Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 48 – 58

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Paralingua - a New Speech Corpus for the Studies of Paralinguistic Features

Katarzyna Klessa*, Agnieszka Wagner, Magdalena Oleśkowicz-Popiel, Maciej Karpiński

Institute of Linguistics, Adam Mickiewicz University, Al. Niepodległości 4, 61-874 Poznań, Poland

Abstract

This paper introduces "Paralingua" - a new speech corpus created within a larger ongoing project whose primary aim was to develop a speaker recognition and identification system for forensics. The present corpus was designed for the purpose of analysis of selected paralinguistic features in continuous speech and for preliminary examination of the vocal display of affective states. The recorded (and annotated) data include conversational speech in the form of task-oriented dialogues, emotional utterances (realized as emotion portrayals), and an acted court scene. As a reference material, a short read text was provided by each of the speakers.

Keywords: corpus design; spoken language resources; paralinguistic features; speaker characterisation

1. Introduction

Many speech corpora used as a basis for speaker recognition and identification systems are primarily expected to provide information related to low-level, short-term frame spectral features that have been proved to be useful indicators of individual human voices (e.g., Juang, Rabiner, 2005). Such systems have been continuously improved over the years and, although still imperfect, they perform well enough for practical applications in many areas of life, e.g., telephony, call centers or personal computer applications. Further improvements of their performance might be related to taking into consideration the mechanisms of sppech perception and the cues underlying recognition or identification of speakers by humans. The processes are often viewed as holistic in nature and drawing from both segmental and suprasegmental phenomena in speech. Thus, an increasing number of studies (e.g., Campbell et al. 2003, Shriberg, 2007, Farrús, 2009) indicate that the performance of the systems might be significantly enhanced by the use of higher-level, longer-term features. The higher-level features in question may be derived both from acoustic-phonetic analyses and also from perception-based tests.

*Corresponding author. Tel.: +48618293663; fax: +48618293662
E-mail address: klessa@amu.edu.pl

doi:10.1016/j.sbspro.2013.10.621

Forensic speaker identification requires a broad range of information including paralinguistic data, the latter being reported to be essential for certain methods of analysis where behavioral information of a person is based on both the verbal and nonverbal channels of communication (Inbau et al., 2004). At the same time a number of limitations need to be faced when considering the actual use of any technically supported speaker recognition or characterisation methods in court, e.g., the lack of population statistics, stimuli presentation techniques, naive vs. expert recognition differences and their potential role in implementing automatic methods of identification (Nolan, 2001).

2. Initial assumptions for designing a corpus for investigation of linguistic and paralinguistic features

As discussed above, the paralinguistic component of speech can be a rich source of valuable data for speaker recognition. Voice quality, non-linguistic prosody, as well as peculiar vocabulary choice and word sequencing in utterances may be distinctive features of individual voice profiles.

A number of available Polish speech corpora have been explored by the project team for the content, including the amount and diversity of paralinguistic features that could be extracted and analysed. *JURISDICT* (Klessa & Demenko, 2009) is a very large corpus of spoken semi-spontaneous and read police reports and court interviews (above two thousand speakers). However, due to the fact that the corpus was recorded for the needs of an automatic speech recognizer directed at taking dicatation of legal texts, the speaking styles included were aimed most importantly to elicit dictation, and thus the amount of fully spontaneous utterances in the corpus is lower and most of them are not longer than two or three phrases. Another analysed corpus was *997 Emergency Calls Database* which is a collection of live recordings from an authentic police emergency call center in Poland. The broad context of the recorded dialogs between a call center operator and a calling person in majority of cases concerns crime or offence notifications and police intervention requests. Altogether there are above 80 thousand recordings of calls from 10 thousand unique telephone numbers. For most of cases there are 2-6 calls from the same telephone number that differ from one another in emotional coloring, emotion intensity, stress level, and situational context. Therefore, the database is a valuable source of authentic emotional responses to real life situations. It provides audio material for the analysis of changes in speaker vocal responses to different situations or of the variability between speakers in their responses to similar situations. Although the advantage of using this database as a source of authentic emotional displays is undisputable as it is characterized by high ecological validity, yet there are numerous drawbacks as well. Most significant problems are as follows: (1) information about speakers' gender, age, location etc. can only be derived from the dialogs themselves; (2) lack of control over recording equipment and environment results in recordings of poor quality and rich in background noise; (3) voices of callers and call center operators overlap during dialog turn taking; (4) fully expressed emotions are not commonly encountered; instead, the recordings represent mostly weak affect and brief emotional displays, with the predominance of rather negative emotions; (5) last, but not least, the access to the corpus is restricted due to legal issues and the resource is available only for few individual researchers. *PoInt* corpus (Karpiński & Kleśta, 2001) contains high quality recordings of both semi-spontaneous monologues and task-oriented dialogues. Although initially it was intended to include emotional speech, the number of utterances tagged for emotions is very limited and the range of represented emotional states is also not significant. Moreover, only a part of the corpus is transcribed. In dialogues, there are frequent overlaps and although the speakers are recorded on separate channels, they were not acoustically separated. *DiaGest2* is a small corpus of task-oriented "origami" dialogues (Karpiński & Jarmołowicz-Nowikow, 2010). It contains both audio and video recordings and it is annotated on multiple tiers, including speech, gesture and gaze direction. The material comprises twenty dialogues (ten in mutual visibility condition and ten in the limited mutual visibility condition) and although it is rich in terms of conversational interaction, speech recordings contain noises caused by the speakers manipulating artifacts and moving around. They also tend to change the distance from the tripod-mounted microphones which influences the acoustic parameters of the recordings. Finally, corpora of acted emotional speech have been explored (e.g., Database of Polish Emotional Speech). With the small number of speakers, the insufficient technical quality as well as the limited size of utterances, their usability for the aims of the project proved to be below expectations.

In sum, although all the explored corpora were found useful for some aspects of paralinguistic features analyses, none of them met all the criteria that would make them perfect resources for the purposes of the present project. Below, the most important issues are summarized and some recommendation for the new corpus are formulated.

The question of balance between the technical perfection of recordings and the freedom and ease of speakers should be always considered. While obtaining immaculate speech recordings is always recommended as their quality can be easily downgraded on demand, it is obvious that the materials dealt with in forensic analysis are mostly of low quality. Accordingly, it seems that steps should be taken to obtain reasonable high quality but not at the cost of speakers' freedom and spontaneity. This, in turn, is related to the chance of gathering "honestly" affective speech. The key issue is to design an appropriate task that would be efficient in evoking in subjects emotions of diverse types and allow to collect them in an ethically and legally acceptable way.

Interactivity is a powerful factor in dialogue that influences the way of speaking in many ways. This factor should be taken into account in the studies of paralinguistic features as their occurrence and quality may depend on the progress of interaction and alignment processes which have been shown to correlate with the success of communication (Pickering & Garrod, 2004). Another aspect that should not be neglected is the multimodal nature of communication. An increasing number of studies on paralinguistic component of communication is based on multimodal data that include both sound and image recording. Paralinguistic features of speech are often bound to some components of gestural behavior as coming from a common source (e.g., Iverson & Thellen, 1999; McNeill & Duncan, 2000). Therefore, it is recommended to take at least the auditory and visual channels into account when analyzing various aspects of communicative behavior.

According to the above considerations, the following assumptions were made regarding the content and the structure of a corpus that would support the study of the paralinguistic component of communication in the wide context of its occurrence:

- Task-oriented semi-spontaneous dialogues focused on tasks that would naturally evoke various affective states in participants;
- Realistic settings, involving, e.g., the use of cellular phones, other popular communication devices or some other artifacts;
- When employing speakers for acted emotional speech, care should be taken to gather subjects of various backgrounds as actors who graduated from the same school as their performance may be similar and result from the techniques they learnt there; other steps towards keeping the coherence of the group are, however, justified; the design of the task as well as the comfort of the speaker also seem to be crucial in this case;
- Independent recording of the interlocutors in separate, acoustically isolated rooms – so that the separation is "natural" and not caused be artificial barriers like acoustic screens or windows;
- Achieving high channel separation in face-to-face interaction recordings without using artificial barriers that might influence the process of communication in unexpected or undesirable ways.
- Video recordings of the speakers that would allow at least for facial expression analysis and, optimally, also for gesture and posture analysis; wherever possible, multi-camera set-up would be recommended to capture more details of body motion;
- Monologue samples from all the speakers to control for the difference between interactive and non-interactive speech;
- Rich and detailed metadata on the speakers.

3. Paralingua corpus design and structure

The Paralingua corpus consists of two main subcorpora: Dial (task-oriented dialogues over the telephone) and Emo (emotion portrayals - monologues, affectively marked speech in a dialogue within an acted court scene, and imitations of authentic emotional speech).

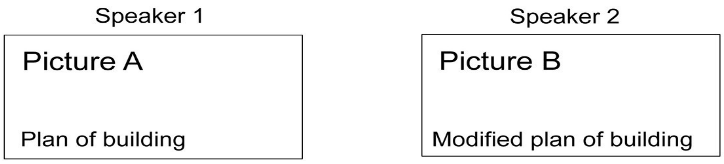

Fig. 1. Recording scenario Dial 1A session settings (a different picture for each speaker), no time restrictions.

1. 3.1 The Dial sub-corpus

The Dial sub-corpus includes recordings of quasi-spontaneous, task-oriented dialogues of 15 pairs of non-professional speakers (approximately 30 minutes of recordings per pair). During the recording sessions, subjects communicated via cellular telephones. The scenarios for the tasks were based on a prior analysis of police reports (with a view to remain consistent with the topics related to other task of the speaker recognition project). Finally, four different scenarios involving a description of a person, a building and a room were designed to be recorded by each pair of the speakers. Three of the scenarios included time restrictions (with a view to investigate possible speech rate modifications).

The main idea behind all recording scenarios for dialogue interaction was to achieve speech material conforming as much as possible to the assumptions for paralinguistic corpora (see section 2). The most important requirements were: interactive character of utterances, focus the speaker's attention on solving the tasks rather than on the fact of being recorded, obtain spontaneous or quasi-spontaneous realizations of vocabulary and syntactic structures related to the target topics, possibly long, uninterrupted stretches of speech (preferably at least several minutes of natural conversation per topic) in order to make possible the investigation of longer-term phenomena on various levels of analysis, e.g. to investigate both segmental and suprasegmental prosodic features as related to speakers, speech rates (viz. the time constraints in three of the scenarios).

3.1.1. Speakers

30 speakers participated in the recordings (22 females, 8 males) aged 20–35. All speakers were students or employees of the Faculty of Modern Literature and Languages at Adam Mickiewicz University, Poznań.

3.1.2. Recording scenarios Dial 1A and Dial 1B: Descriptions of a building and a room. Finding differences.

Scenario 1A: Each of the participants obtained one picture of a building (Fig. 1). The picture given to each person differed in a number of details from the picture given to their interlocutor. The task was to inform each other about the details seen in the pictures in order to subsequently find as many details as possible.

Scenario 1B: Each of the participants was given the same two pictures of a room with a number of differences (Fig. 2). The task was to co-operate with the interlocutor in order to find as many differences as possible in the shortest possible time (the speakers were informed that the time needed for the realization of the task was measured).

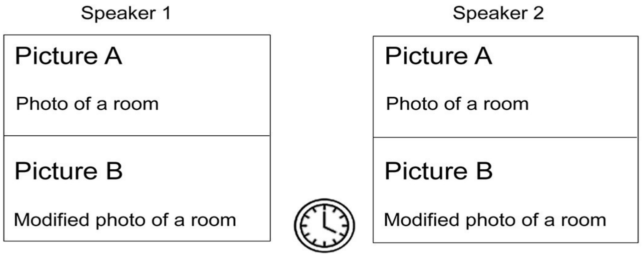

Fig. 2. Recording scenario Dial 1B session settings (the same picture for both speakers). Time restrictions applied.

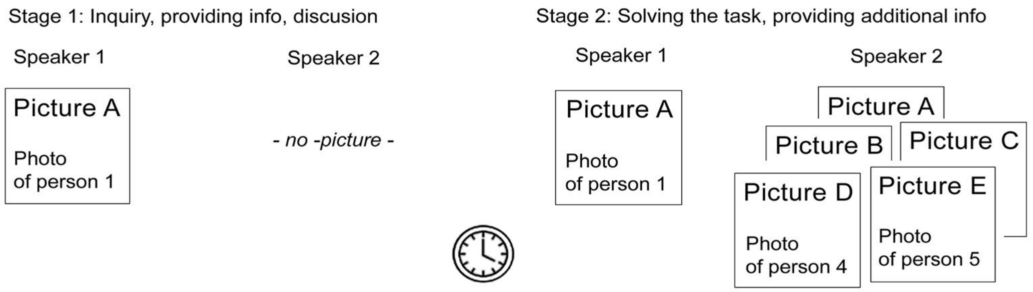

Fig. 3. Recording scenario Dial 2A session settings (person description and identification). Time restrictions applied.

3.1.3. Recording scenarios Dial 2A and Dial 2B: Descriptions of a person. Investigating details of appearance.

Scenario 2A: One of the speakers was given a picture of a person, and was instructed to provide their interlocutor with as many details of the person's appearance as possible (Fig. 3). The other speaker's task was to ask questions about the person in the picture. Both speakers were informed that the time of task completion is measured and that after their discussion there will be an additional task to perform based on the results of their discussion effectiveness of information exchange. Finally, after the discussion the person previously instructed to ask questions was presented a number of pictures of various people and was asked to identify the person described.

Scenario 2B: The speaker whose role in Scenario 2. A. was to ask questions was given a new picture of a person, and was instructed to provide their interlocutor with as many details of the person's appearance as possible. The other speaker was supposed to enquire about the details (Fig. 4). Both speakers were informed that the time of task completion is measured and that after their discussion there will be another task to perform based on their conversation results. However, in order to avoid repeating dialogue schemes from the previous scenario, it was emphasized that this task was to be different than that in Scenario 2. A. Then, the person previously instructed to ask questions was presented a number of pictures of the same person but in different postures and using various hand or head gestures and was asked to identify the one that had been described.

3.1.4. Recording conditions in the Dial sub-corpus

During the recording sessions, subjects communicated via cellular telephones. For each pair of speakers, one person was sitting in a university office room and their interlocutor was in an anechoic chamber, they had no eye contact. Dialogues were recorded using in-built telephone recorders, head mounted close-talking microphones and camcorders. For both speakers the following channels were used:

- a close-talking microphone (Sennheiser PC26 USB), output recorded using laptops (software: Sony Sound Forge Audio Studio ver. 10.0, laptops: Toshiba Satellite C660-1NZ+, Fujitsu Siemens Amilo Pro) (recording quality: 44.1 kHz, 16 bit),
- in-built telephone recorders (software: Record My Call, ver. 4.11 Android, telephones: Samsung Galaxy Ace GT-S5830 (office), LG GT540 Swift (chamber)), (recording quality: 8 kHz, 16 bit),
- camcorder (BENQ DVM23): both audio and video channel.

3.2. The Emo sub-corpus

3.2.1. Speakers

Nine speakers participated in the recordings. Among the speakers, seven are professional theatre actors working in the same theatre in Szczecin, Poland (5 females, 2 males). Additionally, two male speakers (trained speakers but not professional actors) (also living in the region of Szczecin) were recorded. Three recording scenarios were used in this sub-corpus: (1) emotion portrayals , (2) imitations of authentic emotional speech, (3) quasi-spontaneous court scene.

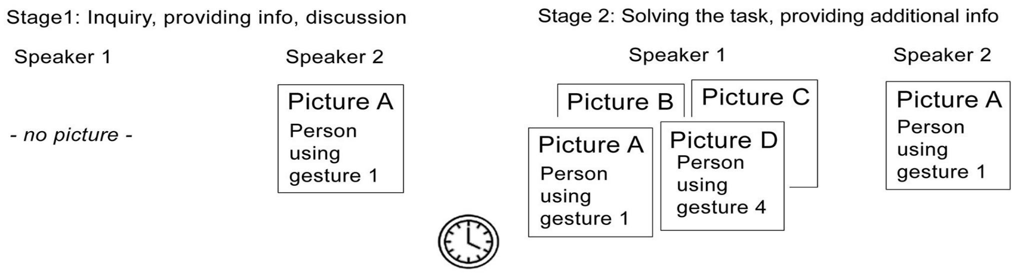

Fig. 4. Recording scenario Dial 2B session settings (person description and identification, focus on gestures). Time restrictions applied

3.2.2. Recording scenario Emo 1: Emotion portrayals

The actor portrayals sub-corpus consists of 5 sentences realized by 9 speakers in 12 emotions. Speakers were instructed to produce each of the sentences as often as they liked and until they felt satisfied with the emotional expression, which resulted in 637 utterances. The sub-corpus was designed according to detailed guidelines concerning speech material, speakers and emotion summarized further in this chapter.

Speech material guidelines for emotion portrayals:

- “Real” (no nonsense) utterances should be recorded in order to avoid overacting, producing unnatural prosody and to make it possible to place the utterance in the emotional context. The advantage of using “real” utterances instead of nonsense ones is also that they can be easily memorized, so there is no need to read them off a paper and speech production is more spontaneous. Since such utterances may appear in everyday life, it is easier to imagine that the utterance occurs in a situation whose appraisal results in a particular emotional state (this is not always the case if actors read nonsense speech material).
- Semantic and syntactic structure should be controlled: The utterances should contain semantically neutral words to avoid allocation of vocal signs of emotion to emotionally significant words. The syntactic structure should be “default” (e.g. excluding narrow focus or interjections), so that different paralinguistic information, particularly emotion, can be conveyed by changing prosody (stress patterns and intonation) and modifying voice quality.
- Semantic and syntactic structure of utterances should be neutral also in order to ensure that in perceptual annotation of speaker state using emotion labels and activation-valence dimensions judges will base their inference only on vocal cues.
- Utterances should consist of at least one prosodic phrase, so that the speaker can manipulate intonation, prominence and rhythm according to intended message (including intended emotion).
- Utterances should have high frequency of appearance in daily conversation – both syntax and vocabulary of the sentence should be commonly used in everyday communication.

Emotions guidelines:

- Basic emotions (Ekman, 1992) and emotions of a different valence and activation (Russel, 1980) should be represented as well as pairs of emotions that differ in valence or activation.

Speakers guidelines:

- A reasonable number of speakers should perform all emotions to make generalization of observed acoustic patterns of vocal communication of emotion possible. The use of acted expressions in the study of emotion was defended on many occasions by Klaus Scherer (e.g., Bänziger & Scherer, 2007, Goudbeek & Scherer, 2010).
- All speakers should utter the same verbal content in order to make comparisons across emotions and speakers possible.

The speech material in the actor portrayals sub-corpus represents members of the same emotion family with similar valence (or quality) and different activation (or intensity), i.e. *anger & irritation*, (panic) *fear & anxiety*, *despair & sadness*, emotions of a similar activation, but different valence, i.e. *interest & boredom*, *pride & shame*, and two more positive emotions: *joy* and (sensual) *pleasure*.

The texts used in the actor portrayal sub-corpus were five "real" (no nonsense) sentences consisting of one or two prosodic phrases, characterized by high frequency of appearance in a daily conversation and relative emotional neutrality of the semantic and syntactic structure (Table 1). The sentences were created in such a way as to make comparisons across emotions and speakers possible, to ensure that in the emotion recognition and classification listeners will rely primarily on vocal cues and to allow the speakers for manipulation of intonation, prominence and rhythm according to the intended message.

Altogether nine speakers were recorded including seven professional actors (five females and two males) and two trained male speakers (non-actors). Beside the definition of the emotions to be portrayed, actors were provided with two prototypical scenarios (based on scenarios created for the GEMEP corpus, Bänziger & Scherer, 2007) to support the use of acting techniques such as Stanislawsky or Acting Method.

Table 1. Sentences used in the emotion portrayals recordings

Original text	Free English translation
Teraz wszystko rozumiem.	Now I understand everything.
Dzisiaj jest poniedziałek.	Today is Monday.
Od rana pada deszcz.	It has been raining since morning.
Powiedział, że nic się nie stało.	He said that nothing had happened
Jedziemy na wycieczkę do Grecji.	We are going on a trip to Greece.

3.2.3. Recording scenario Emo 2: Imitation of authentic emotional speech

The imitated emotional speech sub-corpus is a set of actual, spontaneous dialogue utterances and their imitations made by actors (and one non-actor).

In the first step of the sub-corpus preparation 20 authentic recordings from the *997 Emergency Calls Database* (see section 2, above) were selected coming from 10 callers (5 males, 5 females, 2 phone-calls by each of the callers). For each caller one of the recordings was an exemplar of a strong negative emotion such as fear, anger or sadness, and the second one - distinctively different emotional state than the first one (e.g., emotional state close to neutral or distinctively weaker than in the first recording). Parts of the recordings in which the voice of a center operator or any personal data appeared were removed. Then, 2 to 5 utterances were extracted from each recording depending on the total duration and contents of the original dialogue. Each utterance consisted of at least one prosodic phrase but in contrast to the actor portrayals sub-corpus, neither semantic nor syntactic content of the utterances were controlled. Therefore, the text of the utterances contained authentic vocabulary effectively describing the ongoing situation, which was often emotionally loaded. Those utterances were eventually employed as models for imitations.

In the second step, nine imitators (the same speakers as in the emotion portrayals sub-corpus) were asked to repeat phrases of each callers in a fashion as close to the original as possible (prosodically, emotionally and lexically). It was decided that the imitators would imitate only the speech of callers of the same gender. Before listening to phrases and repeating them one by one, actors were asked to listen to the whole recording from which the phrases were extracted in order to familiarize him/herself with the overall situational and emotional context. Right before the recording session the imitators were also provided with a written script of each of the phrases that they were to imitate but were instructed not to read them while speaking.

3.2.4. Recording scenario Emo 3: Quasi-spontaneous court scene

The speech material included in this sub-corpus represents "quasi-natural" affective speech and was created according to the following specifications:

- Speech material in this subcorpus should be spontaneous (not read), interactive (dialogue, not monologue) and emotional (semantically, structurally, etc.). Acted speech is often "read", not spoken, and therefore it has distinctive characteristics. In both scenarios (acted and semi-natural) text should be spoken from memory and not read.

- Text should consist of dialogues rather than non-interactive monologues, because dialogue is the basic and typical setting for vocal expression of emotion.
- The use of dialogues makes it possible to take into account the socio-cultural context which significantly affects the expression of emotions (e.g., in everyday life more moderate forms of emotions can be expected rather than full-blown emotions; the expression of emotions is controlled by push and pull effects, Scherer & Bänziger, 2004). It also enables to obtain a larger semantic, structural and temporal context manifested by use of emotionally significant words, emotion-specific syntactic and stylistic forms (e.g., long vs. short phrases, occurrence of interruptions or repetitions, pauses) and intonation patterns. As regards temporal context, emotions have episodic character, therefore it seems necessary to take into account not only the emotional episode, but also its development (so called emotional build-up).
- Speech material in this subcorpus should be appropriate for the study of the relation between verbal and phonetic markers of emotion. It should contain passages of neutral speech as well as passages that show how emotion develops (till the "emotional peak") and fades away (back to neutral speech).

In order to ensure as much naturalness and spontaneity as possible. The script prepared for the actors was inspired by a pseudo-documentary TV court-show. The script is not a full screenplay, but includes detailed description of the situation, everything that happened and was said and comments concerning emotional state of the person (e.g., *you respond angrily…*). On the basis of this script actors played a short scene, so that the resulting speech material is authentic in terms of semantic and syntactic context and contains vocabulary and phrasing appropriate to given emotions (it can be expected that the resulting speech will be an adequate reflection of reality). Three actors were engaged in the recordings and they acted two scenes presenting testimony of husband and wife before judge in a divorce case. The resulting material contains approximately 20 minutes of speech.

This acted "semi-natural" speech has some limitations, but it also has some advantages over authentic speech data (e.g., collected from radio or television broadcasts) – it can be controlled to a certain extent (in terms of the type and strength of emotions or recording conditions) and can be made fully accessible (copyright problems are avoided).

The goal of collecting the "semi-natural" sub-corpus was to obtain speech material for investigation of systematic differences between acted and natural emotional speech (Jürgens et al., 2011, Barkhuysen et al., 2007).

3.2.5. Recording conditions in the Emo sub-corpus

Recordings of speech material included in Emo corpus took place in an anechoic chamber of a professional recording studio, which ensured high audio quality and minimum background noise necessary to obtain reliable spectral measurements in the future (44.1 kHz, 16 bit). The material collected includes only audio channel. The distance between mouth and microphone was controlled (the actors were instructed not to change it significantly), but it was not constant. Moreover, the recording level had to be adjusted between very loud and very quiet speech, which makes reliable signal energy analysis difficult. As concerns actor portrayals of emotions, the speakers were instructed to avoid extreme realizations, i.e. shouting (while expressing anger) or whispering (while expressing anxiety or panic fear) in order to ensure that resulting speech material is appropriate for voice quality analysis and that the comparison of voice quality between emotions is possible.

3.3. Read speech data: The North Wind and the Sun

Apart from the target types of utterances (conversational speech in Dial sub-corpus and affective or emotionally marked speech in Emo), a short read text was provided by each voice included in the database: *The North Wind and the Sun*, a fable by Aesop. The reading was recorded: (1) as a reference material for the quasi-spontaneous or elicited speech in the two other parts of the "Paralingua" corpus described in the previous sections; (2) with a view to study rhythm phenomena in the Polish read speech; (3) in comparative studies (as the corresponding text of the fable is available in a wide range of languages).

For the speakers participating in the Dial subcorpus, in each pair of interlocutors, the read speech was recorded by 15 speakers in office conditions, and by the other 15 in an anechoic chamber. The reading task was recorded using head mounted close-talking microphones, and the video camera channel. Each speaker was asked to read the text twice. The recording of the first reading took place at the very beginning of the whole recording session, before starting the dialogue tasks. The second reading was recorded after completing all the dialogue tasks by both speakers. The speakers were not given any specific instructions as regards the style of reading nor the tempo. One of the aims was to look at the potential differences between the realizations of the two reading tasks before and after performing the dialogue task.

The speakers participating in the Emo subcorpus recordings were also asked to read the text twice, according to the following instructions: (1) for the first time, please read the text in your own habitual, unmarked way; (2) for the second time, please read the text in such a way as if you were requested to read it to someone quickly but yet understandably, i.e. like in a situation when you are in a hurry, right before leaving your home for an important meeting, and on one hand you don't want to be late for that meeting but on the other hand it is important that you also read the text before you go.

4. Annotation tools, procedures and file format

4.1. Transcription and segmentation of texts

All the recordings were annotated and transcribed on the orthographic and phonetic level. First, they were transcribed orthographically and segmented into phrases by expert phoneticians using *Annotation System* software tool developed within the present project (Klessa et al., 2013b) producing XML-based annotation files. The phrase boundaries were inserted based on subjective perception judgments using pauses, pitch changes and also syntactic cues as the main boundary indicators. As the next step, the data have been automatically transcribed on the phone level and time-aligned using techniques developed in *Polphone* (Demenko et al., 2003) and *Salian* (Szymański & Grocholewski, 2005) integrated into *Annotation System*.

4.2. Annotation of prosody and phonetic description of voice quality

For the purpose of continuous annotation of perceived *F0* and prosody graphical representation was used (Fig. 5a): it depicted a circle divided into six areas corresponding to pitch level, pitch variability, tempo, loudness, pitch range and articulation. Between the center and the edge of the circle, a smaller circle was drawn which corresponded to the unmarked value of a given feature, e.g., normal tempo or moderate pitch range, whereas the edge and the center corresponded to the maximum and minimum of the feature respectively.

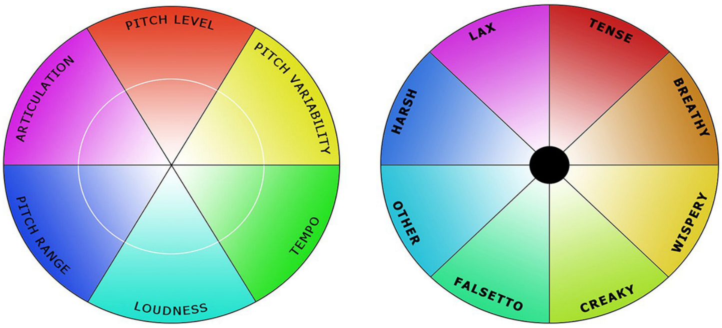

Fig. 5. Graphical representation of the feature space for annotation of perceived F0 and prosody(a, left), and voice quality (b, right)

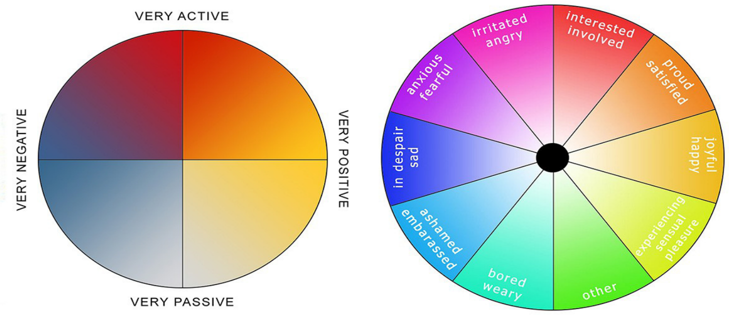

Fig. 6. Graphical representations used for annotation of perceived emotional state: valence and activation (a, left), emotion families (b, right)

Preliminary annotation of perceived voice quality was also carried out using graphical representation (Fig. 5b). On the basis of the phonetic voice quality description (Laver, 1980) the following qualities were distinguished: lax, tense, breathy, whispery, creaky, falsetto, harsh, modal and other. The middle of the circle corresponded to modal voice and the distance from the center to the edge indicated the intensity of a given voice quality (increasing towards the edge). The task of the labeler consisted in placing the cursor in the area corresponding to perceived prosodic and voice quality features.

4.3. Speaker state (emotions)

Perceptual annotation of speaker state was carried out using two graphical representations. One of them illustrated a circle divided by *x* and *y* coordinates into four areas corresponding to emotion dimensions of valence (*x*-axis) and activation (*y*-axis, Fig. 6a). The other one depicted a circle divided into ten areas corresponding to emotion labels describing perceived speaker state (Fig. 6b): *irritated/angry, interested/involved, proud/satisfied, joyful/happy, experiencing sensual pleasure, bored/weary, ashamed/embarrassed, in despair/sad, anxious/fearful* and *other*. Emotions which belonged to the same family and differed only in activation (intensity) were collapsed into one category. They could be distinguished by the distance from the center of the circle which corresponded to greater or lesser intensity of the perceived emotion (i.e., intensity decreased from the center to the edge of the circle). Perceptual annotation of emotional state of the speaker consisted in placing the cursor in the appropriate area of the graphical representation. The resulting coordinates were automatically displayed on the associated annotation layer, saved in a text file and then exported to a spreadsheet.

5. Conclusion and future work

Currently, a higher level annotation for prosody perception, voice quality, and speaker state (using categorical and dimensional representation of emotions) is carried out using the proposed methodology of annotation supported by graphic representation of feature spaces. Preliminary results obtained from the Emo subcorpus show that the expression of different emotions is highly speaker-specific, but some consistencies in the use of pitch level, pitch range and speech rate variation as vocal cues of emotions can also be observed. The Dial corpus has been examined for individual variation in the choice of lexical units, the length and type of phrases, and also voice quality parameters (Klessa et al., 2013a). Initial results also confirm speaker-specific character of the features under study. The observed tendencies will be verified in subsequent statistical analyses based on the entire material.

Acknowledgements

(1) This work was supported from the financial resources for science in the years 2010–2012 as a development project (project no. O R00 0170 12). **(2)** The authors wish to thank the speakers recorded in the Dial sub-corpus who voluntarily agreed to participate in the recordings. **(3)** The anechoic chamber used for the recordings of the Dial sub-corpus was made available by the Center for Speech and Language Processing at Adam Mickiewicz University, Poznań.

References

Bänziger, T., & Scherer, K. (2007). Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. *Affective computing and intelligent interaction*, 476-487.

Campbell, J. P., Reynolds, D. A., & Dunn, R. B. (2003). Fusing high-and low-level features for speaker recognition. *Proceedings Eurospeech Vol. 166.*

Demenko, G., Wypych, M., & Baranowska, E. (2003). Implementation of Grapheme-to-Phoneme Rules and Extended SAMPA Alphabet in Polish Text-to-Speech Synthesis. *Speech and Language Technology, Vol. 7*, 79-97, Polish Phonetic Association, Poznań.

Farrús, Mireia. (2009). Fusing prosodic and acoustic information for speaker recognition. *International Journal of Speech Language and the Law 16.1*, 169-171.

Inbau, F., Reid, J.E., Buckley, J.P., & Jayne, B.C. (2004). *Essentials of the Reid Technique: Criminal Interrogation and Confessions*. Jones and Bartlett Publishers.

Iverson, J., Thelen, E. (1999). Hand, mouth and brain: The dynamic emergence ofspeech and gesture. *Journal of Consciousness Studies, 6*, 19-30.

Juang, B. H., & Lawrence R. Rabiner. (2005). *Automatic speech recognition–a brief history of the technology development*. Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara.

Karpiński, M., Jarmołowicz-Nowikow, E. (2010). Prosodic and Gestural Features of Phrase-internal Disfluencies in Polish Spontaneous Utterances. *Proceedings of Speech Prosody 2010*, Chicago.

Karpiński, M., Kleśta, J. (2001). The project of an intonational database for the Polish language. In St. Puppel, G. Demenko (Eds.) *Prosody 2000*. AMU Faculty of Modern Languages and Literature.

Klessa, K., Demenko, G. (2009). Structure and Annotation of Polish LVCSR Speech Database. *Proceedings of Interspeech*, 1815-1818, ISCA 2009, Brighton.

Klessa, K., Karpiński, M., Wagner, A. (2013a). Annotation Pro – a tool for annotation of linguistic and non-linguistic features. Manuscript submitted for publication.

Klessa, K., Wagner, A., Oleśkowicz-Popiel, M. (in press). Using "Paralingua" database for investigation of affective states and paralinguistic features. *Speech and Language Technology, Vol. 16.* Polish Phonetic Association. Poznań.

Laver, J. (1980) *The Phonetic Description of Voice Quality*. Cambridge Studies in Linguistics. Cambridge University Press.

McNeill, D., Duncan, S. (2000). Growth points in thinking for speaking. In: McNeill, D. (Ed.) *Language and gesture.* Cambridge: CUP, pp. 141-161.

Nolan, F. (2001). Speaker identification evidence: Its forms, limitations, and roles. *Proceedings of the conference 'Law and Language: Prospect and Retrospect'*. Retrieved from: <http://www.ling.cam.ac.uk/francis/LawLang.doc>.

Pickering, M., Garrod, S. (2004). Toward a mechanistic psychology of dialogue. Behavioral and Brain Sciences, 27, pp. 169-190.

Scherer, K. R., & Bänziger, T., (2010a). On the use of actor portrayals in research on emotional expression. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 166-176). Oxford, England: Oxford university Press.

Scherer, K. R., & Bänziger, T., (2010b). *Scenarios - GEMEP (GEneva Multimodal Emotion Portrayals)*. Retrieved from: <http://www.affective−sciences.org/system/files/page/2289/GEMEP%20Scenarios.pdf>.

Shriberg, E. E. (2007). Higher level features in speaker recognition. In C. Muller (Ed.) *Speaker Classification I. Lecture Notes in Computer Science / Artificial Intelligence*, Vol. 4343 (pp. 241–259). Heidelberg, Berlin, New York: Springer.

Szymański, M., & S. Grocholewski. (2005). Transcription-based automatic segmentation of speech. *Proceedings of 2nd Language and Technology Conference*, Poznań. 11–15.