Some Aspects of Annotation of Sound Data within the Framework of the "Multimedia Language Documentation and Language Research Laboratory (MLL)"; a Research Project for a Scientific Multimedia Database

Ralf Vollmann#, Werner A. Deutsch*, Andreas Koechert+, Sylvia Moosmüller*, Anton Noll*, Simone Pribbenow%, Joachim Schalthöfer%

#Inst. of Linguistics, University of Graz,
*Acoustics Research Institute, AAS, Vienna,
+ Department of Mathematics and Computer Science, University of Bremen,
and Institute of Amerindian Languages and Cultures, University of Hamburg,
% Department of Mathematics and Computer Science, University of Bremen

Abstract

This paper deals with some of the major aspects of sound annotation within the framework of the MLL, as it will be provided by the S_Tools acoustic phonetic work stations. Within this system, sound data can be referenced via so-called sound file directories which contain time links and additional information. These data are created interactively in a fully-fledged phonetic working environment and can be processed by external database systems. Additional problems of digital sound archives are mentioned.

Introduction

The *Multimedia Language Documentation and Language Research Laboratory (MLL)*, a research project proposed to the Volkswagen Foundation within the framework of the programme *"Documentation of Endangered Languages"*,[1] will develop an archive system which is capable of managing language documentations whose data will consist of texts, pictures, audio and video recordings as well as an extensive amount of secondary data. This archive will be made accessible in an open, 'virtual lab' (VIRTLAB) by means of the internet, for the immediate use of researchers in different fields of study. It will make data from different media equally available and easily applicable by offering a systematic range of applications and tools, and it will facilitate the evolution of special application programmes by external users.

The data collected during the linguistic studies will contain written texts as well as speech recordings (audio data), photographs (pictures) and video recordings (video data). The processing/editing procedure will add to these *raw data* other forms of information — transcriptions, phonetic systems, dictionaries and other texts (e.g. translations). Consequently, the archive to be designed must be able to incorporate and deal with those different media.

The design of such a database poses a number of challenges both to hardware and software requirements; the current contribution will deal with the annotation of sound data.

[1] The aim of the MLL is the documentation of minority languages threatened from extinction, but its functionality can be seen in a wider range of possible applications. The originality of its approach lies in the multimedia representation of the data collected for the scientific use of researchers in the fields of linguistics, ethnology, religious studies, ecology etc.. For this purpose, the MLL will develop a multimedia database, grounded on the data and the information on the particular languages in question gathered by different groups of linguists.

Data formats

Based on experiences from earlier work (cf. Deutsch et al. 1998, Vollmann 1998) such as various sound databases for speech and music (cf. the HARMONICA project), the MLL is designed as an *open* (structurally dynamic) database system combining different kinds of digital data on the user level. All kinds of data will be stored in files of their respective (standard) formats, avoiding compression utilities which are lossy coders.

On the other hand, users may prefer to be faced with a multimedia data format which combines text, picture, sound and video data in a uniform representation. This *multimedia representation format* of VIRTLAB will not be explained here.

In order to be able to manage diverse data from different languages, a universal classification and a format of representation independent of the different schools of theory seems necessary or tempting (cf. Zaefferer 1995). By contributing their information, the linguists involved in the documentation projects and possible future users will cooperate on the evolution of this ontology.

It has to be considered, however, that uniformity of description may not be achieved. All kinds of standardization at this level create new problems or define standards incompatible with earlier data. Additionally, it is highly questionable whether it will be possible to make all researchers collecting new data adopt the same kind of classifications, the same kind of database system etc. for their work, especially if they would like to contribute existing data — thus requiring a reformatting of data; secondly, researchers from different fields may emphasize quite different aspects in their data, so that their respective meta-data may not be easily comparable; lastly, it may also be thought of relating data from other information systems to the database, which then would appear in another format; to give an example, child language researchers may prefer to present their data in CHILDES format (cf. MacWhin-

ney 2000).
Thus, uniformity of description can only be achieved on a software level, a text database with intelligent text retrieval functions and the capability to combine various data formats during the retrieval process being the only solution to this problem. This is another meaning of the term "open system". It is obvious that there shall not be created a system of annotation tags, but rather a system which allows to analyse data for annotation tags. And it must be possible to unify different structures at a software level and not necessarily at the text level.

(Digital) Sound recordings and recording techniques

The audio workgroup of the MLL is expecting recordings on standard analogue and digital sound carriers; new recordings shall be performed by means of digital recording techniques. Field and studio recordings for the collection of speech material should be performed by means of semiprofessional recording techniques; DAT recordings 48 kHz / 16bit stereo, in binaural technique has been proven as appropriate. In order to get good material, the audio work group will introduce researchers into more advanced recording techniques.
The digitization of analogue sound data by the acoustic work station S_Tools (cf. Deutsch et al. 2000) is performed by a high quality analogue to digital conversion of a continuous audio stream generated from an analogue signal source as well as converting the digital audio stream back to an analogue signal. The quality of the conversion is determined by the resolution available from the analogue to digital (A/D) and digital to analogue converter (D/A). Although analogue tape recordings usually are limited in frequency range up to 16 kHz at a dynamic range of = 65 dB, analogue to digital conversion is recommended to be performed at 48 kHz or 44.1 kHz with 16 bit (i.e. ~ 90 dB dynamic range) resolution. Before digitisation, no prefiltering or signal conditioning at all should be applied. In case of signal enhancement or any other signal processing unavoidable, as denoise, decrackle, dehiss, these functions should be made on working copies or on the fly during usage. The archive copy has to be maintained as linearly transferred into digital format as possible.

Digital Streaming Format to File Format.

Digital Audio Tape (DAT) recordings are copied by means of the digital output of the DAT recorder on WAV files on the Digital Audio Workstation (DAW) disk. As hardware interface AES/EBU or Sony-Philips S/PDIF are available. Provided the recordings have been done carefully, the digital to digital copy needs no operator interaction because a one-to-one copy is created. In case sound files have already been created by another digitisation process, file copies are performed. Standardisation of different sample rates can be performed by high quality sample rate converter programs. Standardisation of different recording levels have to be considered as offline procedures. The production of audio CDs needs sample rate conversion to 44.1 kHz, 16 bit linear PCM. The data are thus digitized in high quality by the S_Tools workstations in WAVE format, which has become a standard. It should be emphasised that sound data of standard linear coded WAVE files remain playable with any wav-player available whether or not additional chunks are included. Whereas the content of user chunks needs to be managed by special application software components.[2] For the original signals, no compression is used in order to avoid data loss. The data are stored on a server system.

File naming conventions.

Before any digitisation takes place the project has to determine definitive file naming conventions for sound, text, video, images and descriptive information to be used throughout the project, in order to faciliate the integration of different media sources in the data base management system. File names and directory trees are crucial issues for long term storage and migration strategies as well as for access purposes.

Annotation of Sound in S_Tools

Additionally to the sound files, special application programs can create metadata related to passages of the sound files. In the case of S_Tools, this is done by creating one or more so-called directory files which contain lists of virtual sound segments by giving a name, begin and end of the segment (internally in samples, but also accessible in seconds, etc.), and further information, e.g. a transcription or various classifications. These sound segments themselves are not ordered hierarchically, but can be ordered by programs. These lists are represented in a delimited format and can thus be used as references in the multimedia database.

Metadata.

Additionally to metadata stored already in dedicated chunks of the wave file, raw segmentation of soundfiles and preliminary annotation will be provided on separate

[2] Out of a variety of sound file formats two evolved as a de facto standard: AIFF used in the MAC/UNIX world and RIFF/WAVE in the PC domain. This scenario was found by an EBU project group P/DAPA (Digital Audio Production and Archiving) when negotiating with industry in order to propose a common file format for linear audio quality serving the AES/EBU hardware interface standard. In order to generate and process descriptive information conveniently metadata should also be included in the file format. The group decided to select the widespread RIFF/WAVE format as a proposal for a standard. One major advantage of the WAVE file format can be used: WAVE files are world wide native files on all PC platforms and each PC is able to play and edit them. WAVE files are also used for audio data import and export on several other computer platforms. In order to enable standardised audio programme exchange the group developed the so-called Broadcast Wave Format. The main issue consisted in the agreement on a specially designed 'Broadcast Extension Chunk' (BEXT Chunk) for storage of additional metadata and descriptive information in sound files.

text files in ASCII ready for import in a database application. HTML or XML versions with linkage to the sound can be implemented. Further descriptive information (recording protocols, side information) has best to be linked to raw segmentation data and sound segments initially at generation time of the sound files. The completion of narrow segmentation and transcription data occurs usually at separate work sessions.

Segmentation.

Completely automatic segmentation of the speech waveform has been performed with varying degrees of success. The accuracy of automatically extracted segment boundaries as well as the recognition of segments leaves much to be desired. Depending on speaker variation which introduces large differences in actual realisations and depending on the quality of language models available, the accuracy of measurements and recognition rate occasionally decrease to unacceptably low levels. As a consequence, operator interaction during automatic segmentation or manual segmentation is still unavoidable.
For this reason, the S_TOOLS software supports manual segmentation, especially necessary for insufficiently described languages or dialects, such as are addressed by the current project. Best results have been obtained by applying a semiautomatic segmentation procedure which reliably extracts basic acoustic features such as pauses, voiced — unvoiced segments, pitch contours, formant frequency candidates and stressed vowel-like sounds (including quality).

While in some cases, the reference linking of an entire speech file to a transcription or a text file might be sufficient just for "listening to", the correspondences of a search on the narrow transcription or text file require a fine-grained audio data segmentation. Segmentation on the utterance and word level should be performed at the time of creation of text transliterations.
The basics of narrow phonetic transcription are composed of individual micro sound segments (audio micro objects) which are addressed sample by sample from the beginning of a sound file. Usually, sample number offset and duration of the segment is referenced. The segment structure is not limited to a single segmentation layer and relative addressing of segments is supported. In order to facilitate context related access to sound segments, overlapping segmentation is implemented. Narrow phonetic transcription can be performed in cumulative procedures.
Archived sound data usually do not change anymore after digitising. What has to be updated in regular intervals are segmentation data and metadata links, the location of suitable cue-in points, segment sequence procedures for rapid browsing, the creation of clips and several further archive staff accessible functions. The concept to manage the metadata separated from the sound files enables fast and easy access and virtual (non-destructive) processing of speech sounds.

Acoustic Analysis and Extraction of Speech Parameters.

On condition that segmentation of speech data has been performed accordingly, automatic acoustic analyses and speech parameter extraction can be carried out.[3] S_TOOLS provides a batch processing facility which enables the user to compute the relevant parameters under visual and audio control as well as to import them into a statistics software package for consecutive processing.
Parameter values such as fundamental frequency, formant frequency candidates, RMS values etc. are written into so-called 'dataset files' which serve as input for automated statistical processing procedures. S_TOOLS cooperates with Mathcad and several other third party statistical program packages.

Combination of "primary data" and "secondary data"[4]

Within MLL, a multimedia representation format will be provided to the users, which combines all kinds of data in a uniform environment. Based on earlier databases (cf. Deutsch et al. 1998), a possible relational model as in fig. 3 can be proposed: languages are represented in metadata and documents to which transcriptions belong; each transcribed entity can be linked to a sound segment in a sound file (or to pictures or video sequences).
Several standards, industry-standards and proprietary standards, can be used to express the links needed to refer to segments. However, since in many cases neither the audio recording nor the description or other audio segment linked to it can or should be changed, a simple hyperlink scheme as in HTML is not sufficient. Instead, it is necessary to use so-called 'independent' hyperlinks, which are external to the files they link. In addition, these hyperlinks

---

[3] Currently, the following speech signal display and parameter extraction modules are - among others - available: Free selectable display and editing of speech waveforms ranging in duration from several hundred samples up to file lengths of 4 Gbyte, variable frequency (resolution) analysis (FFT) ranging from 5 ms window length up to 2 s without averaging and up to the file lenghth with averaging, LPC analysis (autocorrelation method), Cepstrum analysis, SIFT fundamental frequency extraction and harmonic grid analysis, Pseudo Wigner Distribution spectrogram (in preparation), RMS - amplitude display with selectable integration intervals, RMSB - frequency band RMS amplitude, frequency bands selectable from FFT channels.

[4] The possible input for the archive consists of "primary" and "secondary" data; primary data are corpora consisting of various text types and data about the speech community. Written texts, sound, images and videos may be used therein. Secondary data may contain additional information and analyses such as "phonology of the language", "practical orthography", interlinear transcriptions (at morpheme level), "free translations", etc. Additionally, information about "special lexical data", documentations of informant interviews, situational or cultural descriptions of texts, picture sequences and/or communicative situations, i.e. the interpretation of primary data as well as "abstracts" can be made accessible.

must be bi-directional (description to audio and vice versa) to allow both, applications like querying the text and playing back the speech, as well as playing the audio and switching e.g. to the display of the score at an arbitrary point in time. Currently, the most advanced linking mechanism is the HyTime ilink (cf. Goldfarb et al. 1997); similar functionality can be provided by other implementations and will likely be provided by XML linking mechanisms. Until now, the entries in the 'sound file directories' were used as linking data for (various, external) database systems combining these entries with phonetic transcriptions, headers, metadata, etc.; but S_Tools will soonly be supplied with such advanced linking capabilities.

Literature

CCSDS (ed.) (1999). Reference Model for an Open Archival Information System (OAIS). NASA, Washington, DC, May 1999.

Commission of the European Communities (ed.) (1997). HARMONICA. Concerted Action on Music Information in Libraries. WP 3: Networking and Digitisation. Libraries Programme, May 1997.

Deutsch, Werner A. (2000). STX - Intelligent Sound Processing Tools. Vienna: Austrian Academy of Sciences, Acoustics Research Institute.

Deutsch, Werner A. & Ralf Vollmann & Anton Noll & Sylvia Moosmüller (1998). An Open Systems Approach for an Acoustic-Phonetic Continuous Speech Database. The S_Tools Database Management System STDB in: John Nerbonne (ed.): Linguistic Databases. Stanford/Calif.: CSLI Publications (= CSLI Lecture Notes Number 77), S. 77-92.

Deutsch, Werner A. & Anton Noll (2000). S_Tools (STX) — Intelligent sound processing tools. Vienna: Avcoustics Research Institute, Austrian Academy of Sciences.

Gibbon, Dafydd & Roger Moore & Richard Winski (eds.) (1998). Handbook of standards and resources for spoken language systems. 4 vols. (Spoken language system and corpus design; Spoken language characterisation; Spoken language system assessment; Spoken language reference materials). Berlin, New York.

Goldfarb, Charles F. & Steven R. Newcomb & W. Eliot Kimber & Peter J. Newcomb (1997). Information-Processing — Hypermedia/Time-based Structuring Language (HyTime). 2nd ed.. ISO/IEC JTC1/SC18/WG8 N1920; http://www.ornl.gov/sgml/wg8/docs/n1920/html/n1920.html

MacWhinney, Brian (2000). The Childes Project : Tools for Analyzing Talk. 3rd ed.

Specht, G. (1998) Multimedia-Datenbanksysteme: Modelle - Architekturen - Retrieval. Habilitationsschrift, TU München

Subrahmanian V.S (1998). Principles of Multimedia Database Systems, San Mateo.

Urban, B. (ed.) (1996). Multimedia'96. Springer Computer Science. Wien.

Vollmann, Ralf (1998). The structure of the multimedia database STDB. in: Petr Sojka & Václav Matousek & Karel Pala & Ivan Kopecek (eds.): Text, speech, dialogue. Proceedings of the first workshop on text, speech, dialogue — TSD'98, Brno, Czech Republic, Sept. 23-26, 1998. Brno: Masaryk University, pp. 327-332.

Zaefferer, Dietmar (1995). Options for a cross-linguistic reference grammar database. Paper presented at the Conference on Linguistic Databases, University of Groningen, 23-24 March 1995.

Appendix: Figures

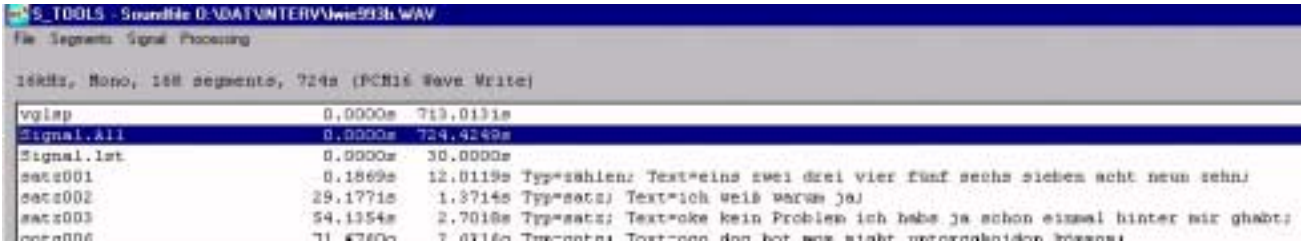

Fig. 1: 'sound file directory' containing in columns from left to right: segment identifier, segment start address, segment duration, segment type, segment content description. Additional categories can be introduced as necessary. The output format shown in this example is ready for use in MS Access.

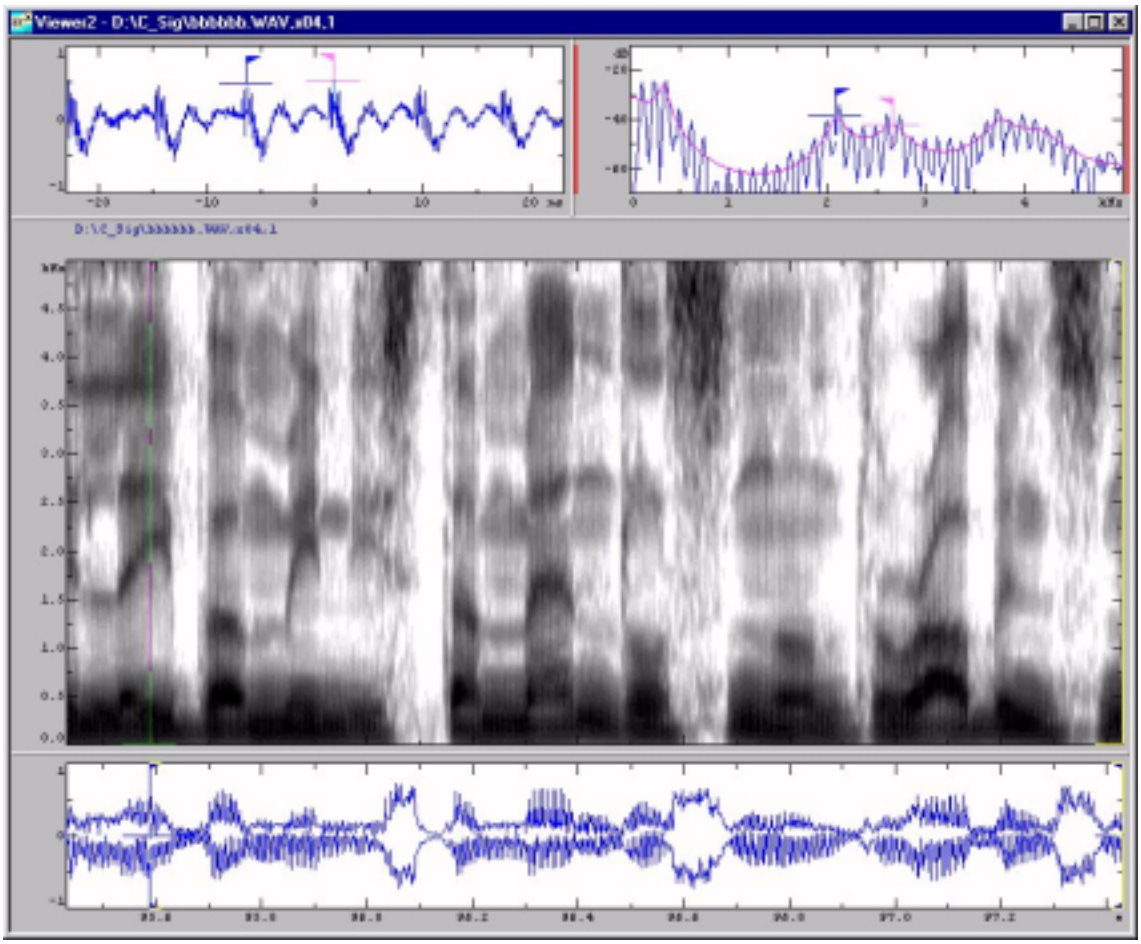

Fig. 2: Spectrogram display window of S_TOOLS. Bracketing of arbitrary sound segments can be performed on the wide band spectrogram under waveform zoom control and LPC-formant candidate extraction.

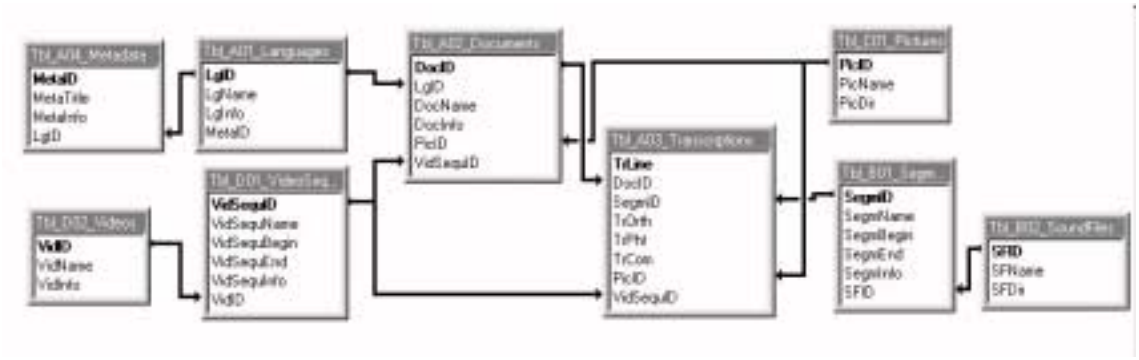

Fig. 3: Possible underlying structure of the multimedia database.