2008-01-01

Metadata Visualisation Techniques for Emotional Speech Corpora

Charlie Cullen
Technological University Dublin, charlie.cullen@tudublin.ie

Brian Vaughan
Technological University Dublin, brian.vaughan@tudublin.ie

Spyros Kousidis
Technological University Dublin, spyros.kousidis@tudublin.ie

Recommended Citation

Cullen, C., Vaughan, B. & Kousidis, S. (2008) Metadata Visualisation Techniques for Emotional Speech Corpora. *AIR: Second International Workshop on Adaptive Information Retrieval,* London, UK. 18 October.

Funder: European Commission under contract FP6-027122, “Semantic AudiovisuaL Entertainment Reusable Objects- SALERO”.

Metadata Visualisation Techniques for Emotional Speech Corpora

Charlie Cullen[1], Brian Vaughan[1], Spyors Kousidis[1], John McAuley[1]

[1] Digital Media Centre, Dublin Institute of Technology, Aungier Street, Dublin, Ireland
charlie.cullen@dmc.dit.ie

Abstract. Our research in emotional speech analysis has led to the construction of dedicated high quality, online corpora of natural emotional speech assets. Once obtained, the annotation and analysis of these assets was necessary in order to develop a database of both analysis data and metadata relating to each speech act. With annotation complete, the means by which this data may be presented to the user online for analysis, retrieval and organization is the current focus of our investigations. Building on an initial web interface developed in Ruby on Rails, we are now working towards a visually driven GUI built on Adobe Flex. This paper details our work towards this goal, defining the rationale behind development and also demonstrating work achieved to date.

1 Introduction

Initial work undertaken as part of the SALERO[1] [1] project used mood induction procedures (MIP) [2, 3] to develop corpora of high quality emotional speech assets obtained under laboratory conditions [4]. The use of Mood Induction Procedures to stimulate emotion has the potential for the same recording conditions to be applied in natural speech as with simulated assets. The difficulties associated with such conditions using MIP's are related to the concealment of recording equipment to avoid revealing the true purpose of the experiment prior to commencement. In Kehrein's experiment, the fact that the participants were seated in separate sound proofed rooms, allowed the conversational interaction to be recorded as two separate high quality audio channels. This allowed both sides of the conversation to be analysed, including overlaps. Similarly, Johnstone [5] used computer games in order to induce real emotional states in test subjects. Johnstone found that computer games were well suited for this purpose as they can be changed and manipulated in order to induce the desired emotions.

A combination of the two experimental designs offers advantages: using computer games as part of a cooperative, task-based MIP offers a high degree of control, either hindering or aiding participants, while the use of separate sound proofed rooms enables high quality audio assets to be obtained. This approach ensures that obtained assets are natural, compared to simulated and broadcast assets, with emotional responses been induced as a result of while the co-operative aspect ensures the social aspect of emotional expression is not neglected. The resulting emotional assets can be claimed to be natural and spontaneous, arising out of the manipulation of the task and

[1] http://www.salero.eu

the interaction of the participants as opposed to voluntary or knowingly coerced attempts to generate emotional states.

2 Corpus Metadata Specification

The only cohesive attempt at corpus metadata standardisation performed thus far has been by the EAGLE/ISLE consortium [6], which has led to the development of the ISLE Metadata Initiative (IMDI). Although not a comprehensive (or universally adopted) standard, IMDI represents the only current standard for speech corpus metadata available. The metadata used by the emotional speech analysis corpus is organised in the following manner (Figure 1):

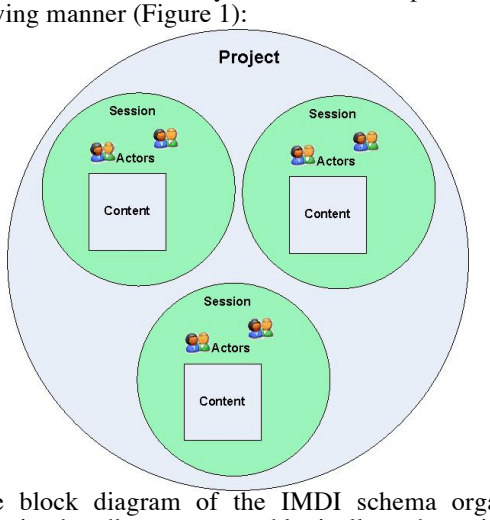

Figure 1: Example block diagram of the IMDI schema organisation. In this example, 3 separate session bundles are grouped logically under a single project. Each session relates to a specific type of content, and involves various actors who deliver the speech acts

2.1 Project

The definition of a particular project allows various sessions to be grouped in a logical form. Thus, in the case of the emotional speech corpus described in this deliverable all sessions are organised relative to the Salero project. By grouping sessions logically, it will allow future expansion of the database to include other corpora developed for different purposes (such as language learning or media production).

2.2 Session

A session is defined as the common bundle for linguistic events within IMDI metadata, and thus all speech assets are defined relative to a specific session. This allows an audio clip to be taken from a longer recording for specific analysis, while still retaining the same overall metadata as all other files in that session bundle. The session definition provides a convenient way to group assets for analysis, allowing assets taken from different experiments to be assessed either in isolation or within a wider common context.

2.3 Actor

The definition of an actor(s) within a session is a very useful aspect of the IMDI standard, as it allows the various participants in a speech recording to be documented for later consideration. In many instances, an actors details may be anonymised to ensure that ethical standards are adhered to (this is given as an option for each testing participant). Having said this, it is also very useful to consider database queries based on metadata such as geographical location or language, to allow broader linguistic analysis to be performed. Future work may consider the multi-lingual definition of assets within a corpus for analysis, and thus actor information would be crucial in this regard.

2.4 Content

The content metadata defined relates to specific activities for a given session. Definitions of genre and sub-genre are open vocabularies, while other terms such as interactivity and planning type are taken from IMDI standard closed vocabularies. By providing more information relating to the type of speech asset being annotated, it is hoped that wider queries can be made in the corpus database as the record set expands over time.

2.5 Asset

Each asset in the corpus is defined in terms of its audio quality, which in turn relates to the LinguaTag SMIL analysis data defined in D6.1.1 and D6.1.2 (Figure 2):

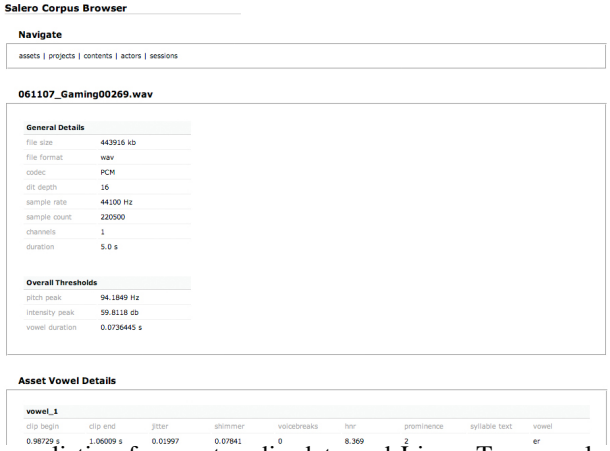

Figure 2: Corpus listing for asset audio data and LinguaTag vowel analysis data. Only the first vowel is shown on this screen

In this manner, an asset could be queried in terms of its emotional dimensions (contained within the LinguaTag SMIL analysis file) or for specific acoustic attributes related to a vowel within the clip. This will allow investigation into the acoustic correlates of emotional speech to be performed using high quality speech assets, subject to emotional clip rating by listener groups (see section **Error! Reference source not found.**). The SMIL data output by LinguaTag also forms the basis of vowel stress animation methods currently under testing in the PGP experimental production 'My Tiny Planets' (see D9.3.2).

3 Implementation

3.1 Requirements

There were, from the outset, several considerations that helped to define the technical architecture of the corpus. Firstly, the prototype must provide editors with the ability to insert assets, in the form of wav files, and related linguatag data, in the form of SMIL files. The prototype must parse the SMIL file and populate the corresponding database tables. The corpus, therefore, necessitates a storage layer or database as a persistent back-end. Secondly, editors require remote access to corpus assets. This allows for the addition, deletion and alteration of corpus assets and related metadata. At first, each asset was to be uploaded and annotated individually. However, following initial trials, it was decided to provide the ability for batch uploads, thereby allowing an editor to upload several assets at the any one time. In this case, each asset is annotated with the same metadata.

Automated Annotation

A central requirement of the prototype was to reduce the overhead associated with annotating digital assets. Often metadata can be reused for multiple assets, and does not require the editor re-entering this data every time they wish to add a new asset. Therefore the approach involved using Ajax auto-suggest as a way to reduce the annotation overhead when entering new assets. An editor can enter any piece of metadata and re-use that piece of metadata through the autosuggest functionality. Similarly, there are two approaches to inserting metadata; the first allows editors to enter metadata and then upload the asset, annotating the asset with autosuggest functionality as outlined above. The second allows the editor upload the asset and annotate the asset on the fly, effectively creating new metadata.

3.2 Architecture

As with traditional web architecture, the corpus is divided into three separate tiers, presentation, application and data:

Presentation

The presentation tier displays the corpus assets and related metadata. As this is a prototype application, a simple style was used to delineate assets from the different forms of metadata.

Application

The application tier contains the business logic of the corpus. The prototype, v0.1, provides two approaches to uploading and annotating assets. The first allows the editor to individually upload and annotate a single asset. The second allows an editor to batch upload assets, therefore reducing time when populating the corpus. As mentioned previously, the second approach requires that all assets are annotated with the same metadata.

Data

The data tier provides persistent storage for the corpus metadata. The audio assets and related linguatag SMIL files are not stored in the data tier. Conversely, each asset and SMIL file is stored on the web server to reduce overhead when querying and retrieving assets from the data tier.

3.3 Technologies

Presentation and Application Tiers

The Presentation and Application tiers were developed using Ruby on Rails[2]. Ruby on Rails is an open source web framework for the rapid application development. The framework is an implementation of the model view controller design pattern and provides an excellent migration mechanism for creating and altering the data tier. Ruby on Rails has support for XML, and consequently SMIL, through the REXML[3] ruby-gem, which is packaged with the latest version of Ruby.

Data Tier

The popular open source MySQL[4] database provides a foundation for the data tier. The creation and manipulation of the database is carried out through the Ruby on Rails framework. The corpus is available at URL: http://corpus.dmc.dit.ie/annotate/login.

4 Ongoing and Future Work

Initial work undertaken as part of the

Acknowledgments: The research leading to this paper was partially supported by the European Commission under contract IST-FP6-027122 “SALERO”.

References

[1] W. Haas, G. Thallinger, P. Cano, C. Cullen, and T. Bu¨rger, "SALERO - Semantic Audiovisual Entertainment Reusable Objects," in *International Conference on Semantic and Digital Media Technologies (SAMT)* Athens, Greece, 2006.

[2] A. Gerrards-Hesse, K. Spies, and F. W. Hesse, "Experimental inductions of emotional states and their effectiveness: A review," *British Journal of Psychology,* vol. 85, pp. 55-78, 1994.

[3] C. Cullen, Vaughan, B. ,Kousidis, S., Wang, Yi ., McDonnell, C. and Campbell, D. , "Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction " in *International Conference on Multidisciplinary Information Sciences and Technologies* Extremadura, Merida, 2006.

[4] C. Cullen, B. Vaughan, and S. Kousidis, "Emotional speech corpus construction, annotation and distribution," in *The sixth international conference on Language Resources and Evaluation, LREC 2008* Marrakech, Morocco, 2008.

[5] T. Johnstone, C. M. v. Reekum, K. Hird, K. Kirsner, and K. R. Scherer, "Affective speech elicited with a computer game," *Emotion,* pp. 513-518, 2005.

[6] ISLE, "IMDI (ISLE Metadata Initiative), Metadata Elements for Session Descriptions," Draft Proposal Version 3.0.3 ed, 2003.

[2] http://www.rubyonrails.org/

[3] http://www.germane-software.com/software/rexml/

[4] http://www.mysql.org/