**Abstract**

MTRANS, a freely available tool for annotating multi-channel speech is presented. This software tool is designed to provide visual and aural display flexibility required for transcribing multi-party conversations; in particular, it eases the analysis of speech overlaps by overlaying waveforms and spectrograms (with controllable transparency), and the mapping from media channels to annotation tiers by allowing arbitrary associations between them. MTRANS supports interoperability with other tools via the Open Sound Control protocol.

**Index Terms:** speech tools, multi-channel speech annotation, conversation analysis.

1. **Introduction**

The analysis of multi-channel speech material such as that available in the ICSI [1] and AMI [2] corpora, or recordings which result from exploring the effect of background conversations on foreground conversations [3] often requires the transcription and annotation of recordings of more than two talkers, here denoted ‘multi–channel’ conversations.

While some tools can handle multi-channel media, they rarely possess the flexibility in audio output and visual display required when transcribing and annotating multi-channel conversational speech. Overlaps, frequent in natural conversations, are particularly interesting, and their analysis would benefit from support for visual signal separation in both time and frequency domains. A further issue is the mapping from annotation tiers to channels, where typical one-to-one associations are too restrictive in cases where the tier corresponds to events, such as inter-turn gaps, which are only meaningful with reference to more than one channel. To support rapid and flexible annotation of multi-channel speech with many annotation levels, we have developed MTRANS, a free and open source tool available from www.laslab.org/tools/mtrans.

2. **Tools and Frameworks**

Tools such as Praat [4] (probably the most widely used software for speech analysis), Speech Filing System [5], or arguably MATLAB® [6] (via its Signal Processing Toolbox), offer solutions to speech-signal analysis including modules for capturing, displaying, editing, annotating, and analysing media. These powerful tools also benefit from user-contributed plug-ins and libraries such as the ‘voicebox’ collection of MATLAB functions [7]. In some cases, these applications are forced to compromise between flexibility and generality yielding them somewhat cumbersome for tasks required in the analysis of multi-channel speech.

Numerous software applications for editing signals can be used for conversational analysis. A complete review of these applications is beyond the scope of this article: only the most relevant tools will be discussed. Anvil [8], a video-oriented annotation tool, integrates multimodal information including motion capture data. The CLAN programs [9] adhere to the CHAT transcription convention, and are useful for frequency counts, co-occurrence and interactional analyses. ELAN [10] is a multimedia (video and audio) annotation tool supporting multiple channel video. The EMU Speech Database System [11] allows signal editing, hierarchical annotation handling and automated signal modification via batch processes. Being database oriented, large corpora can be analysed in detail and through its interface with it is possible to perform sophisticated statistical analyses and create figures in an integrated environment. EXMARALDA [12], through its Partitur-Editor allows multiple level annotation of single audio and video files. Folker [13] shares many similarities with EXMARALDA’s Partitur-Editor. LAB-CAT (a.k.a. ONZE Miner) [14], offers monochromatic spectrogram display (indirectly) for single channel analysis. The commercial application Transana [15] is another video-annotating tool which is simple to use and works with single sources. Transcriber [16], designed specifically for the annotation of broadcast news recordings, supports multiple media channels via the ‘channelTrans’ extension. Note that there is a homonymous version of Transcriber currently not maintained, which allows monophonic transcription of audio files, with spectrogram, signal, and energy plots. Wavesurfer [17] supports multi-channel audio analysis and spectrogram view; its development, stopped for about six years, has been recently resumed. This tool is frequently used, e.g., EMU signal editing functionality is built upon Wavesurfer.

The aforementioned applications provide the user with common features such as time-alignment, basic signal processing (intensity control, lateralisation, etc.), basic editing (cut, copy, and paste), interfaces to other programs and several import/export formats. Table 1 summarises some of these software features.

| **Handling multi-channel audio is achieved in some applications (e.g., EMU and Praat) by synchronously rendering several media files in independent windows. Such a mechanism creates flexibility and scalability problems.** |

---

1. A representative list of annotation tools can be found at annotation.exmaralda.org/index.php/Linguistic_Annotation
2. www.icsi.berkeley.edu/Speech/mr/channeltrans
Figure 1: MTRANS interface. This example has four audio channels of which two spectrograms and 14 tiers are shown.

Table 1: Some applications related to MTRANS (all are freely available, multi-platform, and feature waveform representation.)

<table>
<thead>
<tr>
<th>Specification</th>
<th>other media</th>
<th>multi-channel</th>
<th>open source</th>
<th>scriptable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anvil</td>
<td>No Video, motion</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>CLAN</td>
<td>No Video</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>ELAN</td>
<td>No Video</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Emu</td>
<td>Yes jaw, tongue</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>EXMAREALDA</td>
<td>No Video</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Folker</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>LaBB-CAT</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>MTRANS</td>
<td>Yes Partial</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Praat</td>
<td>Yes No Partial</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Fps</td>
<td>Yes No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Transcriber</td>
<td>No No Yes Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Wavesurfer</td>
<td>Yes No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

3. MTRANS philosophy

One of the goals in creating MTRANS was to cover the gap left by existing applications in the analysis of natural, real-world, multi-party conversations (rich in overlaps) by allowing the visual inspection of overlaid signals in different representations (spectrographic and waveform). In its design, the following features were considered important:

Ease of experimentation: As an example, spectrogram parameters (e.g., maximum frequency, frequency resolution, energy range, and individual signal transparency) can be adjusted continuously in real-time. This fine control facilitates visual speech analysis which often demands tuning of display parameters dependent on the speaker, task and segment of speech material currently under analysis.

Flexibility: Arbitrary tier grouping and ordering allows a logical relationship between tier type and channels (e.g., overlaps between channels) and the determination of which signals are required for playback. Tier type and tier association provides useful semantic information in subsequent processing of annotation files.

Information filtering: To better focus on a particular task, MTRANS allows the user to display/hide information such as individual tiers, tier groups, waveforms, annotation, and spectrograms. Aurally, one can lateralize each audio channel as well as increase its relative sound intensity (reflected in its spectrogram transparency).

Interoperability: MTRANS specialises in creating annotations from multiple audio channels but it is capable of sending real-time transport information (e.g., playback starting point, segment duration, and speed rate) to other applications via the Open Sound Control (OSC) protocol. Provided that OSC-enabled clients exist, there is no need to switch between tools for a particular task nor to overload MTRANS with peripheral features used infrequently. The use of OSC allows MTRANS to focus on its core functionality while recruiting other specialised tools for tasks such as video playback (see later) with no additional coding.

Speed and reactiveness: Regardless of media length, MTRANS is designed to be fast. By implementing resizable panels, users can examine speech in more detail, or conveniently shrink them for improved performance. Since only the visible spectrogram pixels are computed at given time, reducing the size of this panel increases overall performance.

\[4\text{opensoundcontrol.org} \]
4. Main Features

MTRANS runs on platforms supporting MATLAB (such as Windows, Mac OS X and Linux) and it is extensible via MATLAB scripts. It consists of two windows: the main display and a control & search panel. Table 2 summarises the main features of MTRANS.

<table>
<thead>
<tr>
<th>Browsing</th>
<th>Audio channel lateralization and volume control, transparency control, several levels of zooming, resizable panels, overlaid hue-encoded spectrograms and waveforms, text and regular expression search, help via tool-tips.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interoperation</td>
<td>Import/Export tiers to Praat or HTK annotation files. Real-time inter-operation with OSC-enabled software. Supports tier lexica to ensure uniformity amongst transcribers. Automatic generation of overlap and turn tiers as the starting point for manual correction.</td>
</tr>
<tr>
<td>Navigation</td>
<td>Several ways to move between annotations (skip to the previous/next non-empty annotation, or with the same label, etc.), access annotations through search results.</td>
</tr>
<tr>
<td>Playback speed</td>
<td>Audio streams are processed with a vocoder to control the reproduction speed with no significant formant artefacts.</td>
</tr>
<tr>
<td>Real-time</td>
<td>Spectrogram and waveforms are rendered online according to the current parameters, e.g., panel size, gain, transparency, zoom, frequency resolution.</td>
</tr>
</tbody>
</table>

Table 2: MTRANS features

4.1. Main Window

The resizable panels inside the main window (shown in Fig. 1) correspond to spectrogram, waveform, and tier view of the multiple channels to analyse. Hue-encoded spectrograms, with controlled transparency, facilitate visual source separation at overlaps as shown in Fig. 2. Identifying crosstalk between audio channels (common even when the audio is captured using close-talking microphones) improves the analysis quality, and controlling each channel transparency favours crosstalk detection.

Annotations can be created, deleted, or edited in place using standard GUI conventions; their contents can be typed in or selected from user-defined lexica specific to each tier type. In fast annotation mode, a single click inserts annotation boundaries and, optionally, a default symbol.

Tiers map to multiple or individual media channels, e.g., in Fig. 1, overlap tiers are associated with two audio channels. By clicking on a tier one can add, remove, and edit it. Support for large numbers of tiers is achieved via grouping, flexible ordering, hiding/display, and colouring options.

4.2. Control and Search Window

The control tab (highlighted in left window in Fig. 1) provides a mechanism for selecting what and how to display. The top panel allows online adjustments to, for example, frequency resolution, upper frequency limit, and playback speed. These parameters can be adjusted to permit visual analysis of formants and $f_0$ contours. Through the middle panel users can select what information is displayed for each audio channel including waveform, annotation, spectrogram, etc. Deselecting unwanted in-

4.3. OSC functionality

To demonstrate the benefits of this feature, a simple four-channel video player (shown in Fig. 4) was created. This OSC-client performs synchronised video playback corresponding to the annotation currently being played in MTRANS. For instance, the application can reproduce the video segment at the playback speed specified in MTRANS.

MTRANS users can specify server, port, and OSC-bundling (by default localhost, 2012, and individual messages, respectively). If media and OSC-applications are available, one could for example enrich the analysis of conversations by also looking at participants’ video, motion capture, articulatory settings, as well as other synchronous data such as EEG, etc., without necessarily harming each individual application’s performance and

Figure 2: Hue-encoded spectrograms facilitate visual source separation at overlaps and $f_0$ contour visualisation. This feature is especially useful in combination with transparency control in the analysis of overlaps when the signals are very similar.

formation improves performance and usage of the visual space. Display control of tiers and tier types is possible via the bottom panel.

Figure 3: Search & Browse window (fragment). The resizable bottom panel shows the result tree.

Through the Search tab (shown in Fig. 3), users can navigate a set of retrieved annotations by freely selecting them from the result tree or consecutively by using arrow-keys.

3239
crucially without additional coding.

5. Experience with MTRANS

The need for a detailed analysis of multi-channel speech data motivated the development of MTRANS. Speech material came from an experiment on speaking in the presence of competing speech. MTRANS was used to annotate pairs of simultaneous conversations recorded in three sessions of 20 minutes each [18]. Five channels of synchronous audio data (four close-talking microphones and an omnidirectional table microphone) were available during annotation. Annotation involved more than 20 tiers grouped into seven tier types (session structure, speech/non-speech, orthography, interactional events, inter-turn gaps, turn-types, miscellaneous). More than 7000 individual annotation symbols were created, including nearly 1300 events (dysfluencies, alignments, etc.) and more than 2000 turn components (back-channels, interruptions, smooth change, etc.).

6. Further developments

MTRANS currently saves transcriptions in plain text files for ease of use in subsequent text processing pipelines, but they could be converted to XML files or stored in a database. Retrieving corpora from central databases via a database connection is also an attractive feature to implement given the current trends in data storage, retrieval and shared documents in the speech community.

Much of the information necessary to generate a conversation analysis style of transcription (e.g., following Jefferson’s conventions [19]) is available in MTRANS. This output format will be automatically generated in future versions.

EFX commands can be directly inserted in annotations to produce the most common diacritics and digraphs used with the Latin alphabet, Greek letters, and some phonetic symbols. A full support of other languages and IPA symbols would be available via Unicode encoding (UTF-8) but is yet to be fully implemented in MATLAB. The current version of MTRANS is a hybrid of MATLAB and Java code. Full migration to Java is planned for future versions.

So far, MTRANS has OSC server capabilities. Implementing client capabilities is currently under consideration.

Acknowledgements. The authors thank EU Future and Emerging Technology (FET-OPEN) Project LISTA (The Listening Talker) and EU Marie Curie Research Training Network S2S (Sound to Sense) for support and encouragement.

7. References