NanoTrans – Editor for Orthographic and Phonetic Transcriptions

Ladislav Seps, Petr Cerva, Jan Silovsky

***Abstract*— NanoTrans is an application for creating and editing orthographic (textual) and phonetic transcriptions of audio and video files. It was created with the focus on easy-of-use and quick-to-learn user interface and it is indent for the users without any previous experience in linguistics or speech processing. It allows for playing the video track and comes with a simple XML-based structure for the transcription files and with features designed to enable quick and easy editing making. The second but not less important use of NanoTrans is reviewing and presenting outputs of our speech recognition system. Interaction with this system and with other tools is achieved by the flexible file import and export plug-in system. Other notable features are provided, like e.g. automatic update system, support for specialized peripherals including the Transcription foot pedal, easily localizable user interface and wide audio format support.**

***Keywords*—orthographic transcription, phonetic transcription, time aligned text, speech processing.**

I. Introduction

The automatic speech recognition (ASR) technology has attracted a lot of attention over the past decades. Various systems have been developed for voice control and dictation, spoken document transcription or key-word spotting. The core of the state-of-the-art ASR systems is formed by a speech recognition engine which employs hidden Markov models (HMMs) and n-grams for acoustic and language modeling respectively.

The HMMs are usually trained using large speech databases containing recordings of hundreds or even thousands of persons. Their total duration can be higher than one thousand hours. This huge amount of different data allows for training robust speaker independent models. Unfortunately, all the training utterances have to be phonetically annotated before training can be conducted. Similarly, orthographic transcriptions of some other data can be used a) to boost the performance of the given n-gram language model or b) for measuring the recognition accuracy of the resulting system.

It is clear that mentioned enormous amounts of data can be efficiently annotated just with the help of a sophisticated software tool. Therefore, several tools for making transcriptions of media files have recently been developed, like e.g. Transcriber[1], TranscriberAG [2], ANVIL [3], ELAN[4] and others.

One of our current projects aims at transcription, indexing and accessing recordings of academic lectures [5]. For the development of an ASR system for lecture transcription, large amount of lecture recordings have to be annotated to improve acoustic and language modeling accuracy. Therefore, an effective annotation tool was needed. Our main requirements were a) the possibility of playing video and b) user-friendly interface which would be not suitable just for users with prior experience in linguistics or speech processing.

Of course, we consider using some of the existing tools at first, but no solution adapted exactly to our needs was available at the time of our choice. From previously mentioned applications, Transcriber was the closest to our requirements, but it did not have the support for playing videos. This feature was added later in TranscriberAG, but not in a very user-friendly way (floating windows). Moreover, the support for the multiple layers in TranscriberAG adds unnecessary features for our use. Further, the ANVIL software was ruled out because of its approach for displaying the transcription. While Transcriber employs interface with the "wall of text", where the transcription is well readable as it occupies most of the screen, the ANVIL`s interface displays each type of transcript just in one long line of text. Moreover, this line is under the window showing the sound content of the recording, which occupies the biggest part of the screen. The last mentioned ELAN tool is intended particularly for experienced users as it relies on keyboard shortcuts. The work without the knowledge of these shortcuts can be considered uncomfortable and ineffective. There are also a lot of features and different views that can be confusing when the given user does not need them.

Finally, given all the previously mentioned reasons, we chose the option to develop a new tool called NanoTrans. This application and some aspect of its use are detailed in the following sections.

II. Main features of NanoTrans

NanoTrans comes with a simple and user-friendly interface allowing for video playing, where everything can be controlled by a mouse. Moreover, its look and functions are designed to be similar to widely used applications, like e.g. word processors, all in the user's native language.

The next important feature of NanoTrans is that it is distributed with the large number of integrated multi-media codecs to avoid manual conversions. It can also be automatically updated over Internet.

Manuscript received February 11, 2013. The research described in this paper was supported by the Technology Agency of the Czech Republic (project no. TA01011142) and by the Student Grant Scheme (SGS) at the Technical University of Liberec.

Authors are with The Institute of Information Technology and Electronics, Technical University of Liberec, Liberec Czech Republic (phone: +420 485353037 e-mail: ladislav.seps@tul.cz).

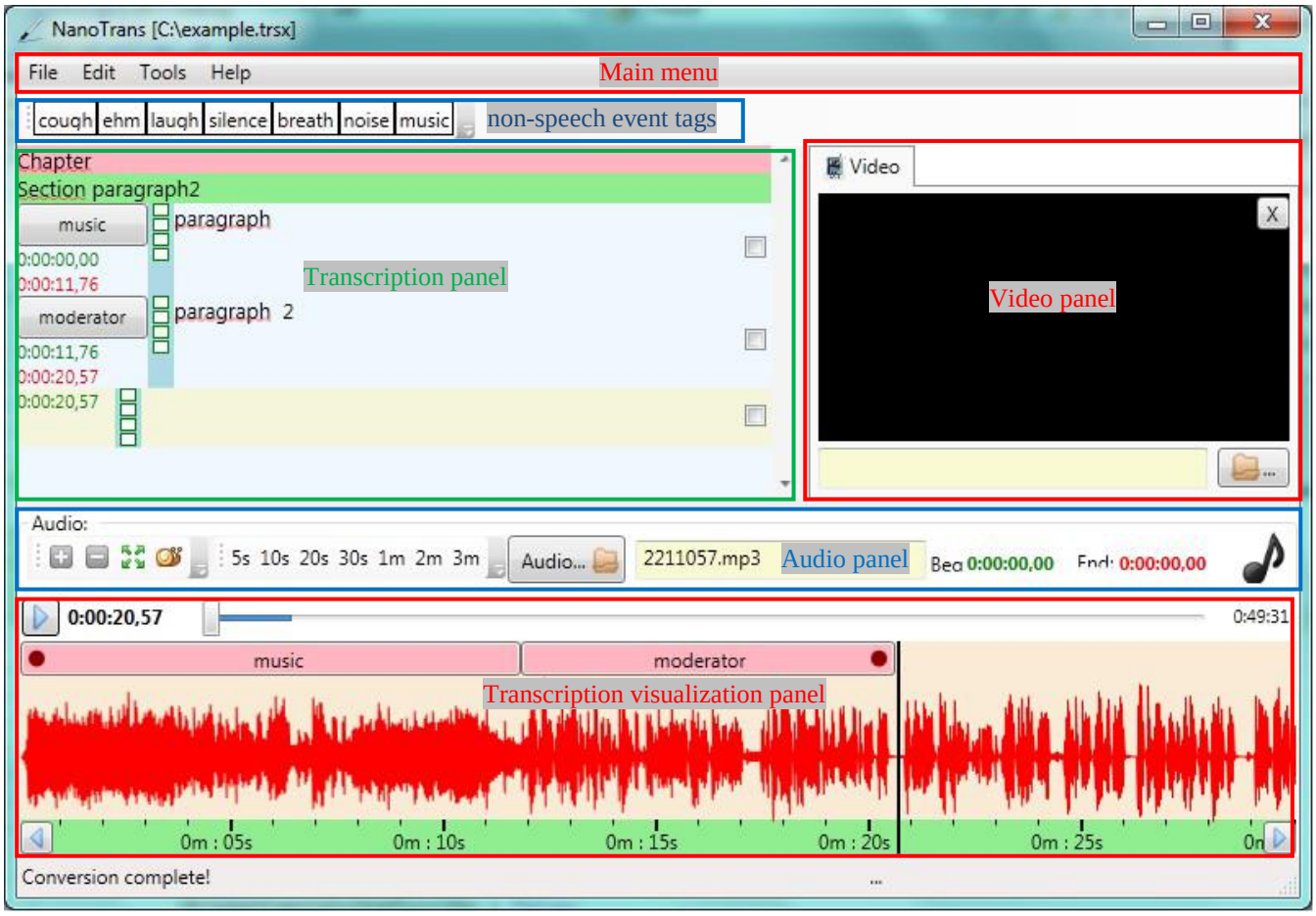

Fig. 1 NanoTrans user interface

To simplify interaction with our recognition system and other tools, import and export plug-in system was implemented. The support the development of plug-ins in a wide number of programming languages, every plug-in can be represented by any executable file or script that accepts command line parameters.

NanoTrans takes advantage of XML-based file format as text or table formats (like CSV - comma separated values) have many problems related to character encoding and whitespace handling that can be different across applications. This file format is related to the internal structure of NanoTrans, which allows for different ways of enumerating through the transcription.

NanoTrans can also be used as a library and in combination with the LINQ (language integrated query) technology, available in .NET languages. Therefore, complex searching and filtering operations can be made with just few lines of code and they can be executed on archives containing thousands of files.

At last but not at least, the use NanoTrans is not limited only to creating and editing transcriptions. It can also be employed for reviewing and presenting outputs of our ASR system.

III. Internal File Structure

Every orthographic transcription of an input recording can theoretically be represented as a plain text (or text with some time related marks). However, it is clear that transcriptions should have a fixed structure as one recording can contain multiple unassociated parts and it should be possible to mark them, separate them and give them some captions.

Here, we took inspiration in Transcriber[1],[2]. Hence NanoTrans handles two layers, called Chapter and Section, for structuring the transcription. Every transcription contains one or more Chapters and Chapter is composed of one or more Sections. Further, Sections may include Paragraphs. Every Paragraph can then contain one utterance (but it is not enforced).

Every Paragraph contains one hidden layer that can't be directly edited by the user. This layer is formed by Phrases. In NanoTrans, the Phrase is the smallest part of the transcription which also contains timestamps of start and end. These timestamps allow that Phrases can be highlighted during playback. In addition, Phrases can be generated by an external forced-alignment tool. The structure of transcriptions reflects in the structure of output files.

As mentioned, the file format depicted in Fig. 2 is XML-based and it evolved from a simple serialization of the inner memory structure. The phonetic transcription of the given phrase is not shown here, but it can be stored as the value of the "fon" attribute. Please note that storing phonetic transcription of phrases is optional. Beside transcription, the file can also contain some other items. For example the "meta" item (at the beginning of a file) can be used to store

any content. This item can't be modified or deleted when the file is processed in NanoTrans. Next, not only the shown attributes, but also the custom ones can be added into the tags. Similarly as meta items, they also cannot be modified or deleted. (except the case when the tag containing them is deleted).

The support of additional items allows for adding task-specific information into the structure. For example, an ID attribute can be added to each tag. After editing by NanoTrans simple by comparing tags with same ID can show what tags was modified, added or deleted.

NanoTrans also supports shortened file format. Within this format, all tag names and attributes are composed just of one or two characters. The format was created to save storage capacity, make loading and saving files faster and increase readability of the phoneme aligned transcriptions (every phrase contains only one phoneme). The size of the shortened files can be as low as 20% of the size of the basic format, which we also call strict format, when dealing with phoneme aligned transcriptions.

```
<?xml version="1.0" encoding="utf-8"?>
<transcription mediaURI="AKSenatTUL-201
    20313.mp4" version="2.0" style="strict">
  <meta>
    <broadcastTime value="1970-01-01T01:00:00"/>
    <groups/>
  </meta>
  <chapter name="recording">
    <section name="recording">
      <paragraph begin="0" end="231770" speakerId="1">
        <phrase begin="0" end="620">welcome</phrase>
      </paragraph>
    </section>
  </chapter>
  <speakers>
    <speaker surname="None" id="1" sex="X"/>
  </speakers>
```

D. Video Panel

The video panel allows for playing the video track of the recording which is being annotated. Of course, video playback is synchronized with audio and it makes use of Window Media Player 10. Therefore, the supported formats are limited to the codecs present in the system.

The feature of video playing is used particularly in situations, when the resulting transcript has to contain speaker labels with corresponding timestamps. The video panel can be hidden when any video track is not available.

E. Audio Panel

Similarly as the previous panel, the audio panel can be used to load and play the audio stream that is being transcribed. Although NanoTrans internally uses sampling frequency of 16 kHz and resolution of 16bits per sample, it can load any audio or video file as it employs the FFmpeg [6] for conversion. (The given version of FFmpeg can be replaced by the user).

The panel also allows the user to change the playback speed. This option is adjustable from the application settings as well as the properties of the transcription visualization panel.

F. Transcription Visualization Panel

The Transcription visualization panel is located at the bottom of the NanoTran`s window. It shows the timeline, waveform of the audio signal, appropriate paragraphs (with the name of the speaker) and the current playback position (the black vertical line). The position of every paragraph, the length of the paragraph and the playback position can be modified here.

G. Phonetic Transcription Panel

This panel is simple as it contains just the text editing box with the phonetic transcription of the selected Paragraph. When playing, the current phrase is highlighted, similarly as the orthographic transcription.

VI. Orthographic Transcription

The transcription panel is the main part of NanoTrans as it is the place where the user does most of the work. It also represents the most complicated part.

It contains three types of elements:

- Chapters – with pink background
- Sections – with greenish background
- Paragraphs – without any special background

Each element is represented by one horizontal block in the transcription visualization panel. Text editor component used is AvalonEdit [7].

AvalonEdit belongs to a group of text editors called code editors that serve namely for text highlighting in programming IDEs. They are similar to rich text editors known from word processors (e.g. Microsoft Word), but they lack the support for advanced formatting. For example, the do not support placing independent blocks (e.g. images) inside the text or using more than one font size. On the other hand, they excel in highlighting parts of the text by different means, like e.g. text color, boldness, background color or underlining.

NanoTrans takes an advantage of these features for spell-checking, correction suggestions and phrase highlighting during playback.

A. Speaker Selection

The name of each speaker is displayed on the button in the top-left of the Paragraph (it is hidden in the subsequent paragraphs with the same speaker). It can be changed by clicking on the button or by the right-click menu.

B. Paragraph Attributes

To mark the characteristics of an entire paragraph, attributes are used. These are displayed as small rectangles with green border on the blue column between time marks and the text of the Paragraph. If the attribute is selected, it has a different color than white. Currently, there are only four following attributes, but others can easily be added:

- Narrowband – for the part of the recording with a lower sampling frequency like a telephone call
- Background speech
- Background noise
- Junk – for incomprehensible paragraphs

C. Spell Checking

All text elements in the Transcription panel supports spell-checking. The words that are not recognized by the spellchecker are highlighted by red wavy line (as can be seen in Fig. 1 and Fig. 4). The tool NHunspell [8] is used as the spell-checking engine. Custom dictionaries for the spell-checking engine can be installed from the application settings. Direct import of OpenOffice and LibreOffice dictionary plug-ins is also supported.

D. Correction Suggestions

NanoTrans can suggest corrections for words in the edited transcription. Unlike most of text editors, NanoTrans don't suggest corrections based on grammatical rules, but using the big list of regular expressions. These are created specifically to detect and suggest corrections to most common errors created by our ASR systems. This list can be modified for specific tasks.

Words that have one or more variants suggested by the correction system are highlighted by green wavy line. Selection tool-tip (see in Fig. 5) with variants for selected text can be opened by keyboard shortcut or middle mouse button.

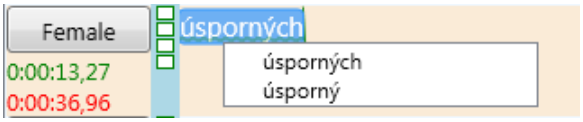

Fig. 5: Correction tool-tip

VII. Phrase Highlighting During Playback

One of the most used features of NanoTrans, which simplifies the work, is the highlighting of words during playback. For this purpose, the existing transcription must be preprocessed by an external tool at first. The software we

utilize is described in [9] and it is based on forced alignment.

In the opposite case, when the transcription of the recording does not exist, it is necessary to process the recording using a speech recognizer at first. The resulting recognized transcription is then just corrected. This approach saves time needed for annotation as the correction of the time-aligned automatic transcript is much faster the standard process of manual transcription.

Technically the minimum length of a phrase can be 0 characters (highlights line between two characters) but it is not recommended to have Phrases shorter than words, because it is affecting the file size of the transcription and it isn't easy to follow that detailed transcription when editing.

While playing, the current phrase is highlighted by the green background (Fig. 6)

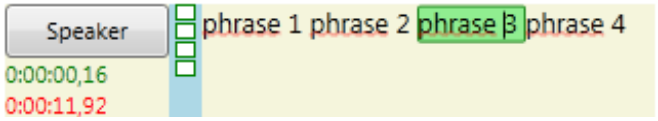

Fig. 6: Highlighted phrase

VIII. Installer and updater

NanoTrans is distributed as an executable installer file. The installer runs just on Microsoft Windows as mentioned and it downloads and installs automatically .NET framework 4.5.

The updater checks the update server each time, when NanoTrans is launched. If a new update is available, all NanoTrans instances are closed at first and the update is downloaded and installed automatically after that.

IX. Specialized Peripherals Support

NanoTrans supports external devices like foot pedals. The support is provided by running an external executable and reading the events from its standard output. The support for some types of USB HID 3-pedal Foot pedals is also included.

Other control devices can be supported by replacing the included exe file by the executable distributed with the given device.

X. Conclusion

We have presented NanoTrans, a tool developed with the main aim to enable effective and user-friendly phonetic as well as orthographic annotations of speech databases. It comes with several features, like e.g. automatic updater or video playing, that are not common in most existing applications for transcription of audio recordings.

The current version of NanoTrans is used on daily basis by members of our laboratory as well as by external staffs who are not experienced in the speech processing technology.

The NanoTrans`s user Interface is easily localizable by creating language specific resource files, so there are no obstacles for users from different countries. The source code is undergoing heavy refactoring in order to be usable for non-Czech users. A public version of NanoTrans will be released in the near future. We also plan to add some additional features like import and export plug-ins for the existing transcription-file format.

The main feature planned in the long term is the integration of an ASR system into NanoTrans. This should allow for automatic generation of phonetic transcriptions or phrase highlighting based on forced-alignment. The integration will make the use of an external (and replaceable) tool as integrating an ASR system directly into NanoTrans would unnecessarily increase memory usage and the size of distributed application.

References

[1] Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. In: *Speech Communication Special Issue on Speech Annotation and Corpus Tools*", Vol. 33, No. 1-2. 2000.

[2] Geoffrois, E., Barras, C., Bird, S., Wu, Z.: "Transcribing with Annotation Graphs". In: *Second International Conf. On Language Resources and Evaluation (LREC)*, pp. 1517–1521. 2000.

[3] Kipp, M., Anvil – "A Generic Annotation Tool for Multimodal Dialogue". Proceedings of the *7th European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1367-1370. 2001.

[4] Brugman, H., Russel, A., "Annotating Multimedia/ Multimodal resources with ELAN". In: Proceedings of *LREC 2004, Fourth International Confe.rence on Language Resources and Evaluation*. 2004

[5] Cerva, Petr and Silovsky, Jan and Zdansky, Jindrich and Smola, Ondrej and Blavka, Karel and Palecek, Karel and Nouza, Jan and Malek, Jiri, "Browsing, Indexing and Automatic Transcription of Lectures for Distance Learning". In: proc. *Of MMSP 12*, pp. 198 – 202. Banff, AB, Canada 2012

[6] FFmpeg project (2013, February) Available: http://www.ffmpeg.org

[7] SharpDevelop project (2013, February) Available: https://github.com/icsharpcode/SharpDevelop/wiki/AvalonEdit

[8] NHunspell project (2013, February) Available: http://nhunspell.sourceforge.net/

[9] Bohac, M., Blavka, K.: "Automatic Segmentation and Annotation of Audio Archive Documents". In: proc. Of *ECMS 11*, pp. 61 – 66. Liberec 2011.