Annotation Pro - a new software tool for annotation of linguistic and paralinguistic features

Katarzyna Klessa, Maciej Karpiński, Agnieszka Wagner

Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland
{klessa,maciejk,wagner}@amu.edu.pl

Abstract

This paper describes the design, development and preliminary verification of a new tool created for the purpose of annotation of spoken language recordings. The software extends the potential of a typical multi-layer annotation system with a new component based on the graphical representation of feature space that supports annotation of continuous and non-categorial features. Apart from the annotation options, the program provides a flexible perception experiment framework aimed especially at testing hypotheses related to continuous and non-categorial features.

The tool was initially tested and first applied for the annotation of a speech corpus composed of conversational and emotionally marked speech data within a larger project confessed to speaker characterisation and recognition.

Index Terms: annotation tools, perception based annotation, paralinguistic features, speech prosody

1. Introduction

We understand the process of speech annotation as assigning tags to selected portions of speech signal that may correspond to various units of analysis, from tiny phonetic segments to complex phrases or paratones. The tags usually come from an explicit, closed set like PoS. Still, there are situations that require more flexibility and where operation on fuzzy categories or gradable features is necessary. For example, in the annotation of emotional aspects of speech, there can be some intermediate affective states between extremes, e.g., between joy and saddness. Their number can be arbitrarily assummed or left for annotators to decide. Another problem pertains the fact that some of labels are two- or multidimensional, i.e., their values can be well represented in a multi-dimesional space. Therefore, they may consist of two or more values (tags) that can be placed on separate annotation layers. This somehow corresponds to the conceptual difference between tags and labels proposed by [1].

Many speech annotation programs (see, e.g., [2]) offer great potential but it is sometimes accompanied by less obvious user interface and complex operation. Some of them were primarily conceived for instrumental phonetic analysis and few of them offer direct support for non-categorial or complex-label annotation. Working on the annotation of spontaneous speech corpora on both linguistic and paralinguistic levels, the authors felt an increasing need for a more intuitive software that would support annotation on multiple levels as well as various types of categorial and non-categorial data.

2. Software design and development

2.1. Assumptions and requirements

Paralinguistic features as well as other non-categorial features pose a challenge in the process of speech data annotation – both for software and for human annotators themselves. The way of defining the space for their annotations may strongly influence eventual results. The type of the scale used for a given dimension (linear, logarithmic, etc.) may be also of importance. Paralinguistic features often remain difficult to define in an unambiguous way, in clear and accessible terms. If "verbal" tags are used (e.g., the names of emotional categories, like "disturbed", "angry"), their understanding by annotators may be strongly influenced by everyday usage of such words. Many of these and similar issues may be only partially solved or alleviated, and solutions will be most often context-dependent, designed or tuned for a particular kind of data and specific scientific aims.

The present program is intended for speech annotation for a range of applications, including those strictly technological (e.g., naturally-sounding speech synthesis, automatic speaker and speech recognition) as well as those focused on the psychology of interpersonal communication. Accordingly, the following functionalities and options were considered essential:

- simple and user-friendly interface, easy installation and configuration;
- multi-layer, synchronised annotation with precise boundary placement;
- various, adjustable annotation spaces and scales available as uploadable images;
- the option of using own spaces and scales represented as images;
- use of complex tags (e.g., for two-dimensional features);
- a slot for plugins that would extend the functionality of the program.

Well-organised software that supports annotation of paralinguistic features may also serve as an experimental tool in perception-based studies. While "top-down" approach, starting annotation with pre-defined categories, may keep annotators and resaerchers "blind" to new, undiscovered phenomena, leaving more flexibility to annotators and offering them non-categorial or continous space may bring new

observations to the daylight or just allow for new categorisations to emerge.

2.2. Implementation and architecture

Annotation Pro was created using C# programming language and the Visual Studio programming environment and (in its current form) is designed for Windows operating system.

The main contruction assumption was to create annotation software of general use, applicable for various types of projects involving both annotation of spoken and written (eg. morphological glossing) resources. On the other hand, the architecture was expected to be extensible and flexible in order to enable annotation according to user-specific needs which has been achieved thanks to plugin technology that enables the users to add their own functionality to the program top menu.

The structure of the system has been developed as a multilayer architecture, in which each application tier represents a specific functional layer. The layers of the program are shown in Figure 1.

n-tier architecture
Presentation
Logic
Database
Shared
Plugin

Table 1. *Programming tiers in Annotation Pro*

The **Database** layer is responsible for the process of writing and reading data on the most basic level. Currently, it concerns writing and reading of XML files and dealing with the software's annotation file format ANT which is in fact a ZIP archive containing the packed XML annotation file (see also 3.3 below). Such solution makes it possible to include various types of content inside the ANT file in future.

The **Logic** layer is an intermediate layer representing data in the form of C# objects that can be used by the programmer for operations on objects and collections of object in the application: Layer, Segment, Configuration.

The highest-level layer is the **Presentation** layer. This layer includes controls representing the elements of the software interface: Spectrogram, Waveform, Layer Collection, Input Device. All these components can co-operate automatically thus making it possible for the programmer to create any clone of the application based on *Annotation Pro's* functionality. The controls of the Presentation layers are treated as components joined by a special control – Synchronizer. The Synchronizer is a special object which controls the state of variables whose synchronization is necessary for the consistent functioning of the application.

Plugin – a layer responsible for plugins. The plugin functionality makes it possible for the user to adapt the software to the individual needs of their own project. Including the plugin technology is a natural consequence of the main construction assumption: only general options that are required for most uses are built-in as fixed parts of the software while every functionality that is more project- or user-specific may be accessible via plugin menu. Any user familiar with C# programmimg language can easily create a plugin thus extending the software's functionality (e.g. speech analysis options, automatic feature extraction from the speech signal or time-alignment procedures or any other desired by the user). While initialising, the program scans the *Plugins* folder located in the user's Documents/*Annotation Pro* folder and updates a list of plugins in the *Plugin* menu based on the contents of the folder. The plugins require a standard C# format (*.cs) with an appropriate structure, shown in an example plugin file, also available for the user in the *Annotation Pro* Plugins folder. The plugin file (*.cs) is compiled and launched at runtime. The user can access and use all controls of the interface, to annotation layers, and data.

Shared – a library including support classes for various layers.

3. User interface

3.1. Annotation interface

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 (e.g. Fig. 2).

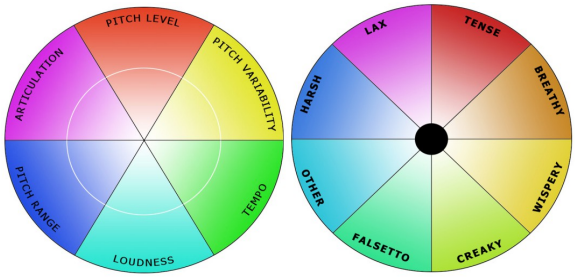

Figure 2: a) and b). Graphical representations used in the description of prosody (a, left) and voice quality (b, right).

Figure 3 (next page) shows the default program interface. The graphic control is visible in the right top corner of the program window. Instead of this particular picture representing two-dimensional space for annotation of emotional states, the user may select another picture (e.g. a *min-max* slider or a set of sliders for perception-based ratings using a continuous scale, etc.). It is also possible to create one's own picture representing any desired two-dimensional feature space. The space 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 both as dots in the picture and as numbers in the related typical annotation layer. While the user clicks on the picture while 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, a collection of coordinates is obtained for which it is then possible to conduct a range of analyses, e.g. cluster analysis (compare also [3] for emotion analysis, and [4] for another examples of another graphic representations used in *Annotation Pro* for both corpus annotation and for conducting perception tests).

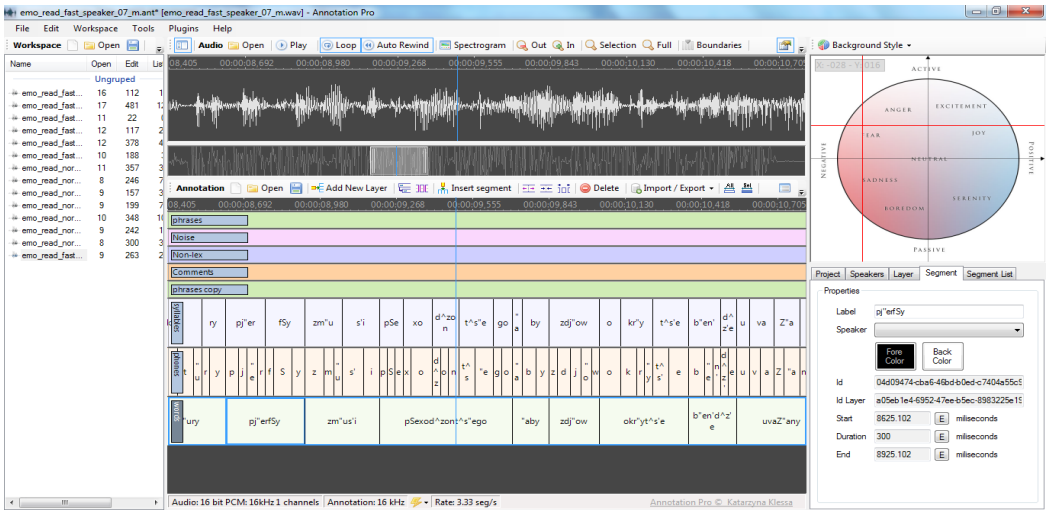

Figure 3: *Annotation Pro* user interface

3.1.1. Annotation of prosodic and paralinguistic features using graphical representation - example

The functionality of the software which enables annotation based on a graphic representation of the feature space was used in the preliminary annotation of perceived prosody and voice quality of emotion portrayals from *Paralingua* database [4] and in the perceptual recognition of speaker state in terms of emotion categories and dimensions [5]. Prosody was annotated in terms of perceived pitch level, pitch variability, tempo, loudness, pitch range and articulation.

The task of the labeler consisted in positioning the cursor in the regions of the circle corresponding to selected prosodic feature and specific intensity of the feature (Fig. 2 a).

In the annotation of perceived voice quality the following labels were taken into account (based on [6]): *lax, tense*, *breathy, whispery, creaky, falsetto, harsh, modal* and *other*. These voice qualities were represented in a circle divided into eight parts with modal voice located in the center and intensity of a given voice quality (to distinguish different levels of e.g. creakiness or breathiness) increasing towards the edge of the circle (Fig. 2 b).

In order to investigate emotional speech production and perception two more graphical representations were created illustrating emotion dimensions of valence and activation (Fig. 4 a), and emotion labels (categories) describing speaker state (Fig. 4 b): *irritated/angry*, *interested/involved*, *proud/satisfied*, *joyful/happy*, *experiencing sensual pleasure*, *bored/weary*, *ashamed/embarrassed*, *in despair/sad*, *anxious/fearful* and *other*. The categorial and dimensional descriptions were based on [7, 8, 9, 10]. In the categorial representation, the twelve emotion labels used in the actor portrayals were collapsed to nine categories (plus *other*), because it was assumed that emotions belonging to the same family, of the same quality and valence, but of a different intensity should be represented together. In the perceptual annotation using the graphical representation (depicted in the Figure 4 b) these differences could be represented 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).

Perception-based annotation of prosody, voice quality and 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.

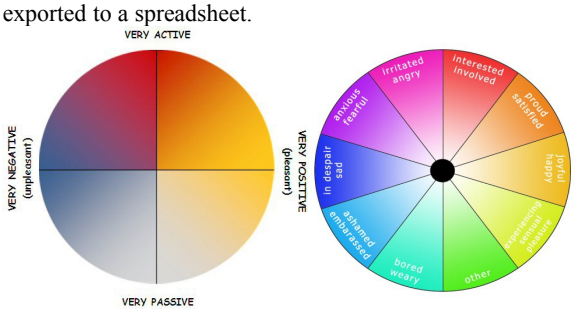

Figure 4 a) and b). Graphical representations used in the classification of emotional speech using valence/activation dimensions (a, left) and emotion labels (b, right).

3.2. Perception test session mode

The annotation interface can also serve as a tool for perception tests as it offers additional options in the *test session* mode. In this mode, the user can use the options for setting-up an experiment. First, it is possible to define options related to participants data management (participant's name or ID, age, gender, region of origin or other features). The perception test set-up is flexible and can be adjusted to particular needs. The experimenter can decide on the number of possible replays of each signal, the order of the signals, the possibility of returning to previous signals after marking the first answer/decision. The original file names can be either displayed or hidden during the test session. The results of the test are written to a CSV file where information about all the actions taken by the subject during the testing session (answers/decisions, opened files, number of listenings, etc.).

3.3. Annotation file format

Annotation Pro annotation (ANT) files are based on the XML format. The format was designed in a way to generalize the

annotation information. Any information narrowing the annotation information to a specific domain or project are introduced by the user via the user interface.

ANT files can store data for annotation using any desired number of annotation layers. The two crucial components of the XML file are **<Layer>** and **.** The first one includes information about annotation layers and the secoind is a universal element that may contain various types of annotation lables (orthographic transcription, tagging of prosody, syntax, discourse markers, paralinguistic annotations, etc.) depending on the user's needs.

Any other relevant information related to the file, speaker, corpus etc. is stored using optional **<Configuration>** elements. This element is of dictionary type, and includes keys and values. The keys should be unique. A number of keys have been reserved for a set of standard properties related to e.g. date of creation, date of modification, version name, title of the project, characteristics of the recording environment, description of background noises characteristic to the project, the name of the collection including the project, the type of corpus, licence, etc.

Annotation Pro also reads XML files created with external tools provided that the format is compatible with ANT. Any information that has not been pre-defined in *Annotation Pro* should be included in the XML file using the <Configuration> elements. *Annotation Pro* 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 *Annotation Pro* on an intermediate, lossless basis.

Apart from the use of the default XML-based annotation files, *Annotation Pro* can import files from the following external formats: *Transcriber*'s TRS [11], and BLF [13], and also from TXT (each verse of the source text file will be imported to a separate segment in the selected annotation layer) and CSV (configurable import, including *Wavesurfer*'s LAB [12]) files.

4. Conclusions

Annotation Pro has already been employed for transcription and annotation of speech data in several research projects. Its applications included the analysis of perception and production of emotional speech. Presently, the technique of emotional speech analysis based on the functionalities provided by *Annotation Pro* is used in a larger-scale study on cross-linguistic perception of vocal communication of emotions. The tool is also used for transcription and annotation of corpora of lesser used languages requiring annotation with non-standard types of font family and morphological glossing. Annotators from both fields confirm that the software's principle is clear and user interface is easy to master while still retaining much flexibility. Although it is clear that all the issues mentioned in 1 and 2.1 cannot be solved by this piece of software, it offers some advantages over other available solutions.

Annotation Pro will be further tested and extended with new plugins. Import/export options will be elaborated in order to accept the data and metadata from other programs (e.g. *Praat* [14], *SPPAS* [15], *TGA* [16]). This is expected to allow to use *Annotation Pro* as a complementary tool for specific purposes and to exchange and integrate data with no information loss. In order to provide higher level of interoperability of the software, it is currently considered to develop a new edition of *Annotation Pro* using *Mono* software platform, thus enabling the use of the program under operating systems other than MS Windows. As for the interface, among options under consideration, there is also video annotation and recording of other types of multiple-speaker data on separate layers.

Annotation Pro is freely available for research purposes from: annotationpro.org (contact e-mail: klessa@amu.edu.pl).

5. Acknowledment

Annotation Pro is developed based on the experiences of the authors gained during the work on an earlier tool named *Annotation System* implemented within project no. **O R00 0170 12** supported from the financial resources for science in the years 2010–2012 as a development project.

6. References

[1] Popescu-Belis, A., "Dialogue Acts: One or More Dimensions?" *ISSCO Working Paper*, 62 – 29th November 2005.

[2] Garg, S., Martinovski, B., Robinson, S., Stephan, J., Tetreault, J., Traum, D.R., *Evaluation of Transcription and Annotation tools for a Multi -modal, Multi-party dialogue corpus.* Southern California, Inst. For Creative Technologies, 2004.

[3] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M. "FEELTRACE: An instrument for recording perceived emotion in real time". *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[4] Klessa, K., Wagner, A., Oleśkiewicz-Popiel, M. "Using *Paralingua* database for investigation of affective states and paralinguistic features". *Speech and Lang. Technology. 14/15,* to be published.

[5] Wagner, A. "Emotional speech production and perception: A framework of analysis". *Speech and Lang. Technology, vol. 14/15*, to be published.

[6] Laver, J. *The phonetic description of voice quality.* Cambridge Studies in Linguistics London, 31, 1-186, 1980.

[7] Russell, J. A. "A circumplex model of affect". *Journal of personality and social psychology, 39(6),* 1161, 1980.

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

[9] Laukka, P. *Vocal expression of emotion: discrete-emotions and dimensional accounts* Phd dissertation, Uppsala Univ., 2004.

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

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

[12] Sjölander, K., and Beskow, J. "WaveSurfer – an Open Source Speech Tool". *Proceedings of 6th ICSLP Conference 2000, Vol. 4* (pp. 464–467). Beijing, 2000.

[13] Breuer, S., & Hess, W. "The Bonn open synthesis system 3". *International Journal of Speech Technology, 13(2)*,75-84, 2010.

[14] Paul Boersma & David Weenink (2009). "Praat: doing phonetics by computer (Version 5.1.05)" [Computer program]. Available: http://www.praat.org/

[15] Bigi, B. "SPPAS: a tool for the phonetic segmentation of speech". *Language Resource and Evaluation Conference*, Istanbul, Turkey, 2012.

[16] Gibbon, D. "TGA: a web tool for Time Group Analysis". *Proc. of the Tools and Resources for the Analysis of Speech Prosody* (TRASP) Workshop, Aix en Provence, August 2013, to be published.