PERCEPTION OF POLISH NEUTRAL AND AFFECTIVE SPEECH BY NATIVE AND NON-NATIVE LISTENERS

Katarzyna Klessa[1], Magdalena Oleśkowicz-Popiel[1], Mariusz Owsianny[1,2]

[1]Adam Mickiewicz University in Poznań, Institute of Linguistics, Deptartment of Phonetics
[2]Polish Academy of Sciences, Poznań Supercomputing and Networking Center
klessa@amu.edu.pl, mmj@amu.edu.pl, mowsianny@man.poznan.pl

ABSTRACT

In this paper a preliminary study of perceptual judgements of Polish dialogue utterances using a two-dimensional (activation-valence), continuous feature space is described. The speech material consists of utterances selected from a corpus of task oriented dialogues, realized before and after a (fake) negative evaluation of the speakers' performance. The group of the test participants includes: (1) native speakers of Polish, (2) participants of various countries of origin without any knowledge of the Polish language.The results show e.g., that the groups of participants tend to differ more in their opinions in the dimension of valence. An additional goal of the study was a verification of the perception test procedures and feature extraction techniques (with Annotation Pro tool) as regards their future applicability for a speaker identification and characterisation project.

Keywords: speech perception, native vs. non-native listeners, correlates of affective states.

1. INTRODUCTION

The objective of the present study is to investigate selected aspects related to the perception of affective and prosodic features in speech based on Polish conversational data. When dealing with perception of affective speech (and perception of paralinguistic features in general, including paralinguistic uses of prosodic features such as pitch, cf. [13]), one of the serious challenges is the choice of speech material and the way of presenting it to the participants of a perception test. On one hand, highly controlled (often acted) stimuli may be preferred [2, 3, 27, 28, 5] (for a corresponding example for Polish data see [31] who used recordings of “emotion portrayals” from the Paralingua corpus [17]). On the other hand, the recent years witnessed a significant development of databases of authentic affective or emotional speech (e.g., Reading-Leeds, Belfast or CREST-ESP databases described in [9]). Such resources are characterized by higher naturalness and credibility of emotions. However, another types of problems may arise, due to e.g., lower recording quality, less control over recording scenarios, uneven representation of speakers and emotional states (only weak emotions may prevail or the opposite: only the extreme ones, for the latter in Polish see [8]). This can result in more difficulties in discerning and discriminating between particular emotions or affective states [9, 7, 27].

Another challenge is reflected by the discussion of the approaches to describing emotions which uncovers a wide variety of views and perspectives [27]. Some of them, which might be referred to as “taxonomic”, are based on various types of discrete emotion categories, degrees (or levels) of affect intensity. A different type of approaches could be referred to as “parametric” as they make use of continuous dimensions and rating scales (cf. also [24]). [7] report on a quite successful work based on applying time-sensitive dimensional representations of emotions while [6] (pp. 751-752) present a critical approach as regards the dimension-based approach in judgment tasks. Treating emotions as separate affective states (or “families” of states [10]) and consequently, using the taxonomic approaches to describing perceived emotions is often expected to be especially useful for canonical, strong emotions. The categorical perception of full-blown or “basic” emotions was confirmed also by [18]. However, with the more subtle states or attitudes usually encountered in everyday conversations, it is admitted that the dimensional approach might still be a justified choice.

The perception (and production) of emotional speech is assumed to depend on the listener's knowledge or level of competence of the language spoken (e.g., [1, 19, 25]). And yet another issue will be the listener's cultural background and mother tongue. Categorization of emotions is reported to be culture-specific at least to a certain extent. However, the character of the categorization and the issue of universals is a subject of discussion (e.g., [26, 10]). One of the reasons for culture-related differences in the perception of affective states or emotions can be related to the cultural dimensions of individualism versus collectivism [20] causing e.g., differences in valence or activation judgments [22]. The acoustic correlates of emotions or affective states are not always clear and have been recognised to a varying extent, especially with regard to the “milder” vocal expressions; e.g., [12] are looking for acoustic correlates of politeness and verify the results by means of a perception experiment; [18] reports on patterns of acoustic cues for both discrete emotions (six categories) and emotion dimensions (*Activation*, *Valence*, *Potency*, *Intensity*); [30] investigates acoustic correlates of emotional dimensions (*Activation*, *Evaluation*, *Power*, and their squares) as well as categories with application to speech synthesis. The results obtained from the analyses of paralinguistic features of speech including features related to affect and emotions are expected to help in the improvements of speech and speaker recognition [29].

In the present study, we investigate features of affective speech based on Polish conversational data (Section 2.1 of the paper) representing rather the subtle or milder affective states, not always obvious to define or label. A brief characteristics of selected acoustic features of the test utterances is presented in Section 2.2. Section 3 describes the perception test procedure and the group of participants. The test results are presented and discussed in Section 4, followed by conclusions and suggestions for further work (Section 5). An additional practical goal of this study is to inspect the potential applicability of the perception test procedures and feature extraction techniques within an automatic speaker identification and characterisation project. We use Annotation Pro [16], a tool employed in the project for corpus annotation and statistical annotation mining. Establishing the perception test procedures based on the same software environment will enable robust processing of both annotation data and perception test results in the future.

2. SPEECH MATERIAL

2.1. Recording scenarios and data selection

The recordings used for the present study come from a corpus of conversational speech composed of 21 task-oriented Polish dialogues [14]. The recording procedure for the corpus was the same for each of the dialogues and resulted in several parts of the recording session differing by the recording conditions and scenarios. Within each pair, one of the speakers was assigned a role of an instruction giver (IG), and the other – instruction follower (IF). The task of IG was to describe a room in his/her house while IF was asked to draw the room based only on the verbal information provided by IG. Both speakers were advised to cooperate in order to achieve the best possible result. Before the final part of the session, they were informed about the evaluation of their performance. The evaluation was fake and always resulted in a negative score (the performance being assessed as “poor”, and the drawing as “inadequately matching the description”). The result had been communicated by moderators only to IG who was then instructed to verbally inform IF. Sharing the information about the result was recorded as the final part of each recording session.

For the present study we selected samples only from the IG's utterances coming from the two session stages described above: (1) the stage directly before, and (2) the stage after the announcement of the negative evaluation. It was decided to use phrases which were as simple and neutral as possible; preferably uttered in the middle of the session stages. In case of the recordings after the evaluation announcement, an additional criterion was to use excerpts from the speakers' comments on their negative results (which was realized only by part of the speakers; some of them actually avoided to pass the failure message at all or gave a very vague record). We preferred not to use phrases explicitly saying that the task was a failure but rather those closely preceding the actual information about the assessment, and thus being more neutral or indirect in terms of lexical or semantic content but still (expected to be) affected by the negative news concerning the task failure and maybe even more by the obligation to pass on the unpleasant information to the interlocutor. Finally, 14 short utterances were selected for the needs of this study. The utterances were produced by 7 female speakers (2 utterances per speaker recorded before/after evaluation). All data were recorded in the same anechoic chamber using a head-mounted Rode HS-1 electret microphone.

2.2. Characteristics of the test utterances

Due to the data selection criteria, the test utterances were not balanced with regard to phonetic or prosodic properties. However, in order to consider the possible influences of such properties, we analysed the speaking rates, and acceleration / deceleration patterns (with TGA + Annotation Pro [11, 15]), as well as F0 and intensity values, formants, jitter, and shimmer (with Praat [4]). The overall mean of speaking rates was higher in the stage after the evaluation than before (5.4 syll/sec and 4.74 syll/sec, respectively), and accordingly, the rates calculated for individual speakers were also higher “after” than “before”, with only one exception. The inspection of acceleration and deceleration patterns showed stronger deceleration patterns for utterances produced in the “before” recording stage manifested by higher positive mean duration difference slopes. For the recordings in the “after” stage, the mean slopes were lower (approx. zero), and in two cases even slightly below zero, thus showing acceleration. The results of F0 and jitter analyses indicate that utterances produced “before” and “after” differ also in terms of these parameters (the rest of the acoustic parameters being less influential). Determining the F0 patterns appeared to be problematic for the “after” recordings due to the presence of paralinguistic phenomena such as slightly trembling voice, more occurrences of hoarse voice, and also laughter overlapping the uttered words. The utterances were characterized by the effect of coinciding periodic component of glottal excitation and other effects, e.g., creaky voice or voice trembling which led to distortions in periodicity of vocal folds vibrations and noises [23]. In order to adjust the F0 measurement parameters we used Prosogram [21]. The resulting measurements showed certain differences in F0 variability and means between the two recording session stages (with the lower overall F0 mean and smaller variability for the “after” condition but also important individual differences). According to ANOVA, the Stage factor was not statistically significant while the significance was confirmed for factors Speaker and Speaker*Stage (p level < 0.000).

3. PERCEPTION TEST PROCEDURE

The perception test was conducted using the interface available in Annotation Pro [16]. Each of the participants listened to the sound signals played from a PC individually, via headphones. The utterances were played in random order with the file names hidden. The participants were instructed to listen to the recordings and decide about the perceived degree of activation (using a continuous scale: very passive - very active) and the valence of the utterance (continuous scale very negative / unpleasant - very positive / pleasant) by clicking on a picture showing an activation-valence feature space. The subjects were not given any elaborate guidelines as regards the judgment criteria; they were only advised that they were expected to rely on their subjective judgments when giving the opinion.

3.1. Participants

The group of test participants consisted of 32 persons (university students or young researchers), including 16 native speakers of Polish (henceforth PL), and 16 listeners of various nationalities for whom Polish was a foreign language (henceforth nonPL). Table 1 provides the information about native languages of the participants in the nonPL group of listeners. Their native languages belong to various language families and can be thus seen as closer or more “distant” to Polish. Moreover, although none of the participants speaks Polish, some of them know another Slavic language(s). Considering this, the participants can be grouped into three categories: Dist1 (habitants of a country immediately neighbouring Poland and/or speaking another Slavic language), Dist2 (all other European languages), Dist3 (non-European languages). The “distance” classification is a tentative one; it was based both on geographical cues and on the interviews with speakers and their reports on their personal linguistic background.

Table 1: Number of participants of the perception test by mother tongue, and “distance” to Polish.

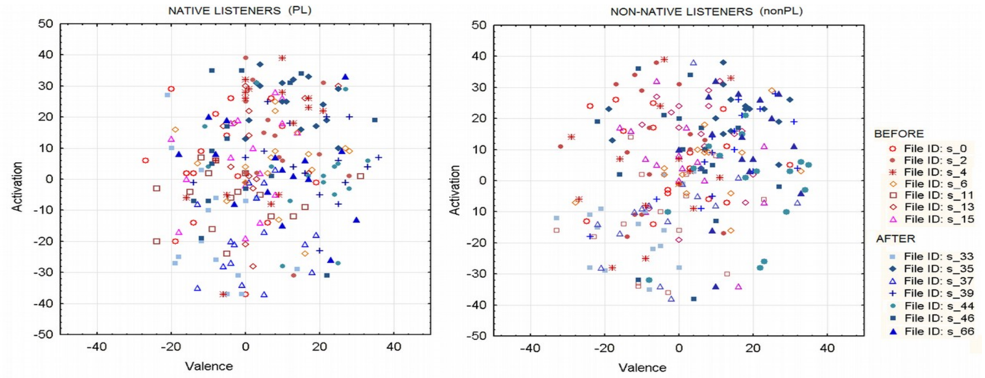
Mother tongue	No. of listeners	“Distance” group
German	2	Dist1
Russian	1	Dist1
Buryat	1	Dist1
Danish	1	Dist2
US English	1	Dist2
Finnish	1	Dist2
Hungarian	1	Dist2
Chinese	2	Dist3
Korean	4	Dist3
Laotian	1	Dist3
Thai	1	Dist3
Total	16	

4. PERCEPTION TEST RESULTS

Differences between PL and nonPL groups’ judgments appear to be more significant on the valence axis: the non-PL group used a wider range of the scale than the PL group. The opinions related to activation judgments are spread similarly for both groups (Figure 1). The results of a one-way ANOVA suggested statistical significance of differences in mean ratings of valence for factors Sound_file* Language PL-nonPL ($p<0.020$), and for the ratings of both valence and activation for the factor Sound_file ($p<0.000$).

Another statistically significant difference was observed when comparing the results obtained for the utterances produced before and after the negative assessment of speakers' performance (the Stage factor, $p<0.000$), it needs to be noted however, that the differences between stages were also strongly influenced by the speaker-related features of the utterances. An interesting observation is that the nonPL participants tended to judge some of the “before” utterances as +active and –valence more often than the PL group.

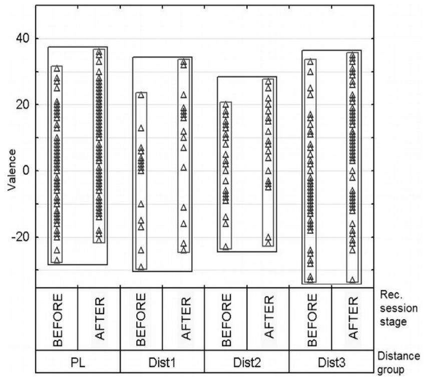
Figure 1: Perceptual responses by native (Polish) and non-native listeners for particular utterances (files IDs and recording session stages (before / after negative assessment) are shown in the legend).



As for the valence ratings, grouped by the language distance groups, the widest range can be observed for judgments given by Dist3 listeners, while Dist2 group judgments for valence were the least spread apart (Figure 2).

The latter could be explained by the fact that Dist2 group includes speakers of the European languages distant enough from Polish language not to understand the exact meaning of the utterances but close enough to share at least some nonverbal means of affect expressions. Dist3 raters' interpretations of certain non-verbal cues may be suspected to differ from European languages' speakers.

Figure 2: Valence judgements variability in two session stages (distance groups as in Table 1).



Asian languages, spoken by the Dist3 speakers, are characterized by tones, aspiration etc. aspects of phonetic means of expression different than those present in the European languages. Dist3 cultures are also especially sensitive to the sphere of politeness which can be expressed either verbally or nonverbally also by manipulating various prosodic or paralinguistic features of utterances. For example, Thai speakers frequently use laughter to mask embarrassment, disapproval, and other feelings of distress. In case of three of the „after" recordings in the present study, where laughter was partially overlapping the words spoken but no broader context was provided, PL and non-PL groups of participants generally agreed in valence judgments (high or rather high valence). The only exception were judgments provided by the Thai and one of the Chinese listeners for file s_35 who rated this utterance as definitely –valence (approx. -19, and -18, respectively, Figure 1). The activation judgments were widely spread for both groups except from the case of one file (s_35 again) attributed similarly high activation by both groups.

4. CONCLUSIONS & FURTHER WORK

This paper described selected issues in the perception of emotional and prosodic features in Polish by Poles and non-Poles, including listeners without any previous contact with both the language and culture in question. The perception tests were conducted with Annotation Pro [16] test interface, using the dimensional approach (cf. e.g., [7]) to the description of affective states, suggested as potentially useful for dealing with "milder" affective states / emotions as opposed to full-blown emotions [18]. The use of discrete categories appear as a perhaps even greater challenge when dealing with perceptual judgments by participants of various cultural background. The differences between the two group of listeners tended to be more noticeable for the valence judgments than for activation (cf. also [22] with regard to valence and attitudes to (dis)pleasure in Western and Asian cultures). The differences between the present groups of listeners became more significant when the factor of the recording session stage (before/after assessment) was taken into account.

The present speech signals were not processed before the perception test thus the native listeners could base their judgments also on the lexical or semantic information as well as on the syntactic structures of the utterances. Future research should include experiments based on extended speech data (i.a. better controlled as regards the role of linguistic contents, perhaps masked with a white noise following the ideas of e.g., [19]).

7. ACKNOWLEDGEMENTS

The present study was supported by Polish National Centre for Research and Development, project no.: **DOBR/0008/R/ID1/2013/03**. The authors also wish to thank the perception test participants from the Institute of Automation, Chinese Academy of Sciences, the participants of *INNET* Summer School in Gniezno, Poland, 2013 (funded by the FP7 grant no. 284415), and the students of the Faculty of Modern Languages and Literatures (Adam Mickiewicz University in Poznań). Special thanks are due to Maciej Karpiński, Dafydd Gibbon, and Jue Yu for their comments which helped to improve our work.

8. REFERENCES

[1] Abelin, Å., Allwood, J. 2000. Cross linguistic interpretation of emotional prosody. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK.

[2] Banse, R., Scherer, K. R. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology, 70*(3), 614.

[3] Bänziger, T., Pirker, H., and Scherer, K. 2006. GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions. *Proc. LREC, Vol. 6*, 15-019.

[4] Boersma, P., Weenink, D. 2009. Praat: doing phonetics by computer (Version 5.1.05). [Computer program] Available: http://www.praat.org/

[5] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B. 2005. A database of German emotional speech. *In Proc. Interspeech*, Lisbon.

[6] Clore, G. L., Ortony, A. Foss, M. A. 1987. The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology, 53*, 751-766.

[7] Cowie, R., Cornelius, R. R. 2003. Describing the emotional states that are expressed in speech. *Speech Communication, 40* (1-2), 5–32.

[8] Demenko, G., Jastrzębska, M. 2012. Analysis of Natural Speech under Stress. *Acta Physica Polonica-Series A: General Physics, 121*(1), A92.

[9] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P. 2003. Emotional speech: Towards a new generation of databases. *Speech Communication, 40*(1), 33-60.

[10] Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion, 6*(3-4), 169-200.

[11] Gibbon, D. 2013. TGA: a web tool for Time Group Analysis. In D. Hirst, B. Bigi (Eds.). *Proc. Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, 66-69.

[12] Grawunder, S., Winter, B. 2010. Acoustic correlates of politeness: Prosodic and voice quality measures in polite and informal speech of Korean and German speakers. In M. Hasegawa-Johnson, A. Bradlow, J. Cole, K. Liviescu, J. Pierhumbert, C. Shin (Eds.). *Proc. Speech Prosody Conference*, Chicago.

[13] Gussenhoven, C. 2002. Intonation and Interpretation: Phonetics and Phonology. *Proc. Speech Prosody Conference*, Aix-en-Provence.

[14] Karpiński, M., Klessa, K. (submitted). *Are you talking to Mike or to the mike? Paralinguistic features of speech under the influence of recording conditions.*

[15] Klessa, K., Gibbon, D. 2014. Annotation Pro + TGA: automation of speech timing analysis. *Proc. 9th LREC*, Reykjavik.

[16] Klessa, K. 2015. Annotation Pro [Software tool]. Version 2.2.1.5. Retrieved from: http://annotationpro.org/ on 2015-04-02.

[17] Klessa, K., Wagner, A., Oleśkowicz-Popiel, M., Karpiński, M., 2013. "Paralingua" – a new speech corpus for the studies of paralinguistic features. In: Vargas-Sierra, Ch. (Ed.) *Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions. Procedia – Social and Behavioral Science*. 95, 48-58.

[18] Laukka, P. 2004. *Vocal expression of emotion: discrete-emotions and dimensional accounts*. PhD Thesis. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Social Science, Department of Psychology, Uppsala University.

[19] Mathon, C., de Abreu, S., Perekopska, D. 2006. Perception of Anger in French as Foreign Language. Experimental Protocol and Preliminary results. *Proc. Speech Prosody Conference*, Dresden.

[20] Matsumoto, D., Consolacion, T., Yamada, H., Suzuki, R., Franklin, B., Paul, S., Ray, R., Uchida, H. 2002. American-Japanese cultural differences in judgements of emotional expressions of different intensities. *Cognition & Emotion, 16*(6), 721-747.

[21] Mertens, P. 2004. The Prosogram: Semi-Automatic Transcription of Prosody based on a Tonal Perception Model. In B. Bel, I. Marlien (Eds.). *Proc. Speech Prosody International Conference*, Nara.

[22] Mesquita, B., Walker, R. 2003. Cultural differences in emotions: A context for interpreting emotional experiences. *Behaviour Research and Therapy, 41*(7), 777-793.

[23] Obrębowski A. 2008, Anatomy and physiology basis of the vocal organ. In Obrębowski, A. (Ed.) *Vocal organ and its importance in the social communication*. Poznań University of Medical Sciences, 9-41.

[24] Osgood, C. E. 1952. The nature and measurement of meaning. *Psychological bulletin, 49*(3), 197.

[25] Pell, M. D., Skorup, V. 2008. Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication, 50*(6), 519-530.

[26] Russell, J. A. 1991. Culture and the categorization of emotions. *Psychological bulletin, 110*(3), 426.

[27] Scherer, K. R. 2003. Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*(1), 227-256.

[28] Scherer, K. R., Bänziger, T. 2004. Emotional expression in prosody: a review and an agenda for future research. *Proc. Speech Prosody International Conference*, Nara.

[29] Scherer, K. R., Johnstone, T., Klasmeyer, G., Bänziger, T. 2000. Can automatic speaker verification be improved by training the algorithms on emotional speech? *Proc. Interspeech*. 807-810.

[30] Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S. 2001. Acoustic correlates of emotion dimensions in view of speech synthesis. In Dalsgaard, P., Lindberg, B., Benner, H., Tan, Z-H. (Eds.) *Eurospeech*, Aalborg, Vol. 1, 87–90.

[31] Wagner, A. 2012. Emotional speech production and perception: *A framework of analysis. Speech and Lang. Technology*, 14/15 (2011/2012), Poznań.