Voice Stress Extraction

Grażyna Demenko

Institute of Linguistic Adam
Mickiewicz University, Poznań
lin@amu.edu.pl

Abstract

The aim of the research was to assess the possibility of voice stress extraction and classification. It was assumed that the study's results could be applied in call centers and could be useful for security services. The authentic Poznan police database with the recordings of the 997 emergency phone calls was used for analysis. Out of 60 000 recordings collected in the database, 20 000 were automatically selected, a few hundred of which were eventually chosen for acoustic evaluation, the basis for that selection being a perceptual assessment. The MDVP analysis confirmed statistical significance of such parameters as fundamental frequency, energy and pitch variations for stress categorization. Some segmental parameters such as tremor and noise parameters were also confirmed to be of some importance. In case of highly stressful conditions a systematic over-one-octave shift in pitch was observed. It was concluded that the range of F_0 *per se* does not seem to correlate with stress whereas the shift in F_0 register constitutes the primary indicator of stress. Linear Discriminant Analysis based on 12 acoustic features showed it is possible to categorize the following classes: neutral, depressive, stressed, highly stressed speech.

Introduction

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 reduce the officers' workload, improve decisions' effectiveness, and save lives ([5],[8]). If voice stress can be automatically detected and analyzed, it could be probably used as a viable diagnostic tool for security services as well as a means of improving speech/speaker recognition capabilities ([8],[16]). 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. similar features are used (F_0, intensity, speech units duration) and arousal dimension of emotions and stress ([1],[4]). Their results agree on the speech correlates that 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. 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, especially on its prosodic structure ([2]). 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 a strong high-frequency energy. When one is bored results in a slow, low-pitched speech with little high-frequency energy. Furthermore, the fact that these physiological effects are rather universal means that there are common tendencies concerning the acoustic determinants of basic emotions. 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 ([7],[11]). 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. The research frequently reports on conflicting results, due to differences in experimental design, categorization of actual or simulated stress, and/or interpretation of results. The studies using actors and simulated stress or emotions have the advantage of a controlled environment. The major disadvantage is, however, an artificial experimental design which can result in producing highly exaggerated misrepresentations of emotions in speech. Another group focuses on the analysis of authentic recordings coming from actual situations. There is usually no doubt as to the presence of stress in these recordings, however there is a problem of categorization of the homogeneous classes of stress. This study focuses therefore on the analysis of stress produced in response to the occurrences in the people's surroundings, perceived by them as unusual, anomalous and impossible to be controlled. The research aims at a preliminary extraction of acoustic features which are produced as a result of a relatively homogenous group of stressors.

Acoustic determinants of stress

Speech Data

Out of 20 000 recordings selected from the database, the recordings of a few hundred of different speakers were chosen for analysis. It was assumed that a particular situation which makes a person call 977 determines more or less notably the way he/she speaks. The following situational contexts were taken into account:

A. Relatively unequivocal contexts

1. Direct life hazard – extreme stress (**ES**). A situation in which the person calling or his/her friends/relatives, etc. are in a direct threat of losing their life produces extreme stress, which can be easily perceived.
2. Indirect life/heath hazard, a threat of losing property – limited stress (**S**). A situation in which the person calling or his/her friends/relatives, etc. are in an indirect threat of losing their life, health or property produces stress whose intensity depends on the individual characteristics of a person.
3. Tragedies of life – depression (**D**). A situation in which the person calling or his/her friends/relatives, etc. are potentially threatened by losing their life/health as a result of a terminal illness (or planned suicide), etc. also other misfortunes in life

can produce depression, whose intensity depends on the individual characteristics of a particular person.
4. Utterance which does not carry emotions – neutral stress (**N**). This utterance type is characteristic of informative calls, when the person calling wants to get some information concerning an address, a telephone number, etc.

B. Ambiguous contexts

When the context of a call is related to complaints, thefts, fights, etc. it can produce very diverse speech styles, depending on the speaker. For instance, a perceptually high stress, disproportionate to the incident reported (concerning e.g. breaking down of some device or having clothes stained) was observed as well as a very low stress recorded for a report of a car stealing or car accident.

C. Special contexts

Many of the recorded calls were made by intoxicated speakers suffering from different kinds of faulty articulation.

For statistic analyses only the context **A** was taken into account as it was considered quite independent of the individual features of particular speakers.

The material was divided into four groups: **ES, S, N, D**. Those groups' speakers were also divided into two groups: males and females, children being excluded. The contexts **B** and **C** which are strongly dependent on the speaker, require some detailed perceptual analyses and thus, at this point of the study, they were not taken into consideration.

Case study

A few cases from each group were chosen for a preliminary analysis whose aim was to determine the type and range of the acoustic variation of different parameters, mainly prosodic ones. Figures 1-5 illustrate F_0 variations in utterances coming from different situational contexts.

Figure 1a and 1b illustrate F_0 contours in female voices for utterances classified through a perceptual analysis as indignation. The observed change in F_0 contour is over 2 octaves (F_{min}=105 Hz, F_{max}=660 Hz). 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 and the tempo is slow. Figure 2 illustrates F_0 contour for an utterance produced by a woman suffering from depression resulting from her being captured and confined. Little variation in F_0 contour, and low signal amplitude are evident. A situation in which a caller or his/her relatives/friends are in a direct danger of losing their life produces extreme stress, which can be easily identified perceptually and acoustically on the basis of the changes in the range of F_0 and its location in the tonal space. The speech rate is exceedingly fast, the F_0 contour is relatively flat and placed in the upper part of the tonal space, the speaker has no time for clear accentuation because the speed at which he/she conveys the message is what matters. In cases of high levels of stress F_0 values can rich extreme values (even up to 750 Hz).

Figure 3 illustrates an example of F_0 contour located in the high register of the tonal space, produced by a woman under great stress resulting from her personal tragedy.

Figure 4 illustrates a gradual increase in stress. In these utterance a shift in F_0, proportional to the levels of stress, can be observed. Also an increase in signal intensity and energy in upper part of spectrum indicate the increasing stress.

Figure 5 illustrates an utterance marked by an extreme stress increase and 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 we may observe certain processes: 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. An objective evaluation of the changes in a particular voice can only be carried out if recordings of the same speaker uttering the same speech under stress-free conditions are available.

Results

Parameters extraction

For each of the four analyzed groups a few dozens of recordings representing the most unambiguous situational contexts were chosen. The recordings were acoustically prepared which consisted in the manual removal of the duty officer's voice from the recordings.

As a result of the MDVP analysis the 24 features were evaluated as following ([6]).

1. Fundamental frequency measurements: *Average (F_0), Highest (Fhi), Lowest Fundamental Frequency (Flo), Standard Deviation of Fo (STD), Phonatory Fo-Range in semi-tones (PFR).*
2. Short-term and long-term frequency perturbation measurements: *Absolute Jitter (Jita)/μs/; JitterPercent (Jitt) /%/, Relative Average Perturbation (RAP) /%/, Pitch/ Smoothed Perturbation Quotient (PPQ) /%/, (sPPQ) /%/, Fundamental Frequency Variation (vFo) /%/.*
3. Short-term and long-term amplitude perturbation measurements: *Shimmer (ShdB), Shimmer Percent (Shim) /%/: Amplitude/ Smoothed Perturbation Quotient (APQ) /%/, (sAPQ)/%/, Peak-Amplitude Variation (vAm) /%.*
4. Noise parameters: *Noise to Harmonic Ratio (*NHR) - a ratio of the harmonic energy in the range 1500-4500 Hz to the harmonic spectral energy (70-4500 Hz), *Voice Turbulence Index (VTI), Soft Phonation Index (SPI)* - ratio of the harmonic energy in the range 70–1600 Hz to the harmonic energy in the range 1600–4500 Hz.
5. Tremor parameters: *Fo-Tremor Frequency (Fftr) /Hz/*: *Amplitude Tremor Frequency (Fatr) /Hz/, Frequency/ Amplitude Tremor Intensity Index (FTRI) /%/, (ATRI) /%/.*
6. Parameters of Subharmonic components: *Number/ Degree of Subharmonic Segments (NSH)/ (DSH) /%/.*
7. Voice break related measurements: *Degree of Voice Breaks DVB /%/, Number of Voice Breaks NVB.*
8. Parameters of Voice irregularities: *Number of Unvoiced Segments (NUV), Number of unvoiced segments detected during the analysis. Degree of Voiceless (DUV) /%/.*

Statistical evaluation

The results of the MDVP analysis were stored in the STATISTICA software for the significance evaluation.

Table 1 shows statistics for the most important acoustic parameters differentiating the groups.

A preliminary assessment of the results allows drawing several conclusions.

1. The tendencies concerning the changes in parameters characteristic of stress are similar in male and female voices.
2. All the F_0 values are greater than they tend to be under normal conditions compare to the average values given in the literature on the subject, e.g. F_0=119 Hz (male voices) and 207 Hz (female voices) in ([13],[15]).

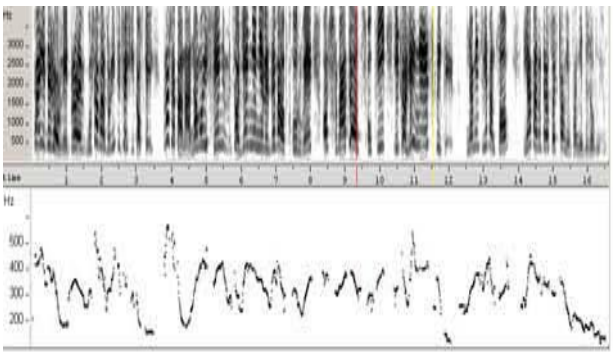

Fig.1a. Fo contour for an expressive utterance: *Mister Policemen! My son is calling me whore, bitch. Who's seen something like that!* (F_{max}=567 Hz, F_{min} =125 Hz). Indignation.

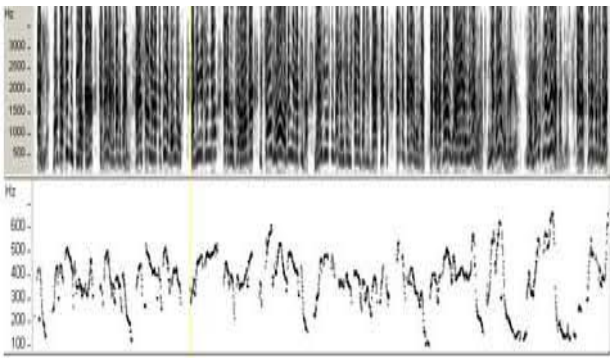

Fig1.b. Fo contour for an expressive utterance: *Hey mister. I've got here such a drunkard, he's maltreating me, I am going to trash him.* (F_{max}=660 Hz, F_{min} =105 Hz). Indignation.

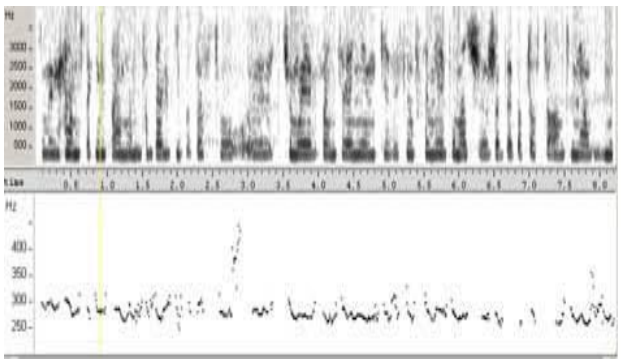

Fig. 2. Fo contour for an utterance produced by a person suffering from depression: *I left for Belgium with one man . now he's detaining me, I'm being held*. (F_{max}=302 Hz, F_{min} =240 Hz).

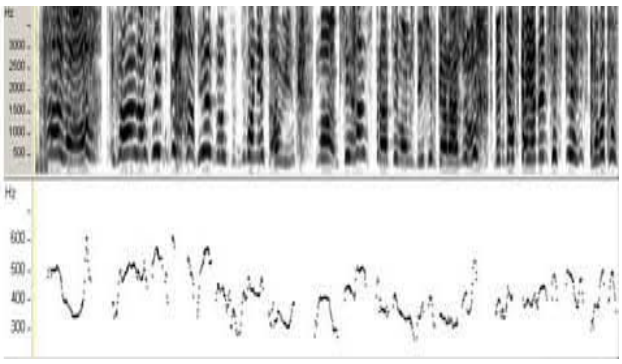

Fig. 3. Fo contour for an utterance: *Please, quickly, I came over from work, my husband is hanging in the garage. High stress.* (F_{max}=589 Hz, F_{min} =277 Hz).

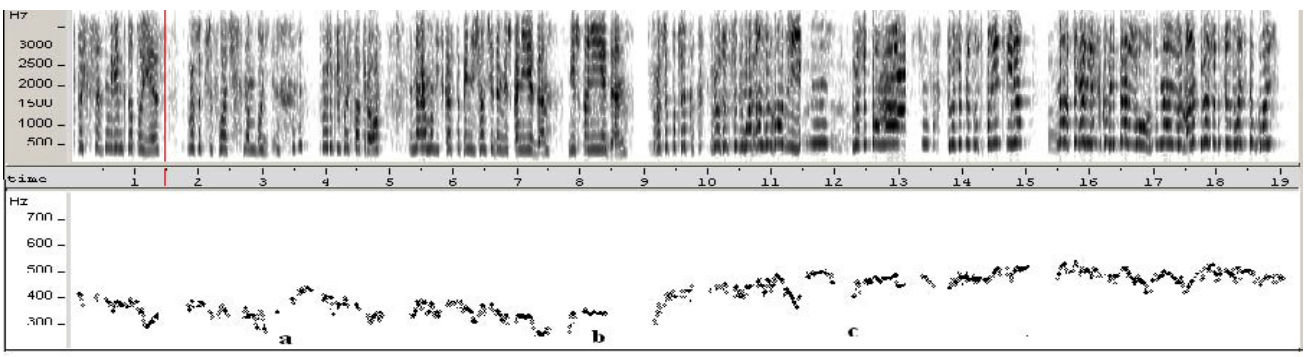

Fig.4. A gradual increase in stress in the utterances: a) someone is entering the apartment (F_{min} =315 Hz), b) someone is entering the apartment, he's masked, (F_{min} =350 Hz), c) he's leaving the room - he is somewhere [here] - direct threat (F_{min} =400 Hz).

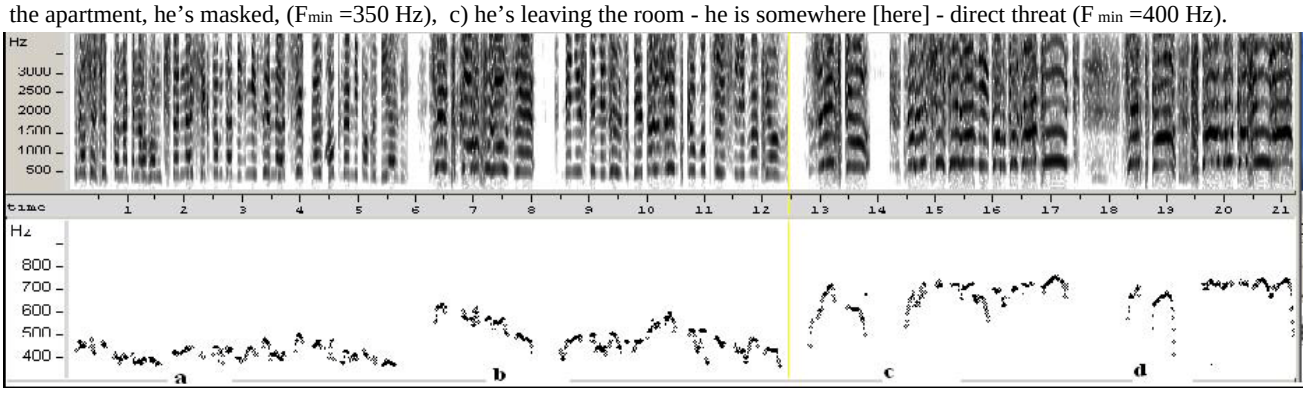

Fig. 5. A gradual increase in stress in the utterances: (a) *Please, [come] quickly to Kanałowa* [street]18, *they want to kill my son, they've broken the window* (F_{min} = 289 Hz), (b) *Miniszkiewicz Elżbieta, quickly, the mobsters have come* (F_{min} = 345 Hz), (c) *Quickly. It's happening, they want to kill him* (F_{min} =495 Hz), (d) *Quickly, Stach is killing him* (scream), (F_{min} = 495 Hz, F_{max} = 748 Hz).

What we can observe here is the Lombard effect, connected with an increase in the amplitude of telephone speech ([9]).
3. A relatively unequivocal pitch shift proportional to the levels of stress was observed and therefore pointed as the statistically most important feature.

	Female voices				Male voices			
	D	**N**	**S**	**ES**	**D**	**N**	**S**	**ES**
SP	37	134	84	48	35	118	61	30
Fo	260	240	300	366	163	143	211	238
Fhi	400	490	550	741	270	380	450	560
Flo	130	116	160	240	100	80	115	145
PFR	18	21	16	15	17	19	17	15
Fftr	4,2	3,6	3,5	3,3	5	4,1	3,5	4
Fatr	5,5	5,7	4 ,4	4,3	5	4,6	4,5	4
Jitt	4	2,9	3,7	3,8	3,9	3,4	4	4
NHR	0,3	0,2	0,3	0,3	0,3	0,3	0,4	0,4
SPI	15	9,5	9,1	8	20	13	11	9
FTRI	1,1	1,2	1,3	2	0,9	1,3	1,4	2,1
DSH	3,2	1,3	1,5	2,3	1,5	0,8	1,6	1,7
RAP	1,5	1 ,7	1,9	2 ,2	1,9	1,9	2,3	2,3

Table 1. Results of statistically significant parameters of the MDVP analysis (SP – number of speakers)

The LDA analysis of 12 parameters enabled a classification with the average 81% accuracy depending on the stress category. As only characteristics of pitch and amplitude were considered, the results showed that extreme stress can be clearly identified only by using the amplitude information with mean and minimum F_0 values.

.

Discussion

The results of the study confirm the significance of the F_0 parameter for investigating stress and agree with the findings by Lieberman ([12]) which point to F_0, and especially F_{max}, as being a particularly important factor affecting the emotional stress perception. However, in the current study it was stated that a shift in the F_0 contour is an important stress indicator, thus an increase in F_{max} in a stressed speech is a result of a shift in the F_0 register. In the study the MDVP software was used, which, in spite of a relatively complex analysis, does not allow precise evaluation of the signal's structure at the prosodic level, i.e. the evaluation of such prosodic features as accentuation and speech tempo. Therefore, on the basis of the current study, it can be only suspected that in case of being under great stress a person's main interest is in conveying his/her message as quickly as possible. As a result the F_0 range is relatively narrow whereas in case of emotional speech the speaker's aim is to express his/her emotions and thus the speech is accentuated carefully and adequately to the particular emotions (F_0 range can increase by even 3 octaves). The assessment of the relationship between stress and emotions requires, however, a complex phonetic and acoustic analysis as well as a perceptual verification ([3],[10],[14]).

Conclusions

In order to perform stress classification three anchor models are needed (one for neutral speech, one for stress and one for depression). These models should be trained with the use of speech obtained from the actual stressful environments.
The interpretation of the numerical measures, used to classify stress, needs multi-level acoustic and perceptual analysis. There is clearly a range of emotions and psychological factors which all contribute to the speaker's 'stress.' In the emergency scenarios a person may experience a combination of fear, panic, anxiety, etc. at the same time. The ability to classify/assess this mixture of speaker traits is necessary in objective determining the level of speech stress.

Acknowledgements

This research has been carried out under grant nr R00 035 02 received from Polish Ministry of Scientific Research and Information Technology.

References

1. Alter, K., Rank, E., Kotz, S. A., Toepel, U., Besson, M., Schirmer, A., Friederici, A. D. 2003, *Affective encoding in the speech signal and in event-related brain potentials,* Speech Communication, 40 (1–2), 61–70.
2. Breazeal, C., 2001, *Designing Social Robots*, MIT Press, Cambridge, MA.
3. Campbell, N., 2005, *Getting to the heart of the matter; Speech Expression of Affect rather than Just Text or Language*, Language Resources & Evaluation Vol 39, No 1, 109-118.
4. Cowie, R. and Cornelius, R.R. 2003, *Describing the emotional states that are expressed in speech*, in Speech Communication, 40, 5-32.
5. Damphousse K..R., Pointon L.Upchurch D., Moore K. 2007, *Assessing the Validity of Voice Stress Analysis Tools in a Jail Setting,* U.S. Department of Justice, Final Report Dc.219031.
6. Deliyski, D. 1993, *Acoustic model and evaluation of pathological voice production*, In E*urospeech'93 Proceedings*, 1969-1972.
7. Grimm, M. et al., *Primitives-based evaluation and estimation of emotions in speech,* 2007, Speech Commun. (2007), doi:10.1016/j.specom.2007.01.010
8. Haddad D., Walter S., Ratley R., Smith M., 2002, *Investigation and Evaluation of Voice Stress Analysis Technology Grant Final Report* (http://www.ncjrs.org/pdffiles1/nij/193832.pdf
9. Jessen, M., 2006, *Forensische Phonetik und die Auswirkungen sprechstilistischer Unterschiede auf Globalmaße der Grundfrequenz,* BKA-KT54DGFS Jahrestagung Bielefeld, 23.02.2006 .
10. Laver, J.,1995, *The phonetic description of paralinguistic phenomena*, Proceedings XIIIth ICPhS, Stockholm,1-4
11. Oudeyer, P.-Y., 2003, *The production and recognition of emotions in speech: features and algorithms*. Int. J. Hum. Comput. Stud. 59 (1–2),157–183.
12. Protopapas A., Lieberman P., 1997, *Fundamental frequency of phonation and perceived emotional stress,* J. Acoust. Soc. Am., Vol. 101, No. 4, 2268-2277.
13. Rietveld T., Vermillon P., 2003, *Cues for Perceived Pitch Register*, Phonetica 2003, 60, 261-272.
14. Scherer, K.R., 2005, *What are emotions? And how can they be measured?* Social Sci. Inf. 44 (4), 693–727.
15. Traunmuller H., Eriksson A.. 2007, *The frequency range of the voice fundamental in the speech of male and female adults* www.ling.su.se/staff/hartmut/F0_m&f.pdf.
16. *The Truth About Voice Stress Technologies Voice Stress Analysis* http://www.voicestress.com/vsa_research.htm.