The AV-LASYN Database : A synchronous corpus of audio and 3D facial marker data for audio-visual laughter synthesis

Hüseyin Çakmak, Jérôme Urbain, Joëlle Tilmanne and Thierry Dutoit

University of Mons, Faculty of Engineering, TCTS lab
20, Place du Parc
7000, Mons/Belgium
{huseyin.cakmak},{jerome.urbain},{joelle.tilmanne},{thierry.dutoit}@umons.ac.be

Abstract

A synchronous database of acoustic and 3D facial marker data was built for audio-visual laughter synthesis. Since the aim is to use this database for HMM-based modeling and synthesis, the amount of collected data from one given subject had to be maximized. The corpus contains 251 utterances of laughter from one male participant. Laughter was elicited with the help of humorous videos. The resulting database is synchronous between modalities (audio and 3D facial motion capture data). Visual 3D data is available in common formats such as BVH and C3D with head motion and facial deformation independently available. Data is segmented and audio has been annotated. Phonetic transcriptions are available in the HTK-compatible format. Principal component analysis has been conducted on visual data and has shown that a dimensionality reduction might be relevant. The corpus may be obtained under a research license upon request to authors.

Keywords: laughter synthesis, audio-visual, database

1. Introduction

In the last years, human-computer interactions have dramatically increased. One of the most important research areas deals with the development of interfaces that behave like humans. The objective is to create as natural interactions as possible for humans, instead of having to adapt to machine specificities. For this purpose, virtual agents should not only be intelligible, but also expressive (i.e., able to convey affective states) and coordinated (among others, this implies a proper synchronization between the audio and visual modalities). Besides verbal capabilities, such expressive agents must also be able to express emotions through non-verbal activities. Laughter being a crucial signal in human communication, it is thus wished for agents to be able to display convincing laughs. Although commercial systems currently use a finite set of prerecorded laughs to choose from when laughter is desired, such a framework is limited to the available laughs and has poor flexibility. Several works recently explored the possibility of synthesizing laughter on the acoustic or visual modalities, separately. In this paper, we aim at offering new possibilities to develop audio-visual laughter synthesis, with the help of a database specifically recorded for that purpose. Possible applications of a laughter synthesis system are numerous in the field of 3D animation (video games, animation movies) and human-machine interfaces (mobile devices, navigation systems, interactive websites).

The database presented in this paper will be used to train laughter models, for both audio and visual modalities, following the statistical parametric speech synthesis framework also known as HTS (Tokuda et al., 2002). The feasibility for the acoustic modality alone has already been demonstrated with a smaller corpus in (Urbain et al., 2013b). In such approaches that are data driven, the available corpus is of primary importance and a trade-off has to be made between the quality, the size of data and the time that the building of such a corpus consumes. This is mainly the reason why the AV-LASYN [1] Database presented here contains only one male subject. However, the pipeline still remains relevant for further recordings of more data from one or more subjects.

Visual laughter synthesis systems are rare. A parametric physical chest model not including the face animation which could be animated from laughter audio signals was presented in (DiLorenzo et al., 2008). In (Cosker and Edge, 2009), authors studied the possible mapping between facial expressions and their related audio signals for non-speech articulations including laughter. HMMs were used to model the audio-visual correlation. In this latter work, the animation is also audio-driven. More recent studies like (Urbain et al., 2013a) or (Niewiadomski et al., 2013) include the animation of laughter capable avatars in human-machine interaction. The proposed avatar (Greta Realizer) is controlled either through high level commands using Facial Action Coding System (FACS) or low level commands using Facial Animation Parameters (FAPs) of the MPEG-4 standard for facial animation. They also proposed another avatar (Living Actor) which plays a set of manually drawn animations.

One particularity of this work is that the visual synthesis for which it is built is a 3D synthesis of motion trajectories. This 3D data offers the ability to drive virtual 3D characters and differs from other visual synthesis approaches that are based on 2D videos. This 3D requirement supposes

[1] For those who wonder what the name AV-LASYN means, it stands for AudioVisal LAughter SYNthesis.

that the available visual data is 3D as well. To meet this requirement, we have chosen to record facial deformation data with a commercially available motion capture system known as OptiTrack. To the best of our knowledge, no database meeting both technical (synchronous audio and 3D visual data) and sufficient size requirements is available for laughter. A similar pipeline for recording audio-visual data has been proposed recently for speech in (Schabus et al., 2012).

This paper is organized as follows : Section 2 gives briefly the motivation then Section 3 describes the recording protocol. Section 4 is dedicated to the post-processing which includes shaping visual data, synchronization of modalities, segmentation and annotations. Then the details of a PCA analysis on the visual data are given in Section 5. Finally, the types of data available in the corpus are summarized in Section 6, before the conclusions.

2. Motivation

The database presented in this paper is built to perform audio-visual laughter synthesis using an HMM-based framework as summarized in figure 1.

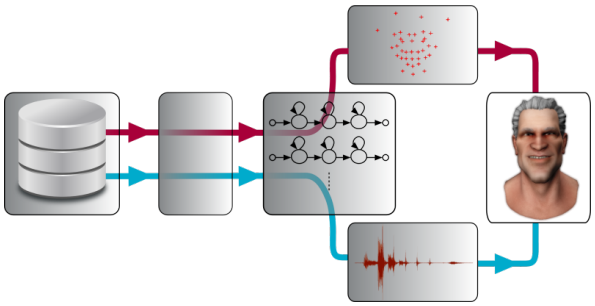

Figure 1: Overview of the pipeline for HMM-based audio-visual laughter synthesis

Basically, the general steps of the pipeline where the AVLASYN Database takes place are as follows :

1. Building a synchronous AV Database
2. Post-processing data to make it suitable for HMM modelling tools
3. HMM-based modelling of acoustic laughter as well as facial expressions
4. Synthesis of synchronous audio and visual laughter
5. Retargeting on a 3D avatar and rendering output

We have already investigated the use of this pipeline for the audio modality only (Urbain et al., 2013b). The AVLC Database has been used for this (Urbain et al., 2010). The HMM modelling tools used were HTK (Young and Young, 1994) and its HTS patch (Tokuda et al., 2002) for HMM-based acoustic synthesis.

This previous work on acoustic laughter synthesis also motivated the building of the corpus presented in this paper. Indeed, to the best of our knowledge and apart from the present work, the only audio-visual laughter corpus with 3D marker data for face motion is the AVLC Database which is made up of recordings from 24 subjects for a total amount of roughly 60 minutes of laughter. Also, 3D facial motion capture data is available only for a subset of subjects among the 24 present in the AVLC Database. This results in a quite small amount of recordings on a per-subject basis. For example, for the subject used in (Urbain et al., 2013b), only 3 minutes of laughter were available. This is drastically less than the amount of data commonly used for HMM-based acoustic speech synthesis. In depth information about the AVLC Database may be found in the related paper (Urbain et al., 2010).

As stated above, the aim is to perform HMM-based audio-visual laughter synthesis by extrapolating the pipeline used in (Urbain et al., 2013b) to audio-visual data. This implies the necessity to record new synchronous audio-visual data as well as some post-processing as explained in the remainder of this paper. While this paper is more focused on the steps 1 and 2 of the pipeline, more information on steps 3, 4 and 5 may be found in (Çakmak et al., 2014) in which a first use of this database for audio-visual laughter synthesis is presented.

3. Recording protocol

This section gives information about the experimental setup used for recordings. Figure 2 gives an overview of the recording pipeline.

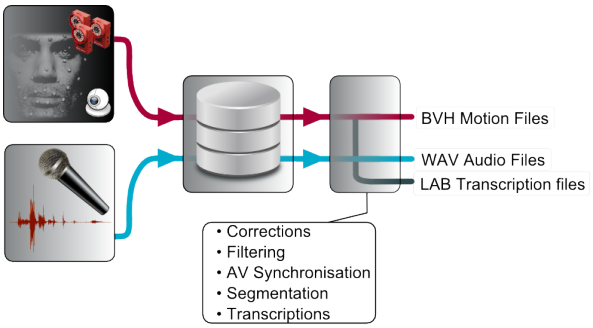

Figure 2: Data recording pipeline

3.1. The stimuli

The laughs were triggered by funny videos found on the web. The subject was free to watch whatever he could find as far as it was funny for him. A total amount of 125 minutes were watched by the subject to build this corpus.

3.2. Hardware and experimental setup

Audio hardware

An external sound card (RME Fireface 400) as well as a professional microphone (SHURE SM58) have been used for audio recordings. The audio was recorded at high sampling rate (96kHz) and encoding (32bits) in order to be able

to study the eventual sampling rate impact on HMM-based audio synthesis quality afterwards (Yamagishi and King, 2010). That being said, since we believe that such a sampling rate is not necessary for most of the applications, the data was further downsampled to 44.1kHz, 16bits encoding. The original data at 96kHz is still available though.

Motion capture hardware
The motion capture hardware that has been used is the Optitrack system from Naturalpoint [2]. A 7-camera setup was used and their placement may be seen on figure 3. These cameras emit and receive IR light at 100 fps. Among the 7 cameras, the one in front of the face of the subject was used to record a 640x480 grayscale video synchronous with all other cameras. This video is saved in a proprietary format by the tracking software provided with the hardware. This grayscale video was used to synchronize audio and marker data with the help of a clapperboard. Audio and motion recording were done on 2 different computers.

Figure 3: Camera placement for motion capture recordings. (Image from www.naturalpoint.com)

We have used a setup with 37 reflective markers, where 33 markers are glued on the face of the subject and the remaining 4 are on a headband. The 4 markers on the headband are used to extract global head motion from the other markers and thus make available head motion and facial deformation separately.

In addition to the motion capture system, a webcam was added to the setup. For each take, a 640x480 AVI file is also recorded at 30 fps. On these videos, the upper body and all markers on the face are clearly visible and this data might be valuable for further image processing if needed.

4. Post-processing

Once the data is recorded, we need to post-process it to make it suitable for HMM-based modeling.

4.1. Cleaning the visual data

The first step at this stage was to check the recorded visual data to get rid of eventual tracking errors that may occur in the form of gaps (discontinuity in trajectories) or swipes (brutal unexpected movements of tracked trajectories). All the recorded data was analyzed and corrected in this regard.

[2]http://www.naturalpoint.com

4.2. Removing head motion

To make head motion available independently from facial deformation data, we have used the 4 headband markers data. Assuming that these 4 markers are always separated by a fixed distance from each other, the movement of the pattern that they form together represents the movement of the head. Therefore we can subtract this head motion from all other markers' motion so that after this process the 4 head markers will stay still while the rest of the trajectories will only contain facial deformation data.

We have chosen to save the data into the Biovision Hierarchy (BVH) format for the structure it provides. For this particular work, the main advantage of using this format is to have in the same file head motion data, neutral pose and facial deformation data with the ability to play them together in a third party software such as Autodesk Motionbuilder. Figure 4 summarizes the process of building the final motion files.
The data is also available in the C3D format as well as the proprietary FBX format but further processing explained in the next sections were only applied to BVH files.

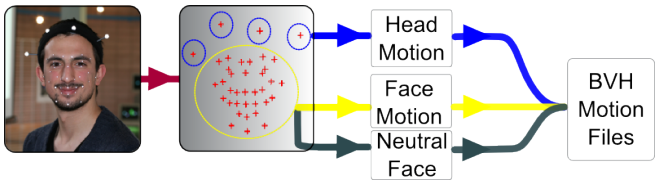

Figure 4: 3D data processing.

4.3. AV synchronization

As mentioned above, the synchronization between audio and marker data has been done using a clap signal which is clearly visible on the audio waveform as well as on the synchronized greyscale video. Since the latter has a frame rate of 100 fps, we are performing a synchronization accuracy of ± 5 ms.

4.4. Segmentation

At this stage, the data is stored in files containing several minutes of recording with several occurrences of laughter. What we call segmentation here is cutting these files into smaller files containing only one occurrence of laughter each. This segmentation has been done manually based on the video signal. The choice of doing it on the video comes from the fact that laughter has an effect on facial expression before it becomes audible and the audio laughter signal stops before the facial expressions disappear (Ruch and Ekman, 2001). Moreover, in this work, we also consider smiles which do not have any impact on the audio modality.
After this segmentation process, we end up with 251 occurrences of laughter including smiles with the distribution given in table 1. In this table, the *Smile* category refers to recordings where there is a facial expression but no sound related, the *Smile & Laugh* category refers to recordings

with relatively long periods without sound but distinguishable facial expressions followed or preceded by audible laugh and the *Laugh* category refers to audible laugh without significantly long smiles.

Type	Occurrences
Smile (visual only)	46
Smile & Laugh	35
Laugh	170
Total	251

Table 1: Laughter occurrences in corpus by type.

4.5. Phonetic Transcriptions

Once the data has been segmented as described above, each segment has been further sub-segmented into phonetic classes that describe the laughter with regards to the audio modality. In this paper, the terms phonetic transcriptions refer to the content of files transcribing the sequence of phone(me)s as well as their boundaries in the time domain. The format adopted for these transcription files is the format defined by HTK (Young et al., 2006). These phonetic annotations were done manually using the Praat software (Boersma and Weenink, V5351 2013) before being converted to the HTK label format. The phonetic classes used as well as their number of occurrences in the corpus are listed in table 2. You may refer to (Urbain et al., 2010) and (Urbain et al., 2013b) for further information about these transcriptions.

inhalation or exhalation	Phonetic Class	Occurrences
e	silence	899
e	fricative	861
e	ə	527
e	a	630
e	nasal	262
e	nareal fricative	226
i	nareal fricative	165
i	fricative	121
i	ə	9
e	o	9
e	plosive	9
i	plosive	3
e	glottal	3
i	nasal	2
i	silence	2
e	grunt	1
i	a	1

Table 2: Phonetic classes in the corpus and their number of occurrences

5. PCA analysis on 3D data

As pointed out by (Schabus et al., 2012), there are many strong constraints on the deformation of a person's face while speaking. This is still true when laughing. The full motion vector at each frame contains 99 dimensions for the face (x, y, z coordinates of 33 markers) and 6 dimensions for the head motion (x, y, z coordinates and x, y, z rotations). This allows 105 degrees of freedom which seems far too much to describe visual laughter. To verify this as well as to de-correlate the data, Principal Component Analysis (PCA) was performed. The PCA was carried out on all dimensions except rotations of the head because they are from a completely different nature (angles instead of lengths).

Let us consider the matrix M with n rows and m columns that contains all the 3D data of the corpus. The rows represent the frames while the columns represent the dimensions. We thus have in our case a n by 102 matrix. The PCA will provide us with a 102 by 102 matrix U where each column contains coefficients for one principal component. Another useful element given by the PCA is a vector V (1 by 102) with the principal component variances (the eigenvalues of the covariance matrix of M).

We have :

$$M_{PCA[n*102]} = \overline{M}_{[n*102]} \cdot U_{[102*102]}$$

where

M_{PCA} = the representation of M in the PCA space
$\overline{M}$ = the mean normalized M matrix
U = matrix of the coefficients of PCA components

One of the reasons for using PCA is that the resulting components are sorted according to their contribution on the variability in the data. Based on this consideration, it might be possible to reduce dimensionality by keeping only the first k components and still correctly represent the data. To determine how many dimensions to keep, we can compute the reconstruction error as a function of the number of kept dimensions k. Let the projection matrix that reduces dimensionality to k be denoted U_k. We thus have :

$$M_{PCA_k[n*k]} = \overline{M}_{[n*102]} \cdot U_{k[102*k]}$$

The reconstructed data M_{REC} from reduced dimensionality data M_{PCA_k} is then defined as :

$$\overline{M}_{REC[n*102]} = M_{PCA_k[n*k]} \cdot U^T_{k[k*102]}$$

to which we still need to add means of each dimension to finally obtain M_{REC}.

Figure 5 gives the Root Mean Squared Error (RMSE) of reconstruction as a function of k with RMSE defined as :

$$RMSE = \sqrt{\frac{1}{102 \cdot n} \sum_{i=1}^{n} \sum_{j=1}^{102} (M_{ij} - M_{RECij})^2}$$

We can see on figure 5 that with 5 principal components, the RMSE is below 1 mm and with the first 20 components it is below 0.2 mm.

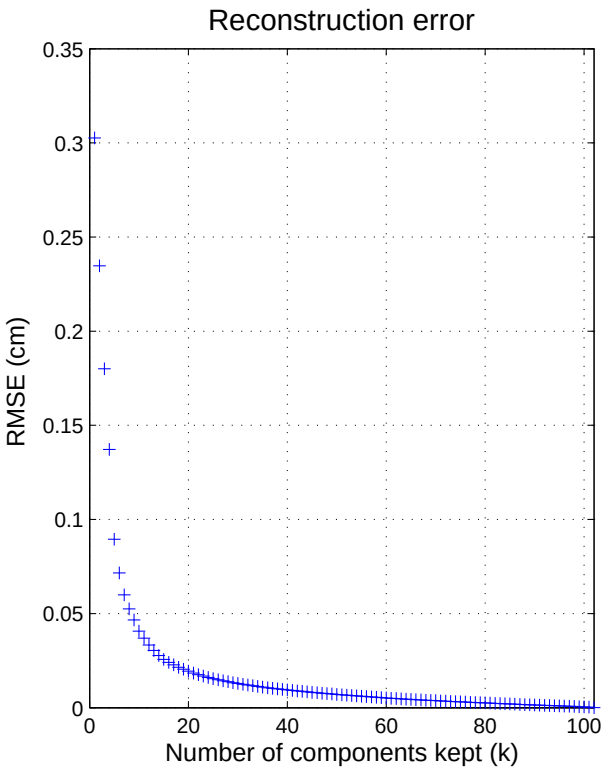

Figure 5: Reconstruction error as a function of the number of PCA components kept.

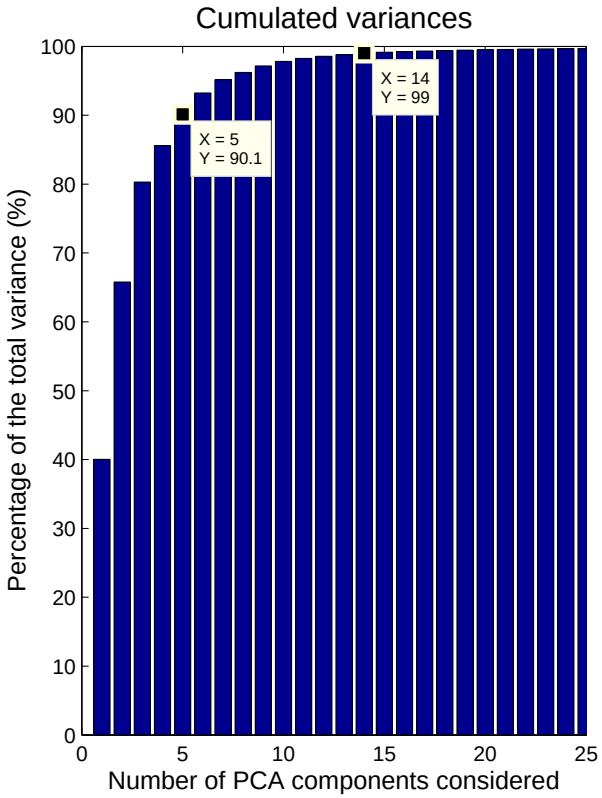

Figure 6: Cumulated contributions of each PCA components for the 25 first components.

Another way to determine the number of components to keep is to study the cumulated variances of the principal components. Figure 6 gives this information for the 25 first components of the PCA. We can see that the first 5 components represent more than 90% of the total variance while the first 14 components represent more than 99% of the total variance. This confirms that the initial 102 dimensional space is not necessary to accurately represent the data.

From the previous considerations, we can tell that it should be enough to keep between 5 and 20 PCA components to work with the visual data in this corpus. Further analysis on the contribution of each component might be relevant tough.

6. Contents of the database

The corpus contains 251 instances of laughter uttered by one male laugher while watching funny movies. This corresponds roughly to 48 minutes of visual laughter and 13 minutes of audible laughter. For each utterance, the corpus contains :

- A WAV audio file [44.1kHz, 16 bits]
- A BVH motion file that can be loaded in common 3D software which contains :
  - The neutral pose
  - 6 channels for head motion (3 translations and 3 rotations)
  - 3 channels for each of 33 facial markers (3 translations)
- A binary motion file containing the same data as in the BVH to make it easier to load programmatically to avoid parsing the BVH
- An HTK label file containing phonetic transcriptions and temporal borders for each laughter phone

In addition, the corpus contains the transformation matrix U, the variance vector V and the vector of means from the PCA described in this paper. The corpus also contains video data from the webcam integrated in the setup. These videos are in AVI format (30 fps, 640x480) and they are not segmented. Unsegmented 3D data is also available in FBX and C3D format, as well as the original audio recordings (96kHz, 32 bits).

7. Conclusion

In this paper, we have shown a recording protocol and the fundamental post-processing steps to follow in order to prepare data for audio-visual laughter synthesis. The PCA on visual data confirmed our thoughts that 99 dimensions (33 markers times 3 coordinates) are not necessary to describe the facial deformation during laughter. This does not mean

that we do not need 33 markers but that the motions of these markers are related to each other and therefore a dimensionality reduction may be applied to data while still correctly representing the motion as shown in (Çakmak et al., 2014). In future work, we are planning to extend this corpus to more subjects in order to have a bigger variety of laughs, to include female laughs as well and possibly investigate adaptation techniques in the HMM-based synthesis framework.

8. Aknowledgements

H. Çakmak receives a Ph.D. grant from the Fonds de la Recherche pour l'Industrie et l'Agriculture (F.R.I.A.), Belgium. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n270780.

9. References

Boersma, P. and Weenink, D. (V5.3.51, 2013). Praat: doing phonetics by computer [computer program].

Çakmak, H., Urbain, J., Tilmanne, J., and Dutoit, T. (2014). Evaluation of hmm-based visual laughter synthesis. In *Acoustics Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*.

Cosker, D. and Edge, J. (2009). Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations . In *Computer Animation and Social Agents (CASA)*.

DiLorenzo, P., Zordan, V., and Sanders, B. (2008). Laughing out loud: control for modeling anatomically inspired laughter using audio. *ACM Trans. Graph*.

Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., Geist, M., Lingenfelser, F., McKeown, G., Pietquin, O., and Ruch, W. (2013). Laugh-aware virtual agent and its impact on user amusement. In *Proc. int. conf. on Autonomous agents and multi-agent systems*, AAMAS.

Ruch, W. and Ekman, P. (2001). The Expressive Pattern of Laughter. *Emotion qualia, and consciousness*, pages 426–443.

Schabus, D., Pucher, M., and Hofer, G. (2012). Building a synchronous corpus of acoustic and 3d facial marker data for adaptive audio-visual speech synthesis. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Tokuda, K., Zen, H., and Black, A. W. (2002). An hmm-based speech synthesis system applied to english. In *Proc. of 2002 IEEE SSW, Sept. 2002*, september.

Urbain, J., Bevacqua, E., Dutoit, T., Moinet, A., Niewiadomski, R., Pelachaud, C., Picart, B., Tilmanne, J., and Wagner, J. (2010). The avlaughtercycle database. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Urbain, J., Niewiadomski, R., Mancini, M., Griffin, H., Cakmak, H., Ach, L., and Volpe, G. (2013a). Multimodal analysis of laughter for an interactive system. In *Proceedings of the INTETAIN 2013*.

Urbain, J., Çakmak, H., and Dutoit, T. (2013b). Evaluation of hmm-based laughter synthesis. In *Acoustics Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*.

Yamagishi, J. and King, S. (2010). Simple methods for improving speaker-similarity of hmm-based speech synthesis. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 4610–4613.

Young, S. and Young, S. (1994). The htk hidden markov model toolkit: Design and philosophy. *Entropic Cambridge Research Laboratory, Ltd*, 2:2–44.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. C. (2006). *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.