The University of Edinburgh Head-Motion and Audio Storytelling (UoE-HAS) Dataset

Citation for published version:
Braude, DA, Shimodaira, H & Ben Youssef, A 2013, 'The University of Edinburgh Head-Motion and Audio Storytelling (UoE-HAS) Dataset'. ???in??? Proc. of Intelligent Virtual Agents. ???pages??? 466-467.

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Author final version (often known as postprint)

Published In:
Proc. of Intelligent Virtual Agents

The University of Edinburgh Head-Motion and Audio Storytelling (UoE - HAS) Dataset

David A. Braude, Hiroshi Shimodaira, and Atef Ben Youssef

Centre for Speech Technology Research,
School of Informatics, University of Edinburgh,
10 Crichton Street, Edinburgh, EH8 9AB, UK
{d.a.braude,h.shimodaira,abenyou}@inf.ed.ac.uk

Abstract. In this paper we announce the release of a large dataset of storytelling monologue with motion capture for the head and body. Initial tests on the dataset indicate that head motion is more dependant on the speaker than the style of speech.

Keywords: Head Motion, Dataset.

1 Introduction

There are very few datasets that have tracked head and body motion during speech. Those that do exist tend to be short and have very few speakers. To address this lack we are making available a dataset of storytelling monologues that was recorded at the University of Edinburgh. This dataset contains speech and motion capture of the head and upper body and so is suitable for research into virtual avatars or body language. We are also in the process of transcribing the speech to open up avenues of research into utilising linguistic information, and enable the data to be used in speech research.

2 Description of the Dataset

The subjects were 16 UK native English speakers. Nine were female and seven were male. Ahead of the recording session the participants were given five stories, these were classical fairy-tales that they should have been familiar with from childhood.

During the recordings the participants were given the story on a teleprompter (Read speech) and then were asked to retell the story in their own words (Free speech). Previous recordings showed when speakers were asked to choose their own stories for free speech but these stories generally lasted less than two min. They were seated during the recording and they were instructed to tell the story as if to an adult native English speaker.

Five motion capture markers were placed on the chest and the participants wore another four markers on a hat to capture the body and head motion respectively. The motion capture was done with the Natural Point, Optitrack system

Table 1. Lengths of recordings (min:sec)

	Read	Free	Total
Total	371:14	323:22	694:36

Table 2. Mean Cross Entropy distance between different utterances

	intra	inter
Speaker	0.97	3.15
Style	2.86	3.23

using seven V100:R2 cameras at a 100 Hz sampling rate. Audio was captured using a free-standing directional microphone. The audio was captured at 44100 Hz with 32-bit depth and down-sampled to 16-bit .WAV format using Audacity. The audio and motion capture start at the same frame. Table 1 shows the total lengths of recordings available.

To obtain the dataset please visit either the SSPNet Project[1] or CSTR[2] websites where further details about the dataset are provided.

3 Speaker and Scenario Dependency

To determine the similarity of the speakers the Cross Entropy distance was used [1]. A multidimensional Gaussian distribution was used to model the data.

The mean distance between examples of the same type (intra) and examples of different types (inter) is given in Table 2. Style refers to whether it was read or free speech.

From Table 2 it is clear that there are differences between read and free speech and head motion from different speakers and the speaker dependence is higher than the style dependence.

4 Future Work

We are currently in the process of recording a large dataset of dialogues. It will also have a large amount of speakers and long samples from each speaker.

Acknowledgements. This work was supported by EU FP7 SSPNet (grant agreement no. 231287) The financial assistance of the South African National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

Reference

1. Helén, M., Virtanen, T.: Audio Query by Example Using Similarity Measures between Probability Density Functions of Features. EURASIP Journal on Audio, Speech, and Music Processing 2010, 1–12 (2010)

[1] `http://sspnet.eu/`

[2] `http://www.cstr.ed.ac.uk/`