Annotation of paralinguistic features for speaker characterisation

Katarzyna Klessa

Abstract

The study aims at describing selected aspects of annotation of paralinguistic features and the possible further uses of the annotated material in speaker characterisation and identification. The ambiguity of theoretical background, various attitudes to categorisation, continuous feature character lead to a wide range of interpretations and numerous implementation problems that make the annotation of paralinguistic features a highly complicated task. Underlying the discussion of paralinguistic annotation and its applicability to speaker characterisation or identification is the fundamental question of the definition of paralinguistic features. Definitions found in the literature vary and sometimes even contradict one another. In the present study, a framework for the annotation of linguistic and paralinguistic features is introduced together with selected details concerning the related data and metadata file format. A paralinguistic profile of the speaker's voice is proposed as the annotation framework's test-bed and a possible future enhancement for speech characterisation process.

1. Identifying paralinguistic features and their use in speaker characterisation

Defining paralinguistic features has been a subject of discussion since quite a long time (Crystal, 1966, 1974; Trager, 1958, 1961). The definitions of linguistic, paralinguistic or extralinguistic features formulated since then still overlap or even contradict one another (Schötz, 2002). What is more, the problem of overlapping and vagueness of categories seems to be present even within the scope of some of the traditionally less controversial feature spaces, such as for example the structure of utterances and the choice of lexical means. When such features are investigated from the perspective of individual realisations of utterances and the specificity of speaker's behaviour (speaker-characteristic repetitions of words, the individual choice of specific phrases or structures), their status becomes rather idiosyncratic. The practically adopted definitions of paralinguistic features might also strongly depend on application, varying between their uses in the fields of forensics, psychology, education, sociology, etc. (e.g. Allen, 1999; Ethier, 2010; Liscombe, 2007; Rose, 2003; Schuller et al., 2010).

For the needs of the present project, a working definition of paralinguistic features has been accepted (after Karpiński, 2012), according to which paralinguistic features are understood as all such features that do not fit in the linguistic system but still somehow contribute to the final meaning of the utterance by

providing cues to its contextually appropriate interpretation and enhancing available characteristics of the speaker.

Most of the currently developed systems for speaker characterisation, recognition or identification are based on short-term spectral features. However the positive impact of adding higher-level or longer-term features has also been reported (e.g. Shriberg, 2007). Using paralinguistic information has been shown to be especially important in certain methods of forensics (Inbau et al., 2004), although a number of limitations need to be kept in mind when considering the actual use of any technically supported speaker recognition or characterisation methods in court (e.g. the lack of population statistics, naïve vs. expert recognition differences, stimuli presentation techniques; Nolan, 2001).

2. Annotation challenges: feature categories and feature spaces

Paralinguistic features might be categorised or grouped in multiple ways depending on the prospective application. A different categorisation might be expected from linguists, phoneticians, speech technologists or engineers, and further differentiation will be seen as a result of the level of analysis. For certain applications it appears sufficient and adequate to distinguish a small number of inherently diversified categories, e.g. to treat “speaker noises” (understood as all “physiological” noises made by the speaker such as sneezing, breathing, coughing, etc.) as one category and “fillers” (non-lexical hesitation sounds) as another (Fischer et al., 2000) without applying any detailed sub-categorisation for the two labels. This type of approach proved to be successful for computer speech recognition based on read, dictation-style and formal speech (e.g. Demenko et al., 2012). However, for the needs of automatically recognising informal, interactive speech as well as for speaker characterisation, a more sophisticated approach is required. The acoustic, phonetic or perceptual correlates of paralinguistic features have been recognised to a different extent, as well as their multilateral interactions and influences (e.g. Geumann, 2001; Grawunder & Winter, 2010; Minematsu et al., 2006; Schröder et al., 2001). Human perception of individual characteristics of speech assumes treating the whole range of co-occurring speech and non-speech events in a holistic way. Although various types of information might be processed separately, they interact during speaker or person recognition (von Kriegstein et al., 2005; von Kriegstein & Giraud, 2006). Not only is it advisable to treat voices as multidimensional objects (Rose, 2003) but also to consider at least some information as regards features related to interactive character of speech communication (e.g. in relation to the interlocutor) as well as information about the environment and situational context

which may significantly influence speaker's vocal behaviour. Considering these cues, it appears justified to look for wider-range of cues to rely on in order to reach the closest possible approximation of the real interpretations by human listeners.

Defining unambiguous boundaries between feature categories and tracking the feature values in the process of annotation of spontaneous or affective speech is problematic even when expert annotators are engaged. With some features it is then useful to use graphic feature continuum or space representation instead of arbitrarily assigning a set of categories or parameters (cf. *Feeltrace*, a tool developed by Cowie et al. (2000) for the analysis of emotional speech using a two-dimensional space representation derived from psychology).

In the present work it was decided to apply the method of using a graphical representation of feature space to annotation of various types of linguistic and paralinguistic features in spontaneous speech. A potential additional effect of using the visual representation of the feature space is the possibility of discovering new clusters of feature values and subsequently defining new categories based on the analysis of the annotation results.

2.1. Software framework: Annotation System

For the needs of the present project a new software tool was developed, named Annotation System[1]. The software was created using C# programming language for Windows operating system. Apart from the "traditional" multi-layer annotation interface (accompanied by both spectrogram and waveform signal display), a universal graphic control was implemented in the program which enables using various graphical spaces as a basis for annotation. Figure 1 shows the program's interface. The graphic control is visible in the right top corner of the program window, and in this case, an example visualisation of the phonation types continuum (based on Ladefoged, 1971) has been selected as the annotation space. Instead of this continuum, the user may select another picture (e.g. a simple feature-degree strip or area (where the target feature is to be specified by the user) or an example representation for emotion-appraisal space and several other). It is also possible to create one's own picture representing any desired two-dimensional feature space.

The plain to which the graphic control picture is related is interpreted by the software as the Cartesian coordinate system. When the user clicks on the picture, the coordinates of the clicked points are stored and displayed in the related "tra-

[1] The software will be made publically available for non-commercial research use after the end of the present project in 2013. Contact e-mail: klessa@amu.edu.pl.

ditional” annotation layer. While the user clicks on the picture during the sound is being played, the subsequent clicks result in the automatic insertion of segments in the annotation layer and the corresponding coordinates as annotation labels. The number of segments and their distribution over the layer’s timeline is directly connected with the selections made by clicking the points in the graphic representation control. As a result of this procedure a collection of coordinates is obtained for which it is then possible to conduct a range of analyses, e.g. cluster analysis.

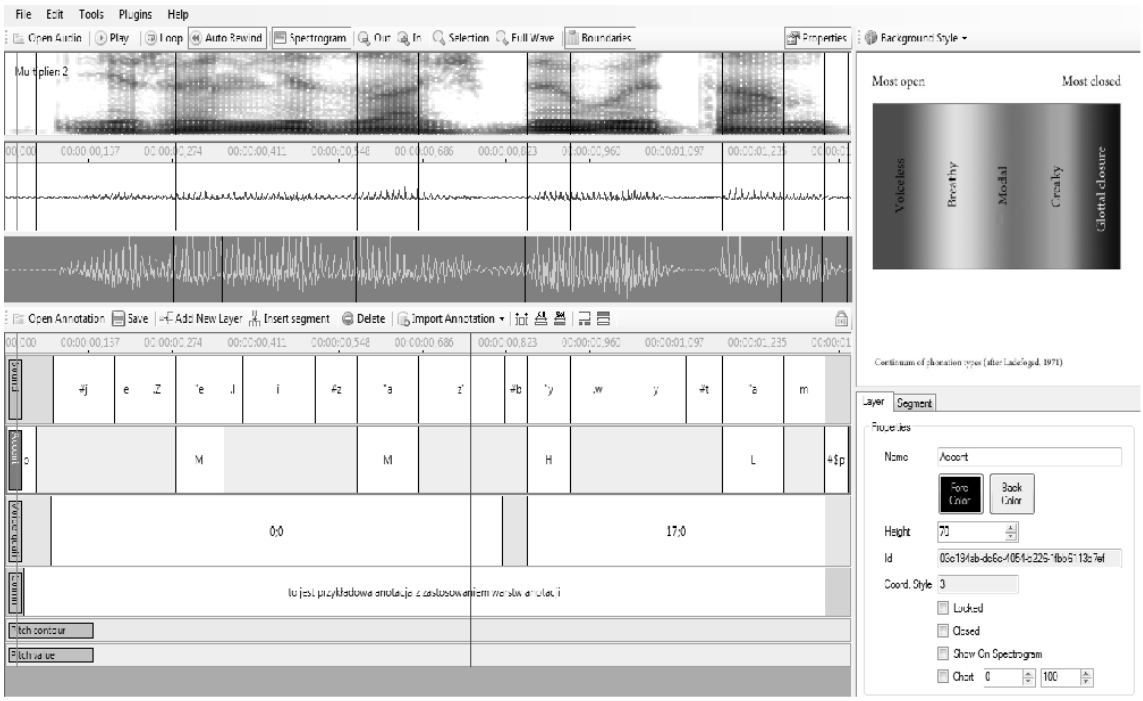

Figure 1: Annotation System interface. The definable graphic control panel (top right corner) might be replaced be another picture selected by the user (here: Phonation types continuum, after Ladefoged, 1971).

2.2. Data and metadata XML file format

The Annotation System file format is the XML format. The format enables storing data for annotation of multiple recordings from multiple speakers using any desired number of annotation layers. It is assumed that the annotation of paralinguistic features will be made independently from the existing annotation of another type (e.g. orthographic) and that usually it will be saved in a separate annotation layer. As a result a label including orthographic transcription in the XML does not differ technically from a label including paralinguistic annotation. The segment including a paralinguistic label may be stored at a separate layer for this particular feature or at any other layer selected by the user. The

three crucial elements of the Annotation System's XML file are <Speaker>, <Label>, and .

Fragments of an example XML file are shown below:

```
<Speaker>
<Id>cd567434-4345-5434-7654-54566778hb5g</Id>
<Name>Speaker1</Name>
<Age>36</Age>
<Gender>male</Gender>
<Origin>Great Poland</Origin>
<Nationality>Polish</Nationality>
<NativeLanguage>PL</NativeLanguage>
<Weight>80</Weight>
<Height>182</Height>
<Custom1></Custom1>
<Custom2></Custom2>
<Custom3></Custom3>
<Description>Opis mówcy</Description>
<Education>higher</Education>
</Speaker>
```

```
<Layer>
<Id>c656097c-6b48-4b24-9d43-265cc26b26a9</Id>
<Name>New layer...</Name>
<ForeColor>-16777216</ForeColor>
<BackColor>-5383962</BackColor>
<IsSelected>true</IsSelected>
<Height>24</Height>
<CoordinateControlStyle>0</CoordinateControlStyle>
<IsLocked>false</IsLocked>
<IsClosed>false</IsClosed>
<ShowOnSpectrogram>false</ShowOnSpectrogram>
<ShowAsChart>false</ShowAsChart>
<ChartMinimum>0</ChartMinimum>
<ChartMaximum>100</ChartMaximum>
<IdSpeaker />
</Layer>
```

```
<Segment>
<Id>708a9f98-e2a7-4d69-9ed1-607ec10d8156</Id>
<IdLayer>c656097c-6b48-4b24-9d43-265cc26b26a9</IdLayer>
<Label>23;12</Label>
<ForeColor>-16777216</ForeColor>
<BackColor>-1</BackColor>
<BorderColor>-65536</BorderColor>
<Start>70500</Start>
<Duration>10000</Duration>
<IdSpeaker />
<Feature>Voice Quality</Feature>
<Language>EN</Language>
</Segment>
```

The fundamental annotation unit is the whose basic parameters are *Label*, *Start* and *Duration*. The <Label> element delivers information about the annotation label. The interpretation of the <Labels> depends on the <Feature> element, i.e. when <Feature> is set as different from default (for example VoiceQuality), then the <Label> is known to include values of a feature different than the direct transcription of the utterance. This way it is possible to easily distinguish the types of labels included in particular segments. Similarly, setting <Language> element as different from the default language for a speaker (<Speaker>) means that this particular segment has been uttered by the speaker in a language other than their native language.

Any other relevant information related to the file, speaker, corpus etc. is stored using <Configuration> elements. This element is of dictionary type, and includes keys and values. The keys need to be unique. The following keys have been reserved for a set of standard properties:

- Created (date of creation)
- Modified (date of modification)
- Version (version name)
- ProjectTitle (title of the project)
- ProjectEnvironment (characteristics of the recording environment)
- ProjectNoises (description of background noises characteristic to the project)
- ProjectCollection (the name of the collection including the project)
- ProjectCorpusType (the type of corpus)
- ProjectCorpusOwner (the name of the owner of the corpus)
- ProjectLicence (licence, information of the availability of the project)
- ProjectDescription (another description of the project)

The Annotation System can open an XML file created in an external tool if it's format is compatible with the above specification. Any information that has not been pre-defined in the Annotation System should be included in the XML file using the <Configuration> elements. The Annotation System will open such files, ignore the "foreign" information, but it will not be lost. Thanks to this solution, it is possible to make use of the Annotation System on an intermediate, lossless basis, e.g. in order to annotate paralinguistic features using the graphic control, and to return to another annotation tool. It is important, though, not to use the reserved keys enlisted above, since these can be modified during file edition in the program.

An example notation of a <Configuration> element in the XML is shown below:

```
<Configuration>
<Key>ProjectCorpusType</Key>
<Value>spontaneous dialogue</Value>
</Configuration>
```

Apart from opening the above XML files, the Annotation System can import files of external formats: Transcriber's TRS (Barras et al., 2001), Wavesurfer's (Sjölander & Beskow, 2000) BLF, and also from TXT files (each verse will be imported to a separate segment in the selected annotation layer).

3. Towards the paralinguistic profile of the speaker

The annotation software introduced in the previous section has been created with a view to support processing data within a speaker characterisation research-development project. The next step is the selection of the features that might serve as an enhancement for speaker characterisation process by adding information based on longer-term features with a special focus on perceptual judgments of multidimensional voice features in conversational contexts. This step will be taken as a supplement to modelling based on short-term spectral information based on automatic feature extraction.

After an investigation into the JURISDIC large vocabulary speech recognition database (Klessa & Demenko, 2009), analysis of the annotation procedures established for the Polish police emergency call database (Demenko et al., 2009), as well as the annotation of a newly recorded dialogue corpus (Klessa et al., 2013), a set of features has been formulated as a basis for the paralinguistic profile of the speaker. A summary of the profile is presented in Table 1.

Table 1: Paralinguistic speaker profile: features related to longer-term phenomena, subjective judgements or meta-data. The abbreviation PUT stands for "Per Unit of Time".

Feature	Description
Gender	male / female
Age	age in years
Region of origin	name of the geographic region
Language	- native / non native - language's ISO code given
Perceived voice quality (VQ)	- stability over a period of time (variability within utterances) - VQ changes as related to the utterance structure - the overall judgment of speaker's VQ on a continuum scale
Perceived expressivity (EX)	- stability over a period of time (variability within utterances) - the overall judgment of speaker's EX on a continuum scale
Perceived stress level	- the perceived level of stress on a continuum scale
Non-verbal fillers	- vowel-like, nasal-like, compound (vowel-nasal), quasi-verbal ("hm", "mhm"), unclassified (number PUT given for

Feature	Description
	each, perceived intensity marked on a continuum scale)
Number of self-repairs	- phrase level repairs - number PUT given - word level repairs - number PUT given
Non-speech speaker noises	- laughter, cough, yawn, breath, sigh, lip smack, sneeze, swallow, unclassified (number PUT given for each, perceived intensity marked on a continuum scale)
Verbal tics	words repeated unconsciously, functioning as verbal fillers or adding emphasis
Specific lexical items	speaker-characteristic lexical item(s)
Specific syntactic structures	speaker-characteristic syntactic structures
Interjections towards the interlocutor	- the number of interjections in the course of the interlocutor's utterance: not related to further turn-taking by the speaker / followed by further turn-taking by the speaker
Reaction to the interjection by the interlocutor	- turn-giving as a result of interlocutor's interjection (number of occurrence) - utterance continuation despite interlocutor's interjection (number of occurrence)
Repetitions after the interlocutor	- number of word-level repetitions after the interlocutor - number of phrase-level repetitions after the interlocutor
Speech rate	- number of speech units (e.g. speech sounds) PUT - subjective judgment of speech rate labelled on a speech rate continuum
Vocal Pitch	- perceived height of voice labelled on a pitch continuum - long-term fundamental frequency mean / variability
Voice Intensity	- perceived intensity labelled on an intensity continuum - long-term intensity mean / variability

4. Conclusions and future work

In the present paper, a framework for investigation of continuous and categorial paralinguistic features has been presented together with a software solution and the corresponding XML-based data and metadata file format. The first use of the framework is the on-going verification of the paralinguistic speaker profile introduced further in the paper. The described feature set is currently being tested in the annotation process of a corpus of Polish task-oriented dialogues. Since this work is part of a larger project also involving modelling based on short time frames of speech coming from very large speech corpora (over thousand speakers), the present results are intended to be incorporated into the more compre-

hensive common framework. The final speaker profile is planned to rely both on short and long-term feature levels. It is aimed to test its usability within the process of characterisation and recognition of speakers for the needs of forensics and military services.

Acknowledgments

This work is supported from the financial resources for science in the years 2010–2012 as a development project (project no. O R00 0170 12). The author also wishes to thank Prof. Maciej Karpiński for his inspiring co-operation and advice within the present project.

References

Allen, L. Q. (1999). Functions of nonverbal communication in teaching and learning a foreign language. *The French Review, 72*(3), 469–480.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication, 33*(1–2), 5–22.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'FEELTRACE': An instrument for recording perceived emotion in real time. In R. Cowie (Ed.), *Proceedings of the ISCA workshop on speech and emotion: A conceptual framework for research* (pp. 19–24). Belfast: Textflow.

Crystal, D. (1966). The linguistic status of prosodic and paralinguistic features. *Proceedings of the University of Newcastle-upon Tyne Philosophical Society 1*(8), 93–108.

Crystal, D. (1974). Paralinguistics. In T. A. Sebeok (Ed.), *Current trends in linguistics*, Vol. 12 (pp. 265–95). The Hague: Mouton.

Demenko, G., Cecko, R. Szymański, M., Owsianny, M., Francuzik, P., & Lange, M. (2012). Polish speech dictation system as an application of voice interfaces. In A. Dziech & A. Czyżewski (Eds.), *Proceedings of 5th International Conference on Multimedia Communications, Services and Security, Cracow 2012* (pp. 68–76). Springer for Research and Development.

Demenko, G., Grocholewski, S., Klessa, K., & Rau, Z. (2009). Polish language resources for speech technology: JURISDIC LVCSR corpora. In Z. Vetulani (Ed.), *Proceedings of the 4th Language & Technology Conference* (Paper ID: 96). Poznań.

Ethier, N. A. (2010). *Paralinguistic and nonverbal behaviour in social interactions: A lens model perspective*. Doctoral dissertation. University of Waterloo. Retrieved from <http://uwspace.uwaterloo.ca/bitstream/10012/5673/1/Ethier_Nicole.pdf>.

Fischer, V., Diehl, F., Kiessling, A., & Marasek, K. (2000). Specification of databases – Specification of annotation. *SPEECON Deliverable*, D214. Retrieved from <http://www.speechdat.org/speecon/public_docs/D21.zip>.

Geumann, A. (2001). Vocal intensity: Acoustic and articulatory correlates. In B. Maassen, W. Hulstijn, R. Kent, H. Peters & P. van Lieshout (Eds.), *Proceedings of the 4th International Speech Motor Conference* (pp. 70–73). Nijmegen: Uitgeverij Vantilt.

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. Pierrehumbert & C. Shin (Eds.), *Proceeding of Speech Prosody 2010*. Chicago.

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.

Karpinski, M. (2012). The boundaries of language: Exploring paralinguistic features. *Lingua Posnaniensis, 54*(2), 37–54.

Klessa, K., & Demenko, G. (2009). Structure and annotation of Polish LVCSR speech database. In *Proceedings of Interspeech 2009* (pp. 1815–1818). Brighton.

Klessa, K., Wagner, A., Oleśkowicz-Popiel, M., & Karpiński, M. (2013). "Paralingua" – A new speech corpus for the studies of paralinguistic features. In *Proceedings of CILC 2013: 5th international conference on corpus linguistics*. Alicante.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago: University of Chicago.

Liscombe, J. (2007). *Prosody and speaker state: Paralinguistics, pragmatics, and proficiency*. Doctoral dissertation. Columbia University.

Minematsu, N., Asakawa, S., & Hirose, K. (2006). Para-linguistic information represented as distortion of the acoustic universal structure in speech. In *Proceedings of International Conference on Acoustics, Speech, & Signal Processing (ICASSP'2006)*, Vol. 1 (pp. 261–264).

Nolan, F. (2001). Speaker identification evidence: Its forms, limitations, and roles. In *Proceedings of the conference 'Law and Language: Prospect and Retrospect', 1*. Levi. Retrieved from: <http://www.ling.cam.ac.uk/francis/LawLang.doc>.

Rose. P. (2003). The technical comparison of forensic voice samples. Expert evidence 99. In H. Selby & I. Freckelton (Eds.), *Thompson Lawbook Co*. Sydney.

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

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Deviller, L., Müller, C., & Narayanan, S. (2010). The INTERSPEECH 2010 paralinguistic challenge. In T. Kobayashi, K. Hirose & S. Nakamura (Eds.), *Proceedings of Interspeech 2010*. Makuhari.

Schötz, S. (2002). Linguistic & paralinguistic phonetic variation in speaker recognition & text-to-speech synthesis. In *GSLT Papers: Speech Technology 1*. Retrieved from: <http://www.speech.kth.se/~rolf/gslt_papers/SusanneSchotz.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.

Sjölander, K., & Beskow, J. (2000). WaveSurfer – an open source speech tool. In *Proceedings of 6th ICSLP Conference 2000*, Vol. 4 (pp. 464–467). Beijing.

Trager, G. L. (1958). Paralanguage: A first approximation. *Studies in Linguistics, 13*, 1–12.

Trager, G. L. (1961). The Typology of Paralanguage. *Anthropological Linguistics, 3*, 17–21.

von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence unimodal voice recognition. *PLoS Biology, 4*(10), e326.

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A. L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience, 17*(3), 367–376.

Next contribution