LARGE-SCALE GATHERING AND TAGGING OF SPEECH SAMPLES

Tiago Miguel Carmo Borba

Project/Dissertation developed under supervising of Ana Aguiar (Prof) in Instituo de Telecomunicações - FEUP

1. Motivation

This master thesis work was developed within the VOCE project's initial task goals – to collect and annotate an academical environment speech database, i.e. a *corpus*, composed of speech samples and physiological sensor data annotation. Furthermore, the collection workflow includes speech stress detection through objective physiological sensor data classification (i.e. tagging), and subjective human, three-folded - psychologist, volunteer and self-assessment - stress tags. After processing, all data will be used in future VOCE tasks, for stress-related feature extraction, and application of machine-learning to produce speech stress detection algorithms through voice input alone.

2. Main Goals

In order to gather consistent and unbiased data for the *corpus*, collection should include a sequential, fail-proof process, which even requiring heavy human-computer interaction, due to its nature, is kept simple and easy to use. For this a software solution, composed of a smartphone and a netbook application interacting through client-server paradigm, was developed, and second stage processing for sample segmentation was planned.

3. Work Description

Data collected includes audio files, and multiple annotation files, for the sensor data, the basic information of each recording, and psychological and demographic questionnaires that are also required. For stress-detection purposes, collection of baseline data is included in this process, complementing the speech event recording. Relying on the format and data of the previous solution, the second stage processing completes the *corpus* workflow. In this stage, data is uploaded to and centralized in a web server, where the application of audio segmentation techniques should take place to split recordings in utterance-like samples. These smaller samples represent the individual *corpus* elements, and to them physiological sensor-based stress detection takes place for objective and subjective tagging. Subjective self-assessment is generalized to all single recording related samples, but volunteer and psychologist stress assessments target single samples.

3.1. *Corpus* Collection and Tagging for Stress Classification

Taking the goals presented in the previous section as foundation, architecture was developed for the full workflow (as opposed to the single recording workflow which includes the second stage segmentation and stress assessment) of each single *corpus's* element acquisition. Fig. 1 presents both the architecture and flow of data until its final storage in a sample database - the desired output. *SCP Sensors* and *SCP Server, both* illustrated in Fig. 1, are the two applications that compose the collection platform module, being the former a smartphone-running application for the collection and relaying of sensor data, and the later a centralized flow control application running on a netbook computer. The collection platform includes other objects, all referring to hardware devices used by the applications. Data collected is initially stored on a file system where *SCP Server* runs, and includes annotation files, audio files, and demographics and psychological questionnaire answers. The annotation data that accompanies output of this component complies with the novelty *corpus* annotation standard EXMARaLDA. The Stress Tagging and Database part of the architecture, also referred to as segmentation and tagging platform, is a second part of the VOCE *Corpus* acquisition that has started

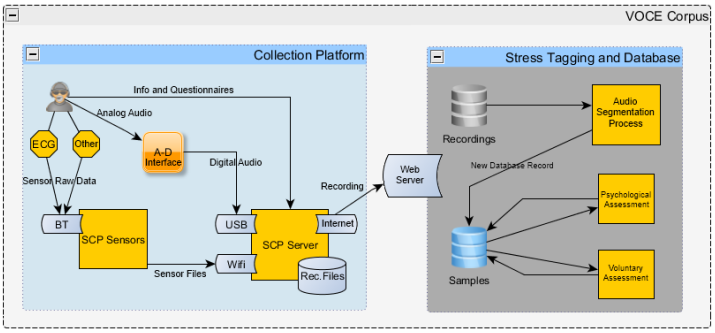

Figure 1 – Proposed solution architecture

development throughout this thesis work, but not fully implemented. It features the segmentation into small samples and the sample assessment components that compose one *corpus* element extracted from the recordings.

3.2. Sample Collection Platform

The sample collection platform is a set of applications developed for the collection of the raw

data that is to be processed and tagged with stress classification. Information includes speech recordings (WAV files), along with annotation of the metadata and physiological sensor data from ECG. The application developed is easily extensible to include other sensors. The application was developed following the requirements of the collection workflow, which includes the initial recording of 2 baseline recordings: the first one has been identified as **pre-baseline**, a recording the day before the speech event that is targeted; and the **baseline**, recorded immediately before the speech event, on the same day. Questionnaires are also included to correlate with specific needs of stress analysis, and also analysis by a psychologist and a definition of the stress self-assessment of the subject. It was built on the client-server paradigm where a netbook computer, on which the main program is run, acts as the server, and a smartphone acts as the client connects to the physiological sensors via Bluetooth. It then runs a program that stores the physiological sensor data and relays parts of it to the server via WiFi for immediate visualization by the collection monitor, and eventually sends all data when the server so requires. The result of each collection workflow is a set of files characteristically structured, in this case a folder with 13 files, where the 3 STAI XML's, the 3 audio WAV's, the 3 sensors XML's, the 3 annotation XML's and the demographics questionnaire are stored and ready for processing by the next steps of the project.

The application integrates a set of devices required for different data acquisition: *SCP Sensors*, the client application, handles physiological sensor data through Bluetooth connectivity from the subject, which; *SCP Server* connects to an audio device from an external analog-digital interface, and this device receives analog audio from a wireless microphone that the subject must also wear. The applications themselves run on an Android Smartphone (*SCP Sensors)* and on an Ubuntu netbook (*SCP Server).* Both of them were developed for the Java framework and specifically the server is OS independent.

Data collected is individual to each recording. After a collection ends, there are 3 sensor files, 3 annotation files, 3 audio files and 3 questionnaire files, one for each stage, and a single demographics file. The recordings can also be uploaded to a web server through the *SCP Server* application. This web server is the destination of recordings, where they will be further processed by the stress tagging and database module.

Extensive acceptance testing was conducted through real scenarios at FPCEUP, and aided development providing fine-tune to minor issues. Assessing all tasks for each stage of a single recording, it could be estimated that any collection will take about **27 minutes for collection the day before**, **and 1 hour the day of the event**, with the reference 30 minute event duration. Since testing took place for both sequential events during classroom work presentations, some measures were also taken to provide more usability in such scenarios, approximating their times to single subject recordings.

3.3. Segmentation and Tagging Platform

The second module of the *corpus* acquisition workflow was fully planned, and initiated development for interface with the first module. Specifically, a database for storing recording information and references to its files was designed and deployed with a web server. The server replies to and receives recordings from the *SCP Server* application, adding them to the database. Segmentation of samples, the first process to be applied to a recording, is planned to use external punctuation and pause detection algorithms as a service, provided by INESC-ID collaborators. The segmentation process is illustrated in Fig. 2.

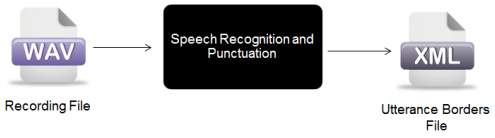

Figure 2 – Segmentation interface with tagging platform

Physiological sensor data can then be used to classify single utterances objectively, using baseline parameters for comparison. After objective classification, volunteer and psychologist assessments must be associated to the samples and complete the sample, since self-assessment was applied to every sample during segmentation (it is the same for every recording sample). Volunteer assessment will be conducted through a website and include multiple samples selected by established parameters, avoiding assessment bias. The psychologist assessment is to be conducted in a similar fashion, but with manual selection of samples to assess.

4. Conclusions

The components of the collection workflow include the collection platform and the sample tagging platform. Fully implemented and thoroughly tested, the collection platform applications provide a new paradigm for simultaneous gathering of physiological sensor data, audio recordings and computer-based questionnaire information. The platform's outline has minor influence on the quality of data retrieved, in comparison with acquisition of all information separately through conventional methods. It furthermore builds upon the simultaneous acquisition paradigm to acquire substantial data that can be used to correlate the different information types, wich could not be so easily acquired through separate gathering. The platform is easily extensible, configurable and partially OS-independent thanks to

the Java language. It provides clear separation of recording subject-centered tasks and recording monitor-centered tasks, providing an intuitive, fail-proof user interface for each task. Real-time data monitoring is acceptably retrieved, and information storage is centralized. This storage, which presents a form of output of the platform, is structured solidly and follows annotation standards for the type of data gathered, specifically the EXMARaLDA *corpus* annotation standard. Connecting the two main workflows of the whole corpus collection acquisition, a server upload feature will provide the output to a remote database, centralizing different sources of collection data and separating the segmentation and tagging tasks from the collection platform core features. Furthermore, completion of the stress assessment and tagging platform using the mentioned guidelines and other parameters defined in the master thesis work are expected to provide compliance with the full necessity of the VOCE *Corpus*.