Acoustic and Biomedical Engineering

Analysis of Natural Speech under Stress

G. Demenko[a,b,*] and M. Jastrzębska[b]

[a]Poznań Supercomputing and Networking Center, Zwierzyniecka 20, 60-814 Poznań, Poland

[b]Uniwersytet im. Adama Mickiewicza w Poznaniu, H. Wieniawskiego 1, 61-712 Poznań, Poland

This paper presents how voice stress is manifested in the acoustic and phonetic structure of the speech signal. Out of 60 000 authentic Police 997 emergency phone calls, 22 000 were automatically selected, a few hundred of which were chosen for acoustic evaluation, the basis for selection being a perceptual assessment. In highly stressful conditions (e.g. panic) a systematic dynamic over-one-octave shift in pitch and significant increase in speech tempo was observed. In states of depression a systematic down shift in pitch and significant decrease in speech tempo was observed. Basic statistical measurements for stressed and neutral speech run over the database showed the relevance of the arousal and potency dimension in stress processing. In speech produced under fear an upward shift in pitch register was significant (in comparison to neutral speech), while speech recorded during experiencing anger was characterized by an increase in F_0 range.

PACS: 43.72.−p, 43.72.Uv

1. Introduction

In many military and civilian applications it is necessary to assess whether or not a speaker is under stress and 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 save lives [1, 2].

Several investigations — separate research on stress and emotion recognition [3] — showed direct application of emotion recognition to stress recognition [4, 5]. Thus differences in acoustical features between neutral and stressed speech brought by a variety of emotions and the Lombard effect have been studied intensively [6, 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 [2, 8]. 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, especially on its prosodic structure. 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 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.

1. There is clearly a range of emotions which all dynamically contribute to caller's "stress". Could emotions be detectable under psychological stress?

2. How to evaluate effects of emotions on stress detection based on separate models trained using speech both from the stressful and neutral environments?

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

2. Speech corpus annotation

The whole set of recordings was automatically grouped into sessions according to the phone number from which the call was made. Using Transcriber, a group of students performed a six-level preliminary annotation on a

* corresponding author; e-mail: grazyna.demenko@speechlabs.pl

few hundreds of recordings with similar length. The annotation included: (1) background acoustics, (2) types of dialog act, (3) suprasegmental description 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) determination of the context (as the recordings come from a police emergency call database, 3 main contexts were discerned: threat, complain and depression) and metalanguage description that incorporated (5) the time aspect (passed, immediate and potential) and (6) descriptions of emotionally colored phrases (each was assigned 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) [11].

3. Emotion assessment

3.1. Case study

Figures 1–6 illustrate spectrograms and fundamental frequency (F_0) variations in utterances coming from different situational contexts.

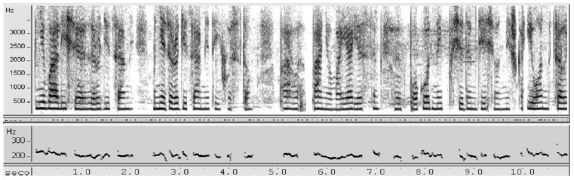

Fig. 1. F_0 contour for an utterance produced by a woman suffering from depression: "They got mad at me and want to throw me out of my apartment" ($F_{\max}$ = 230 Hz, $F_{\min}$ = 192 Hz). Slow speech rate (3.5 syll. per s).

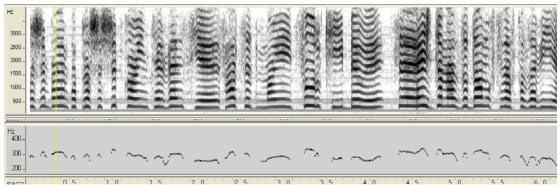

Fig. 2. F_0 contour for an utterance produced under fear (direct threat): He wants to get into our house, he wants to kill us, he is standing there..." Fast speech rate (7 syll. per s). Flat intonation contour ($F_{\max}$ = 310 Hz, $F_{\min}$ = 250 Hz).

Figure 1 illustrates F_0 contour for an utterance produced by a woman who feels helpless in her situation and her state can be described as a state of depression. Little variation in F_0 contour and low signal amplitude is evident.

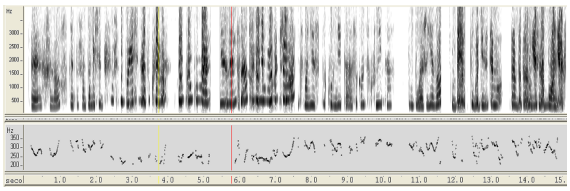

Fig. 3. F_0 contour for an utterance produced under fear resulting from an indirect threat: "I'm so scared, because I'm hiding now..." Variable speech rate. Significant F_0 contour fluctuations ($F_{\max}$ = 330 Hz, $F_{\min}$ = 210 Hz).

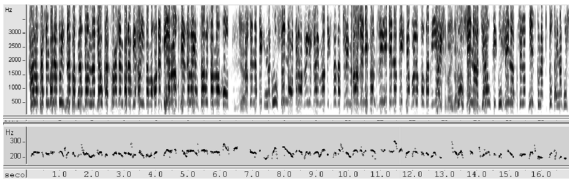

Fig. 4. F_0 flat contour for an utterance produced in a panic voice: "Halo, sir, they're robbing my house...!" Very fast speech rate (9,8 syll. per s).

Both Fig. 2 and Fig. 3 present F_0 contour produced by women suffering from fear. However Fig. 2 illustrates the case when fear results from a direct threat. In such a situation little variation in F_0 contour and higher frequency energy are observed, whereas in the second case (Fig. 3), the source of fear comes from and indirect threat and a significant variation in F_0 contour can be noticed.

The situation from Fig. 4, in which a caller or his/her friends are in a direct danger of losing their properties or lives, results in extreme stress. Hence, 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 (female voices may reach up to 700 Hz). Figure 5 illustrates an utterances marked by an extreme

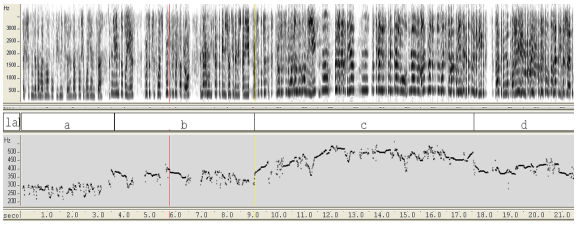

Fig. 5. A gradual increase in stress in the utterances: (a) "Someone is entering the apartment" ($F_{\min}$ = 220 Hz), (b) "He's masked" ($F_{\min}$ = 260 Hz), (c) "he is somewhere [here]" — direct threat ($F_{\min}$ = 320 Hz) (d) "Please come to Kwiatowa Street" — the answer after being asked by a police officer to calm down and tell him the address ($F_{\min}$ = 280 Hz).

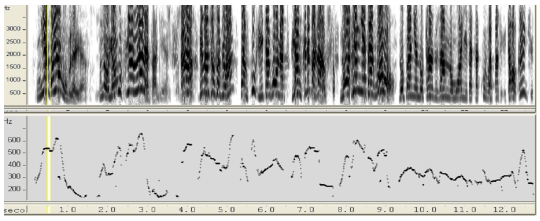

Fig. 6. F_0 contour for an expressive utterance (indignation): "I've got here such a drunkard, he's maltreating me, I am going to trash him..." ($F_{\max} = 675$ Hz, $F_{\min} = 139$ Hz).

stress increase and ended with small stress decrease. 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.

Figure 6 illustrates F_0 contours in female voices for utterances 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.

3.2. Emotion classification

The material was divided into four groups: neutral (N), depression (D), fear (F), anger (A). Speakers from these groups were further divided into two groups: males and females, children being excluded. Table shows numbers of speakers.

TABLE

Statistics of particular voice groups.

	Female voices				Male voices			
emotion	N	D	F	A	N	D	F	A
number of voices	95	37	44	48	70	35	29	30

The acoustical preparation of recordings consisted in the manual removal of the duty officer's voice from the recordings. For the acoustical analysis 9 features have been used: average (F_0), lowest fundamental frequency, standard deviation of F_0, phonatory F_0-range, noise to harmonic ratio soft phonation index, F_0-tremor frequency/, number/degree of subharmonic segments [12].

The LDA analysis of 9 parameters enabled a classification with the average 81% accuracy depending on the emotion category. The results showed that extreme stress can be clearly identified by using only the amplitude information with mean and minimum F_0 values.

4. Stress detection

For the purpose of investigating stress detection a database of recordings from 30 speakers has been collected. For each speaker there were 2 or 3 voice recordings selected, one from situation under significant stress (level of arousal marked as 1 or 0.5) and another in a neutral or close to neutral state (arousal level 0). Each recording consisted of a full sentence or a phrase with no overlapping speech (many were the cases in which the selection of such a phrase was impossible due to an audible police officer's or a third party's voice). In majority of cases the selected phrases or sentences were taken from the first part of the whole phone call conversation.

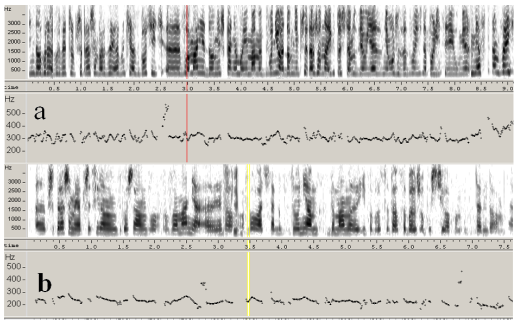

Fig. 7. (a) A constant stress in the utterance: "Please, come over, there's a house-breaking. She's scared to death" ($F_{\min} = 240$ Hz, $F_{\max} = 352$ Hz). (b) Neutral speech. F_0 contour in the utterance: "I called one hour ago, I want to call off the intervention" ($F_{\min} = 167$ Hz, $F_{\max} = 264$ Hz).

Figure 7a shows as example an utterance informing about a burglary and life threat, whereas Fig. 7b illustrates utterances from the same person calling off the intervention (informing that the burglar has left the apartment), recorded 1 h after the first call. In the latter case, a shift in pitch register is approximately 40 Hz (Fig. 7b).

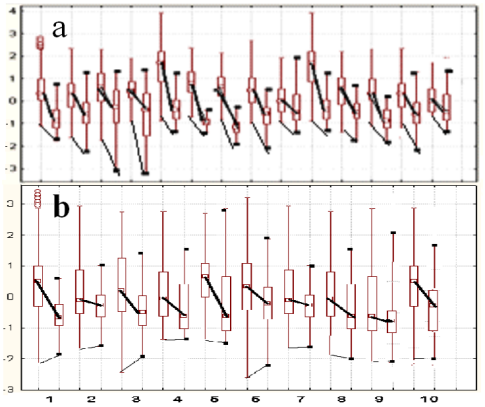

Fig. 8. (a) F_0 variability range for 14 speakers' utterances classified as the fear category. (b) Median, 25–75% F_0 variability range for 10 speakers' utterances classified as the anger category.

Figure 8a illustrates the range of F_0 variability for utterances classified as the fear category from 14 speakers. Shaded areas present the range of F_0 variability for the same speaker but in the neutral state. F_0 contour has been Z-normalized separately for each speaker. Average

and minimal F_0 values obtained for the utterances under stress and under neutral conditions, respectively, have been connected by lines.

In the case of stress related to fear an upward shift in F_0 register can be observed.

Figure 8b illustrates the range of F_0 variability for utterances classified as the anger category from 10 speakers.

There is a systematic increase in the range of F_0 variability for the stress related to anger or irritation.

5. Conclusion

In the study the MDVP software for features extraction was used, which, in spite of a relatively complex analysis, does not allow precise evaluation of the signal's structure at the prosodic level, e.g. the evaluation of intonational contours and speech tempo. 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 [13] which point to $F_{0\,\max}$, as being a particularly important factor affecting the emotional stress perception. However, in the current study 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, 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.

Acknowledgments

This project is supported by the Polish Ministry of Science and Higher Education (project no. OR00006707).

References

[1] G. Demenko, in: *Proc. Speech Prosody Conf. 2008, Campinas (Brasil)*, Eds. P.A. Barbosa, S. Madureira, C. Reis, ISCA Archive, 2008, p. 53.

[2] L. Vidrascu, L. Devillers, in: *Proc. Interspeech*,2005, p. 1841.

[3] R. Cowie, R.R. Cornelius, *Speech Commun.* **40**, 5 (2003).

[4] K. Alter, E. Rank, S.A. Kotz, U. Toepel, M. Besson, A. Schirmer, A.D. Friederici, *Speech Commun.* **40**, 61 (2003).

[5] P.-Y. Oudeyer, *Int. J. of Human-Computer Studies* **59**, 157 (2003).

[6] J. Hansen, C. Swail, A. South, R. Moore, H. Steeneken, E.J. Cupples, T. Anderson, C. Vloeberghs, I. Trancoso, P. Verlinde, Nato report (2007), `http://www--gth.die.upm.es/research/documentation/referencias/Hansen_TheImpact.pdf` .

[7] R. Huber, A. Batliner, J. Buckow, E. Noth, V. Warnke, H. Niemann, in: *Proc. Int. Conf. on Spoken Language Processing, Beijing (China)*, ISCA Archive, 2000, p. 665.

[8] K.R. Scherer, *Soc. Sci. Inform.* **44**, 695 (2005).

[9] A. Batliner, K. Fischer, R. Huber, J. Spilker, E. Noth, in: *Speech Emotion-2000*, ISCA Archive, 2000, p. 195.

[10] P. Ekman, *Cognition Emotion* **6**, 169 (1992).

[11] R.J. Fontaine, K.R. Scherer, E.B. Roesch, P.C. Ellsworth, *Psychol. Sci.* **18**, 1050 (2007).

[12] D. Deliyski, in: *Proc. Eurospeech'93*, ISCA Archive, 1993, p. 1969.

[13] A. Protopapas, P. Lieberman, *J. Acoust. Soc. Am.* **101**, 2268 (1997).