ANALYSIS OF VOICE STRESS IN CALL CENTERS CONVERSATIONS

Grażyna Demenko[1], *Magdalena Jastrzębska*[2]

[1]Poznan Supercomputing and Networking Center, Poznan, Poland
[2]Department of Phonetics, A. Mickiewicz University of Poznan, Poznan, Poland

grazyna.demenko@speechlabs.pl; magdalena.jastrzebska@speechlabs.pl

Abstract

This paper presents how voice stress is manifested in the acoustic and phonetic structure of the speech signal. Out of few thousand authentic Police 997 emergency phone calls, a few hundred were selected automatically according to their duration (dialogs shorter than 3-4 seconds were omitted). Finally, 45 speakers were chosen for acoustic evaluation, the basis for selection being a perceptual assessment.

Basic statistical measurements for stressed and neutral speech run over the database showed the relevance of the arousal dimension in stress processing. The MDVP analysis confirmed statistical significance of such parameters as fundamental frequency and pitch variation, noise-to harmonic-ratio, subharmonics and voice irregularities for stress detection. In case of highly stressful conditions a systematic over-one-octave shift in pitch was observed. Linear Discriminant Analysis based on 9 acoustic features showed it is possible to categorize the following classes: male – stressed and neutral, female stressed and neutral speech.

Keywords: call centers interfaces, stressed speech, fundamental frequency.

1. Introduction

In many military and civilian applications it is necessary to assess whether or not a speaker is under stress. The presence of stress is also becoming increasingly important in the field of multilingual communication and security systems. Emergency call centers and police departments all over the world are bombarded with different kinds of calls, only some of which are of great importance. It would be then of particular interest to detect speech marked by stress in order to improve decisions' effectiveness and to save lives [1, 2, 3].

Several investigations - separate research on stress and emotion recognition [4] - showed direct application of emotion recognition to stress recognition [5, 6]. Thus differences in acoustical features between neutral and stressed speech brought by a variety of emotions and the Lombard effect have been studied intensively [1, 7]. A number of studies have 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 units duration) and arousal dimension [8, 9]. Their results agree that the speech correlates are dependent on physiological constraints and correspond to broad classes of basic emotions, but disagree on the differences between the acoustic correlates of particular classes of emotions [8, 10]. Certain emotional states, which can be controlled by the speaker to some extent, are often correlated with physiological states, which in turn have quite mechanical and thus predictable effects on speech and on its prosodic structure in particular. For instance, when a person is in a state of anger, fear or joy, the sympathetic nervous system is aroused and the speech becomes loud, fast and enunciated with strong high-frequency energy. When one is bored or sad, the parasympathetic nervous system is aroused, which results in a slow, low-pitched speech with little high-frequency energy. Apart from these individual differences, some studies show an increase in intensity and fundamental frequency, a stronger concentration of energy above 500 Hz and an increase in speech rate in cases of stressed speech.

While some progress has been made in the area of stress definition and assessment there is still a number of important research areas that require further investigation.

Our study focuses on the analysis of stress produced in response to the occurrences in the people's surroundings, perceived by them as unusual and impossible to be controlled.

We will analyze third order stressors, psychological ones, which have their effect at the highest level of speech production [1]. An external stimulus such as a threat is subject to individual mental evaluation, but other emotional states like anger, irritation will have also the effect at this level.

Due to methodological difficulties that concern speech under stress analysis, literature presents results that are somewhat at variance with each other. Validity of the studies however, depends heavily on the experimental material.

We assumed that separate models trained using speech from both, stressful and neutral environment, should allow to better determine acoustic stress indicators, in particular, should help answering the question which of the F_0 derivatives are most valuable stress predictors.

The structure of the remaining parts of the paper is as follows: Section 2 is a brief introduction to the speech database and training data, in Section 3 the F_0 range variability is presented, Section 4 describes stress detection and data summarization, while in Sections 5 a short discussion is given.

2. Speech corpus construction and annotation

The 997 - Emergency Calls Database is a collection of spontaneous speech recordings that consists of crime/offence notifications and police intervention requests.

In the first step of the speech corpus construction the whole set of recordings was automatically grouped into sessions according to the phone number from which the call was made, receiving over 8 000 sessions.

In the next step a six-level preliminary annotation was performed by three students trained in phonetics. The annotation included the description of: (1) background acoustics, (2) types of dialog act, (3) suprasegmental features such as speech rate (fast, slow, rising, decreasing), loudness (low voice or whisper, loud voice, decreasing or increasing voice loudness), intonation (rising, falling or sudden break of melody and unusually flat intonation), (4) context (threat, complain and depression) (5) time (passed, immediate and potential) (6) emotional coloring (up to 3 categorical labels and values for 3 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 [9, 11]).

For the purpose of investigating stress detection, only those speakers were considered who in two or more dialogs manifested different arousal levels. For each of the 45 selected speakers two speech samples were collected: one from situation with arousal level marked 0 and another one with arousal level marked 0.5 and above. Each speech sample consisted of sequences of utterances (full sentences and phrases) that were perceptually homogenous in terms of voice quality.

Voice quality features are especially effective for both, perceptual and automatic identification of the paralinguistic information expressing emotions, stress, attitudes [12]. The sequences were obtained by replacing with silence those parts of recordings that contained intrusive noise (police officer's voice, third party's voice, significant background noise).

3. Pitch variability

3.1. Pitch register

A key issue of stress detection is finding such a speech utterance segmentation that would result in units well-defined in terms of perceptual and acoustic homogeneity. Vocal register, which divides whole voice region into the component registers by the voice quality, is an important perceptual category.

There are many approaches to define vocal register, e.g. as perceptually distinct regions of vocal quality that can be maintained over some ranges of pitch and loudness or as a range of consecutive voice frequencies, which can be produced with nearly identical phonatory quality or in intonation research to describe the distance between Low and High tones [13]. Vocal register definition and terminology is one of the most controversial problems. Also modeling pitch range variation within and across speakers is a major problem [14].

In further analysis, which concentrates on the stress-caused pitch variability (leaving out linguistic and discourse features), two terms referring to pitch are introduced: pitch position, defined by F_{min} value and pitch compass, defined by F_{max}-F_{min} range. F_{min} and F_{max} were measured automatically and checked manually by an expert in phrases characterized by voice quality that was perceptually evaluated as consistent (in most cases it was the modal voice register).

Three cases have been presupposed: (1) different pitch position, same compass, (2) different pitch position different compass, (3) same pitch position, different compass.

In the following section of the paper several cases of pitch range variability selected from various situational contexts from the 997 database are presented.

3.2. Pitch range

3.2.1. Different pitch position, same compass

In this case three pitch position settings in the utterance could be observed: (a) relative constant pitch position within the utterance and dynamic pitch position changes within the utterance: (b) pitch position shifted upward, (c) pitch position shifted up and down.

a) Relative constant pitch position within the phrase

Fig.1a shows as example an utterance informing about a burglary and a life threat, whereas Fig.1b illustrates an utterance from the same person calling off the intervention (informing that the burglar has left the apartment), recorded 1 hour after the first call. In the latter case, a shift in pitch position is approx. 40 Hz lower (Fig.1b), as compared to the position in utterance from Fig. 1a.

The utterances in Fig.1a and 1b have similar pitch compasses but different pitch positions (probably caused by stress).

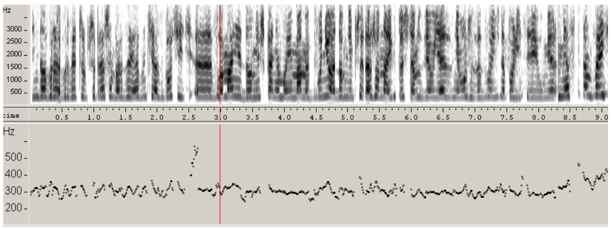

Figure 1a: F_0 contour of constant stress in the utterance: Please, come over, there's a house-breaking. She's scared to death. (F_{min} =240 Hz, F_{max}=352 Hz).

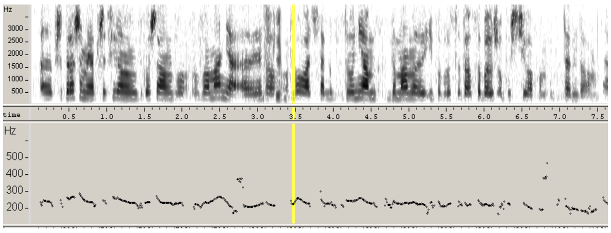

Figure 1b: F_0 contour of neutral speech in the utterance: *I called one hour ago, I want to call off the intervention.* (F_{min} =167 Hz, F_{max}=264 Hz).

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

In cases of high stress levels, F_0 can rich extreme values (female voices may reach up to 700 Hz). Fig.2 illustrates an utterance of a female marked by an extreme stress. At the start stress level increases even more. It only decreases slightly at the end of the recording after hearing a prompt to calm down. As the stress of the speaker increases, certain processes may be observed: an upward shift in the voice pitch as well as a prominence of the higher frequencies in the spectrum, an increase in the signal's energy and rate changes.

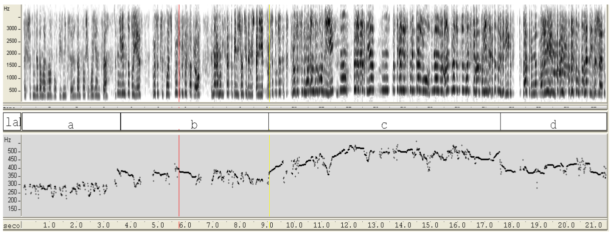

Figure 2: A gradual stress increase in the utterances: a) *Someone is entering the apartment* (F_{min} =220Hz), b) *He's masked* (F_{min} =260 Hz), c) *he is somewhere [here]* - direct threat (F_{min} =320 Hz) d*) Please come to Kwiatowa Street* - the answer after being asked by a police officer to calm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

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

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

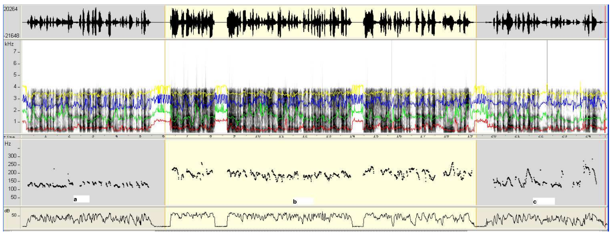

Figure 3: a) *I keep trying to get through….* (F_{min} =121Hz), b) *I've reported it so many times already…* - clearly audible irritation (F_{min} =173 Hz) c) *… so I don't know anything anymore…* - the answer after being asked by a police officer to calm down (F_{min} =115 Hz).

3.2.2. Different pitch position, different compass

In cases of anger and mixed emotions significant changes of both pitch position and pitch compass were observed. Fig.4 illustrates F_0 contour for an utterance in a female voice classified as indignation. The speaker can easily control her emotional state so that her message is clearly perceived by the listener. Each syllable which is lexically permissible is clearly stressed.

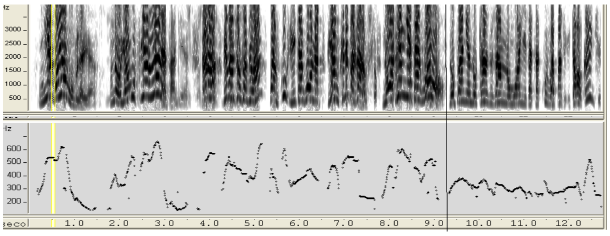

Figure 4: F_0 contour for an expressive utterance (indignation): *I've got here such a drunkard, he's maltreating me, I am going to trash him...* (F_{max}=675Hz, F_{min} =139 Hz, first part of the utterance). *But what can I do …* (F_{max}=275Hz, F_{min} =206 Hz, second part of the utterance).

As a result of discourse, the final part of the recording (beginning of which has been marked by the cursor) has a different F_{min} and pitch range width than its first part.

3.2.3. Same pitch position, different compass

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

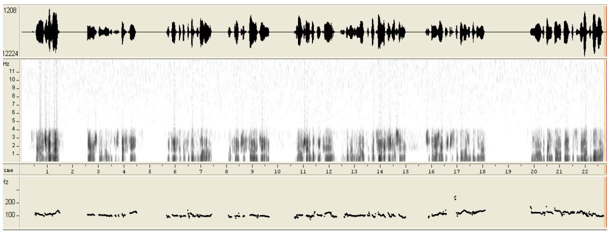

Figure 5a: F_0 contour for a neutral utterance: *Hi, I live on XXX street…* (F_{max}=137Hz, F_{min} =92Hz).

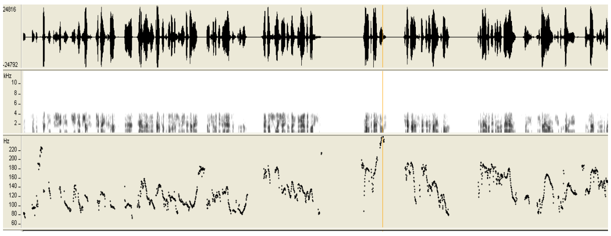

Figure 5b: F_0 contour for an expressive utterance of indignation: *I hear some shouting and name-calling… him...* (F_{max}=252Hz, F_{min} =86 Hz).

4. Stress detection

The material was divided into four groups: G1: male – stress, G2: male – neutral/mild irritation, G3: female – stress, G4: female – neutral.

The acoustical preparation of recordings consisted in the manual removal of the duty officer's voice from the recordings. For the acoustical analysis of 32 MDVP features [15], for LDA Linear Discriminant Analysis only 9 have been used: *Average (F_0), Highest (Fhi) and Lowest Fundamental Frequency (Flo), 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 9 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.

Fig.6 shows z-normalized F_{min} *(Flo)* values for 4 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.

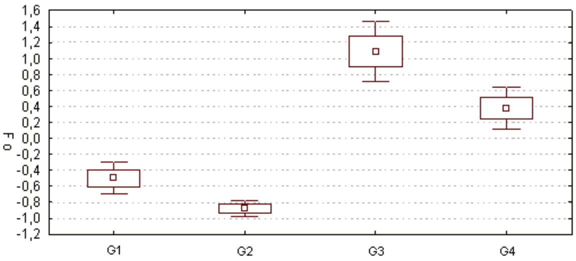

Figure 6: Z-normalized values F_{min} for G1, G2, G3, G4.

Table 1 shows classification results, slightly better for utterances by male voices affected with stress.

	% correct	**G_1:1 p=,23364**	**G_2:2 p=,27103**	**G_3:3 p=,23364**	**G_4:4 p=,26168**
G_1:1	80,000	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. Discussion

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

The results of the study confirm the significance of the F_0 parameter for investigating stress and agree with the findings by Protopapas and Lieberman [16] which point to F_{max}, as being a particularly important factor affecting the emotional stress perception. However, in the current and also previous study [17], it was stated that a shift in the F_0 contour is an important stress indicator, thus an increase in F_{min} in stressed speech is a result of a shift in the F_0 register, especially when caused by fear. A systematic increase in the range of F_0 variability for the stress related to anger or irritation was observed. This suggests, that modeling of pitch range variability, with special attention to F_{min}, should prove useful for expressive speech synthesis.

In the study the MDVP software was used, which, in spite of a relatively complex analysis, does not allow for precise evaluation of the signal's structure at the prosodic level, i.e. the evaluation of such prosodic features as accentuation and speech tempo.

The results confirmed the need of including such phenomena as the shift of pitch position and change in pitch register width into prosodic structures segmentation, particularly for expressive utterances produced under stress.

To enable better explanation of the factors that have an influence on pitch register changes in utterances diversified linguistically and in terms of situational context, it should also be analyzed how linguistic and discourse factors affect pitch position and pitch compass.

6. References

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

[2] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] Ishi, C.T., Ishiguro, H., Hagita, N., "Automatic extraction of paralinguistic information using prosodic features related to F_0, duration and voice quality.", *Speech Communication*, 531-543, 2008.

[13] 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.

[14] Shriberg, E., Ladd, D.R., Terken, J., Stolcke, A., "Modeling pitch range variation within and across speakers: predicting F_0 targets when 'speaking up'.", *Proc. Of the International Conference on Spoken Language Processing* (Addendum, 1-4), Philadelphia, PA, 1996.

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

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

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

7. Acknowledgements

This project is supported by The Polish Ministry of Sciences and Higher Education (project no O R00 0170 12).