shouting and name-calling… him…" (F_{max}=252Hz, F_{min} =86 Hz).*

4. Stress classification

The material was divided into four groups: G1: male – stress, G2: male – neutral/mild irritation, G3: female – stress, G4: female – neutral. Although acoustic analysis of MDVP allows for 32 features [19], only 9 have been used and correlated to LDA Linear Discriminant Analysis. The features used were: Average (F_0), Highest (F_{hi}) and Lowest Fundamental Frequency (F_{lo}), Fundamental frequency variation (vF_0 /%)/, Jitter (Jitt), Amplitude perturbation Quotient (sAPQ)/%/, Degree of Subharmonic Segments (DSH) /%/, Noise to Harmonic Ratio (NHR), Degree of voiceless DUV (%).

The LDA analysis of nine parameters enabled the classification of four groups with the average 80% accuracy, for two groups (neutral and stressed speech, males and female together) the accuracy was a bit higher, 84%. The results showed that extreme stress can be clearly identified by using only the amplitude information with mean and minimum F_0 values.

Figure 6 shows z-normalized F_{min} (F_{lo}) values for four groups: G1, G2, G3, G4. Highest pitch position (F_{min}) values are demonstrated by groups G1 and G3 (speech under stress), whereas F_{min} values for groups G2 and G4 are statistically substantially lower.

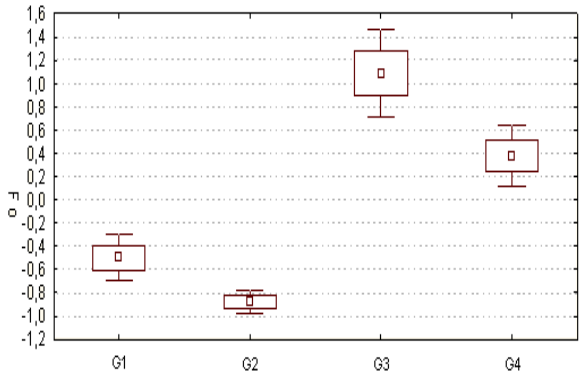

Figure 6*: Z-normalized values F_{min} for G1, G2, G3, G4.*

Table 1 shows classification results. Utterances by male voices affected by stress (G1) obtained better results than those of female voices affected by stress (G3).

	% correct	G_1:1 p=,23	G_2:2 p=,27	G_3:3 p=,23	G_4:4 p=,26
G_1:1	80,00	20	3	2	0
G_2:2	86,20	3	25	0	1
G_3:3	76,00	1	0	19	5
G_4:4	78,57	0	4	2	22
Total	80	21	35	21	30

Table 1*: Classification matrix: rows – classification observed, columns – classification expected.*

5. Stress visualization

An approach to characterize vocal folds (VF) vibrations from HSDI recordings using Nyquist plot was pioneered an detailed in Yan et al., [20, 21], while automatic and robust procedures to generate the glottal area waveform (GAW) from HSDI images were also provided by Yan et al. [22].

The principles underlying this approach are summarized below and illustrated in Figure 7. The HSDI-derived GAW is normalized for all of our analyses to a range of 0~1 with 0 corresponding to complete closure and 1 corresponding to maximum opening. This operation allows for standardized dynamic measurements of VF vibration. The Nyquist plot and associated analyses are used to represent the instantaneous property of the VF vibration, rather than a time averaged one. This property is revealed by the amplitude and phase of the complex analytic signal (e.g. in the form of Nyquist plot) that we generate from the Hilbert transform of the GAW as illustrated in Figure 7 (A, B, C). This operation is applied to as many as 200 glottal cycles taken from 4000-frames of a 2-second HSDI recording (i.e. at a 2000 Hz acquisition rate). Nyqist plots can be also derived from the acoustic signals. Here we submitted to Nyquist analysis the acoustic signals from neutral and stressed segments derived from our samples and to depict the differences in voice stress levels form the same speakers.

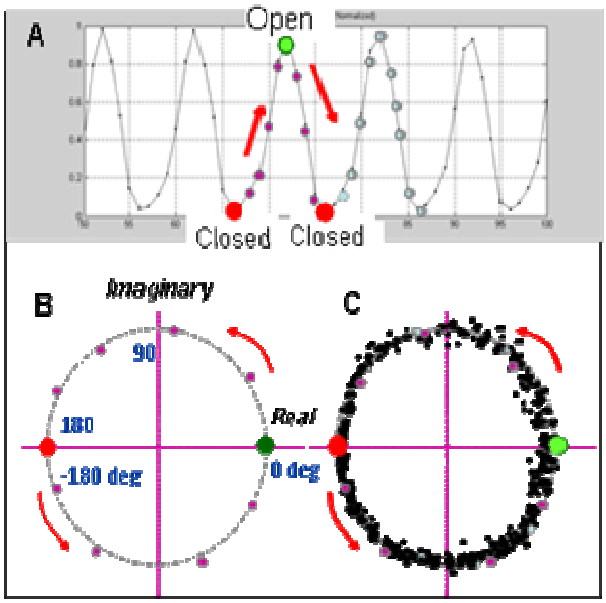

Figure 7*: Concept of the Nyquist plot approach to characterize vocal fold vibrations.*
***A)** a normalized GAW, representing 50 sequential frames (5 vibratory cycles) from a 2000 f/s HSDI recording; the open (0^o) and closed (90^o) glottal cycles are determined from automatic tracing of HSDI images (Yan et al, 2006b).*
***B)** One vibratory cycle is mapped onto the complex plane, where the magnitude-phase of the analytic signal is graphed; the complex analytic signal is constructed from the Hilbert transform of the GAW.*
***C)** Overlays of subsequent vibratory cycles generate a Nyquist plot - deviation of the points from the circle (scatter and shape distortion) reflects the effects of shimmer, jitter and nonlinearity.*

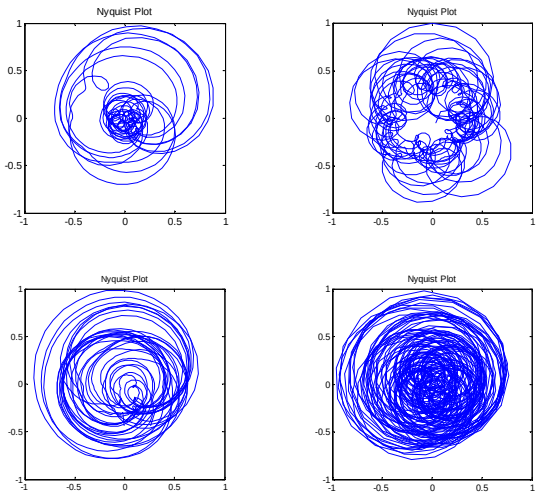

Figure 8: *Nyquist plots for vowel "a" (Fig.8a) and "i" (Fig.8b) in neutral speech (upper plots) and speech under stress (bottom plots)*

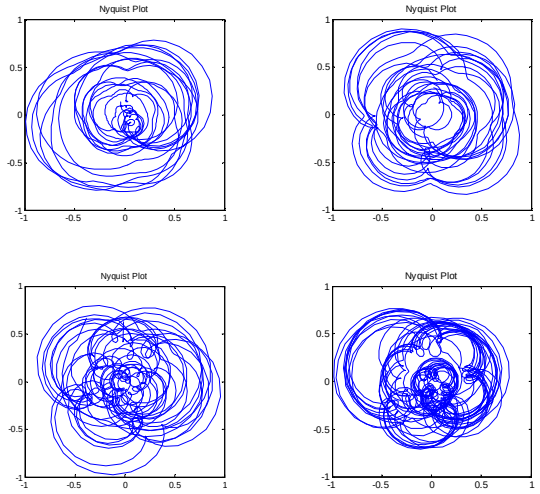

Figure 9*: Nyquist plots for vowel "o" (Fig.9a) and vowel "o" from different phonetic context (Fig.9b), in neutral speech (upper plots) and speech under stress (bottom plots)*

Figures 8a and 8b show Nyquist plots for vowel "a" and "i" in neutral speech (upper plots) and speech under stress (bottom plots).

Figures 9a and 9b show Nyquist plots for vowel "o" from two different contexts both in neutral speech (upper plots) and speech under stress (bottom plots). All vowels were extracted from continuous speech.

The differences in Nyquist plots for vowels in neutral speech (upper plots) and speech under stress (bottom plots) are obvious and very distinctive. Overall, more structured Nyquist patterns (for vowels "a", "i", "o") are observed in speech under stress in comparison to those in neutral speech. Yet, it should be noted that for an objective analysis it is necessary to use a standardized set of speech samples (mainly vowels or sonorants) that will enable evaluation of statistical significance.

6. Conclusion

Despite restricting the study to 45 speakers, a clear tendency in acoustic characterization of speech under stress was observed.

The results of this study confirm the crucial role of the F_0 parameter for investigating stress. Our results agree with literature [2, 5, 10, 23] and point that F_{max} (averaged from several values within the region were maximum was detected, in order to avoid pitch detection errors) must be considered a particularly important parameter in the emotional stress detection. However, this and our previous work [24] showed that a shift in the F_0 contour is also a crucial stress indicator, thus an increase in F_{max} in stressed speech results from a shift in the F_0 register. This holds specifically for vocalizations caused by fear. A systematic increase in the range of F_0 variability for the stress related to anger and to irritation was observed. The results also confirmed the need of including

shift of pitch position and change in pitch register width into prosodic structures segmentation.

We are now preparing to correlate these findings with visual (optical) observations of vocal fold activity using HSDI during production of various emotional vocal components. This will, in our opinion, enable improved explanation of the factors that influence pitch register changes in utterances diversified linguistically and in terms of situational context.

7. Acknowledgements

This project is supported by The Polish Ministry of Sciences and Higher Education (project no O R00 0170 12) and in parts by PVSF funding. We are grateful to Ms. Emma Marriott and Ms. Clara Lewis, for editing of this text.

8. References

[1] Eisenberg, A, "Software that listens for lies," The New York Times, Sunday December 4, 2011.

[2] Hansen, J., et al., "The Impact of Speech Under `Stress' on Military Speech Technology," NATO report. Online: http://www gth.die.upm.es/research/documentation/referencias/Hansen_The Impact.pdf, 2007.

[3] Lefter, J., Rothkrantz, L., Leeuwen, D., Wiggers, P., "Automatic stress detection in emergency (telephone) calls," International Journal of Intelligent Defence Support Systems 4(2), 148-168 (21), 2011.

[4] Vidrascu, L., Devillers, L., "Detection of real-life emotions in call centers," Proc. of Interspeech, 1841–1844, 2005.

[5] Shipp, T., Izdebski, K., "Current evidence for the existence of laryngeal macrotremor and mictrotremor," J. Forensic Sciences, 26, 501-505, 1981.

[6] Cowie, R., Cornelius, R.R., "Describing the emotional states that are expressed in speech," Speech Communication, 40, 5-32, 2003.

[7] Alter, K., Rank, E., Kotz, S. A., Toepel, U., Besson, M., Schirmer, A., Friederici, A. D., "Affective encoding in the speech signal and in event-related brain potential," Speech Communication, 40 (1–2), 61–70, 2003.

[8] Oudeyer, P.-Y., "The production and recognition of emotions in speech: features and algorithms," Int. J. of Human-Computer Studies 59 (1–2), 157–183 (2003).

[9] Huber, R., Batliner, A., Buckow, J., Noth, E., Warnke, V., Niemann H., "Recognition of emotion in a realistic dialogue scenario," Proc. of the Int. Conf. on Spoken Language Processing Beijing, China, 665- 668, 2000.

[10] Izdebski, K. (ed.), " Emotions in the Human Voice", [Vol. 1-3], Plural Publishing, San Diego, CA, 2008-2009.

[11] Ekman, P., "An argument for basic emotions," Cognition and Emotion 6, 169-200, 1992.

[12] Scherer, K.R., "What are emotions? And how can they be measured?", Social Science Information 44 (4), 695–729, 2005.

[13] Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E., "Desperately seeking emotions or: Actors, wizards, and human beings," Speech Emotion-2000, 195-200, 2000.

[14] Izdebski, K, Yan Y., "Preliminary observations of vocal fold vibratory cycle with HSDI a function of emotional load," In progress (ePhonscope, 2013).

[15] Izdebski, K., "SFCM 202 Voice Physiology Manual". In press e-Q&A-p, San Francisco, CA, 2013.

[16] Fontaine, R.J., Scherer, K.R., Roesch, E.B., Ellsworth, P.C., "The World of Emotions is not Two-Dimensional," Psychological Science 18 (12), 1050-1057, 2007.

[17] Frič, M., Šram, F., Švec, J.G., "Voice registers, vocal folds vibration patterns and their presentation in videokymography," Proc. of ACOUSTICS High Tatras 06. 33rd International Acoustical Conference - EAA Symposium, Štrbské Pleso, Slovakia, October 4th - 6th, 2006. ISBN 80-228-1672-8, 42-45, 2006.

[18] Shriberg, E., Ladd, D.R., Terken, J., Stolcke, A., "Modeling pitch range variation within and across speakers: predicting F0 targets when 'speaking up'," Proc. Of the Int. Conf. on Spoken Language Processing (Addendum, 1-4), Philadelphia, PA, 1996.

[19] Deliyski, D., "Acoustic model and evaluation of pathological voice production," Proc. Eurospeech'93, 1969-1972, 1993.

[20] Yan Y, Ahmad K, Kunduk M, Bless D, "Analysis of vocal fold vibrations from high-speed laryngeal images using a Hilbert transform based methodology", Journal of Voice 19(2), 161-175, 2005.

[21] Yan Y, Edward Damrose, Diane Bless, "Functional Analysis of Voice Using Simultaneous High-Speed Imaging and Acoustic Recordings", Journal of Voice 21(5), 604-616, 2007.

[22] Yan Y, Chen X, Bless D, "Automatic tracing of the vocal fold motion from high speed digital images", IEEE Trans Biomed Eng 53(7), 1394-1400, 2006.

[23] Protopapas A., Lieberman P., "Fundamental frequency of phonation and perceived emotional stress," J. Acoust. Soc. Am. 101 (4), 2268-2277, 1997.

[24] Demenko, G., "Voice Stress Extraction," Proc. of Speech Prosody Conference. May 6-9, 2008, Campinas, Brasil, 53-56, 2008.