# Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations

Andrew J. Aubrey, David Marshall, Paul L. Rosin, Jason Vandeventer
School of Computer Science, Cardiff University, Wales, UK
(a.j.aubrey, dave.marshall, paul.rosin, j.m.vandeventer)@cs.cardiff.ac.uk

Douglas W. Cunningham
Brandenburg Technical University,
Cottbus, Germany
douglas.cunningham@tu-cottbus.de

Christian Wallraven
Korea University, Korea
wallraven@korea.ac.kr

## Abstract

*We present the Cardiff Conversation Database (CCDb), a unique 2D audiovisual database containing natural conversations between pairs of people. The database currently contains 30 conversations. To date, eight conversations are fully annotated for speaker activity, facial expressions, head motion, and non-verbal utterances. In this paper we describe the data collection and annotation process. We also provide results of baseline experiments in which an SVM classifier was used to identify which parts of the recordings are from the frontchannel speaker and which are backchannel signals. We believe this database will make a useful contribution to computer vision, affective computing, and cognitive science communities by providing raw data, features, annotations and baseline comparisons.*

## 1. Introduction

Social interaction is a central part of most people's daily life. Increasingly, people are communicating not just with other people but also with a wide variety of automated services. An accurate model of how people convey and respond to different forms of information would be a great aid to the development of socially enabled devices – a core goal in the field of affective computing [22].

Early work on conversational modelling focused on written transcripts of conversations. As a result, traditional models of communication assumed that in any dyadic conversation one person was active (the speaker) and one was passive (the listener). Since at least 1970, however, it has been repeatedly shown that human conversations are very much multimodal. In addition to the words chosen, prosody, facial expressions, hand and body gestures, and gaze all convey conversational information. For example, Bridwhistell has shown that speech conveys only about one-third of the information in a conversation [4, p. 86-87]. The rest of the information is distributed throughout a number of "non-verbal" semiotic channels, such as hand or facial motions [11]. It has also been shown that non-verbal information is often given a greater weight than spoken information: when the spoken message conflicts with facial expressions, the information from the face tends to dominate [6, 12, 15].

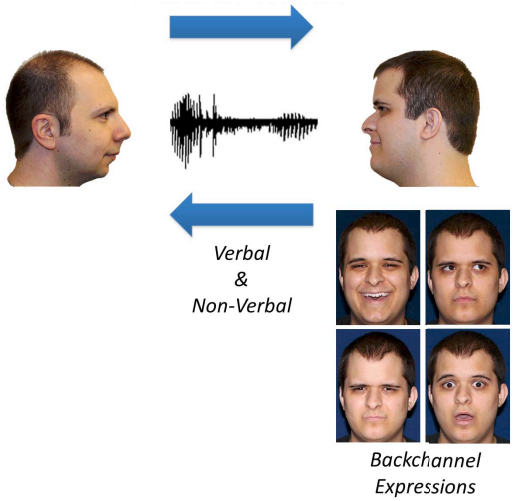


Figure 1. Backchannel signals can have a significant effect on conversational flow. They can be multimodal, including speech (verbal and non-verbal) and facial expressions.

Once real conversations (and not just written texts) are examined, it is clear that listeners are in fact not passive. Listeners provide a lot of feedback to the speaker (see Fig. 1) using what Yngve [21] calls *backchannel communication* (for a recent review, see Xudong, [20]). This feedback can indicate comprehension (e.g., a look of confusion), provide an assessment (e.g., saying "correct"),

control conversational flow or even add new content (e.g., such as sentence completion). For obvious reasons, we use the term frontchannel to refer to the speaker's behavior.

In order to detect backchannel communication, let alone fully model it, it is necessary to have real-world test data. Whilst some conversational databases exist, the general lack of interaction between participants, or poor visibility of the face make these unsuitable for our research. For example, the database in [10] contains pre-defined speaker/listener roles, in [16, 18], the subjects are often too far from the camera for the face to be visible, and in [14], one side of the conversation contains an operator controlled synthesised face. In this paper we present a new multimodal database of natural conversations, designed specifically to allow modelling of front and backchannel elements of dialogues.

### 1.1. Contributions

This paper offers the following contributions:

- The CCDb is a unique, non-scripted, audio-visual natural conversation database where neither participant's role is predefined (i.e. speaker/listener).
- It currently consists of 30 conversations between pairs of participants, equating to 300 minutes of audio-video data.
- Currently, 8 conversations have annotations for facial expressions, verbal and non-verbal utterances, and transcribed speech. Annotation of the remaining 22 conversations is ongoing. This translates to 40 minutes of conversations currently annotated, or 80 minutes if the two sides of the conversation are used independently.
- As well as releasing the data, extracted features, and manual annotations, this paper also offers baseline classification rates for the current set of annotated conversations, obtained using publicly available Support Vector Machine (SVM) software [7].

The data will be of interest to computer vision, affective computing, and cognitive science researchers alike. Furthermore, parts of the database have already been used in experiments to determine the sensitivity to backchannel timing and content [1].

The data is available online as a shared resource for the community, to be used in experiments for developing features, classifiers, and conversational models. To annotate the whole database requires a massive effort, and is time consuming. Therefore, the community's assistance is sought to help complete the task of annotating and validating the remaining conversations. The database can be found at www.cs.cf.ac.uk/CCDb.

## 2. Database

In contrast to several other related databases, the one presented here contains *natural* conversations. While it was collected in a laboratory, the participants had *free rein* to discuss whatever subject they wished. Due to variation in the familiarity of participants with one another, general conversation topics were suggested. However, participants were not required to use them, the conversations were not scripted. Furthermore, the participants did not act in a simulated manner, nor were they prescribed roles to fulfill (i.e. a participant is not given the role of speaker or listener). The conversations were driven by the participant's knowledge (or lack) of the discussion subject, which led to spontaneous behaviour.

Hereafter, *sequence* refers to one-half of a conversation (a single video), whilst *conversation* refers to a pair of sequences.

### 2.1. Recording Equipment

In order to capture natural, spontaneous expressions, the data needed to be captured in as natural a setting as possible. Two audio-video recording systems were set up as shown in Fig. 2. The participants were in the same room and sat opposite each other. To capture *each side* of the conversation the following equipment was used: a 3dMD dynamic scanner captured 3D video, a Basler A312fc firewire CCD camera captured 2D color video at standard video framerate, and a microphone placed in front of the participant, out of view of the camera captured sound (at 44.1KHz). In this paper only the 2D recordings are discussed and used; the 3D system setup and subsequent processing of that data is the subject of future work. To ensure all audio and video could be reliably synchronized, each speaker had a handheld buzzer and LED (light emitting diode) device, used to mark the beginning of each recording session. A single button controlled both devices and simultaneously activated the buzzer and LED. No equipment was altered between the recording sessions, except for the height of the chair to ensure the speaker's head was clearly visible by the cameras.

Figure 2. Setup of recording equipment

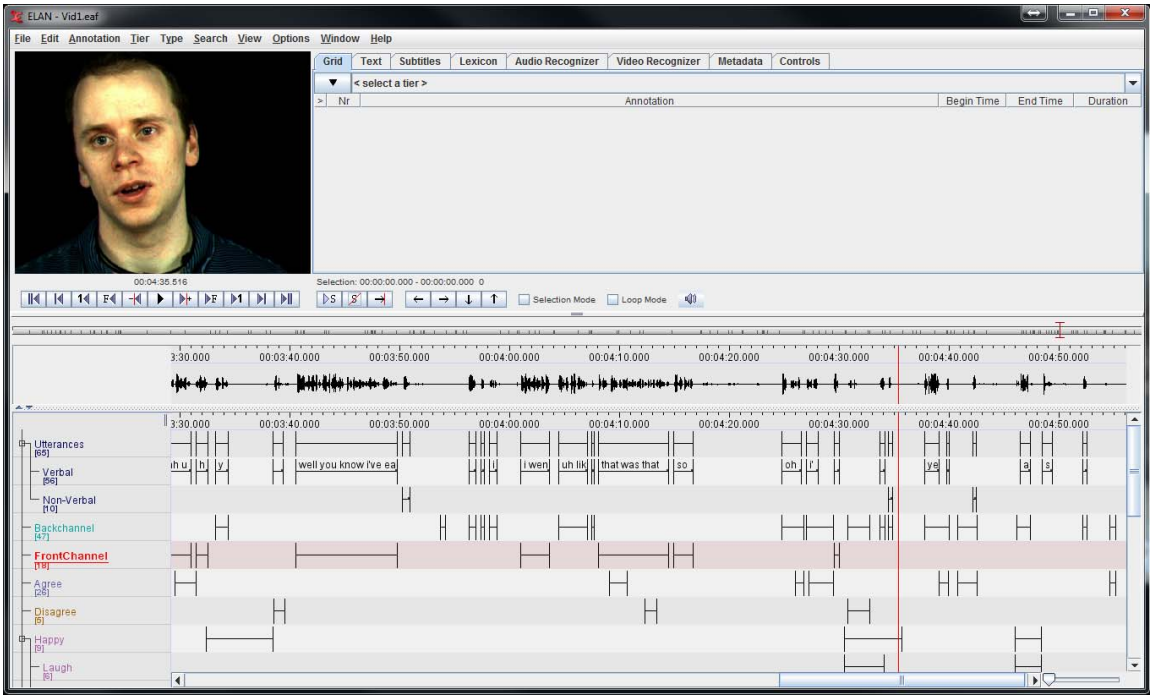


Figure 4. ELAN user interface, showing annotations for a sequence from the database.

Figure 3. View of camera setup during recording.

## 2.2. Recording Method

The full dataset consists of 30 conversations, each containing a pair of speakers, and lasting approximately five minutes. There were 16 speakers in total, 12 male and 4 female, between the ages of 25 and 56 (hence, some speakers were paired multiple times with different conversational partners). The speakers were recruited from within the School of Computer Science at Cardiff University. Prior to the recording session each speaker was asked to fill out a questionnaire. The questions simply required a response on a five point scale from *strongly dislike* (1), to *neutral* (3), to *strongly like* (5), and was aimed at finding out how strongly the speakers felt about possible conversation topics. The questionnaire was used to suggest topics to each pair of speakers for which they had similar or dissimilar ratings, and could if they desired be used as a basis for their conversation. This was done in an attempt to elicit positive or negative expressions. However, it is important to note that the speakers were not restricted to the topics suggested, they could (and did) deviate from the suggestions if desired. Examples of the topics covered in the questionnaire are the like or dislike of different genres of music (rap, opera, jazz, rock etc), literature (poetry, sci-fi, romance, biographies etc), movies, art, sports (rugby, football, ice hockey, golf etc), technology (smartphones, tablets), games, television and current affairs. All participants were fully fluent in the English language.

## 3. Annotations

Manual annotation of the sequences was carried out in ELAN [19]. ELAN is a publicly available, easy to use suite that allows for multiple annotation tracks, hierarchical tracks and also textual annotation on the tracks enabling for example speech sections to be transcribed accurately in time (Figure 4). A variety of facial expressions and gestures were annotated, and the speech for each conversation was fully transcribed. Currently, two annotators have been used for the 16 annotated sequences, (i.e. a single annotator for each sequence). Validation of the annotations was performed by having three people independently annotate two sequences and comparing the results. The annotators were instructed to mark a backchannel as any expression or gesture made in response to verbal or non-verbal action from the other

speaker. These backchannels can occur during or after the action. The annotation tracks we have included so far are (based on those discussed in [17]) :

- **Backchannel.** Expressions, gestures and utterances that would be classified as backchannels.
- **Frontchannel.** Main speaker periods.
- **Agree.** Up and down rigid head motion and/or vocalisation (e.g. 'yeah').
- **Disagree.** Left/right rigid head motion and/or vocalisation (e.g. 'no').
- **Utterance.** The periods of speaker activity, including all verbal and non-verbal activity.
  - **Verbal.** Whole or partial words spoken. Also includes the transcription of the speech.
  - **Non-Verbal.** Both verbal fillers (e.g. 'umm', 'err') and other non-verbal sounds (e.g. 'uh-huh') are identified.
- **Happy.** Smile or laugh.
  - **Smile.** Lip corners move upwards.
  - **Laugh.** Spontaneous smile and sound.
- **Surprise.** Mouth opening and/or raised eyebrows/widening of eyes.
- **Thinking.** Eye gaze goes up and left/right.
- **Confusion.** Slight squint of the eyes, eyebrows move towards each other.
- **Head Nodding.** Up/down rigid head motion. This can be agreement or motion made during speech.
- **Head Shake.** Left/right rigid head motion. This can be disagreement or motion made during speech.
- **Head Tilt.** In plane left/right rotation of the head.

## 4. Baseline Experiments

Whilst the database contains a rich variety of information that researchers investigating and modelling conversational behaviour will find useful, our current focus is on detecting backchannels. To this end we include here some baseline experiments on classifying each sequence into one of three classes, i) backchannel, ii) frontchannel, iii) neutral (neither backchannel or frontchannel). Detecting backchannel signals is not an easy task. While it seems reasonable to assume that prolonged periods of speech are frontchannel, short utterances could be either frontchannel or backchannel. Thus, the length of utterance cannot be used as an unambiguous cue for detecting backchannel signals. Likewise, not all facial expressions are backchannels; many frontchannel communication acts include expressions as well as speech (such as laughing while telling a joke) [5]. Moreover, some facial motions are not in fact meaningful expressions. Thus, the mere presence of facial motion cannot be used an unambiguous cue for backchannel detection. The remainder of this section describes the audio and visual features extracted from the data, and discusses the results of simple experiments. As we are only interested in classifying the frontchannel, backchannel and neutral periods, only these annotations were used as classes for training/testing the SVM.

Figure 5. Example video frames from the database

### 4.1. Audio-Visual Features

A combination of audio and visual features were used to classify the data into frontchannel, backchannel or neither. The visual features were based on shape parameters obtained from an Active Shape Model (ASM) [8]. The ASM tracks landmarks of salient facial features (Figure 6) through the entire sequence. From this data a low-dimensional model (i.e. two) is built that characterises the change in shape over that sequence. The shape parameter for each frame can be described as $I = \hat{I} + \mathbf{P}b$, where $I$ is a vector of landmarks, $\hat{I}$ is a vector of the mean landmarks, $b$ are the shape parameters and $\mathbf{P}$ is a matrix of eigenvectors.

The facial features used here are mouth shape (both inner and outer contour) and global head motion (obtained from the inner eye corners as they are stable points undergoing rigid motion). These facial features are then further processed in the following manner to obtain the features used in classification.

- The first derivatives of the shape parameters are taken.

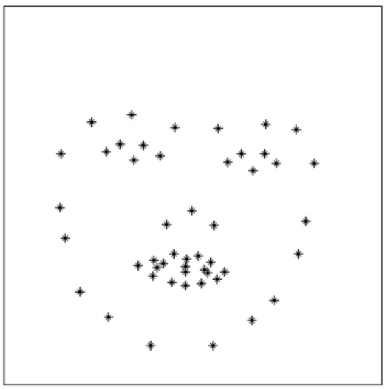

Figure 6. Illustration of facial landmark placement

- Shape features are smoothed at multiple scales. This is useful for highlighting backchannel responses as they can last for several frames (e.g. a quick nod) or to over a second. It also reduces noise in the signals.

The audio features for each frame are extracted from the audio track of the video using PRAAT [2]. They consist of pitch, intensity and the first two formant frequencies ($F_1, F_2$). This is in line with prosody features used in previous backchannel recognition research [13, 16]. The feature vectors obtained from the sequences are then concatenated and used for training and classification.

### 4.2. Baseline Classification Results

The backchannel classification results provided here are intended to serve as a baseline for future research on automatic analysis of conversational behaviour. Support Vector Machines (SVMs) are used to model the audio-visual features pertaining to the 3 classes (backchannel, frontchannel, neutral). Table 1 shows the average proportions of the 3 labels for each of the 16 sequences. Each sequence typically contains between 8-9,000 samples (frames). We define neutral periods to be when a speaker is neither frontchannel nor backchannel. These periods are typically where the listener is paying attention to the speaker, but is not making any expression, gesture or utterance.

| Backchannel | Frontchannel | Neutral |
|---|---|---|
| 22% | 37% | 41% |

Table 1. Average percentage occurrence of each label per sequence.

Two cross validation experiments were carried out on the 16 annotated videos: leave-one-in (LOI) (i.e. choosing one video for the model training and testing on each of the other 15 videos), and leave-one-out (LOO). LibSVM [7] was used with the default parameter settings. For the LOI experiments, each video is used in turn to build the SVM model and to determine the scaling parameters which were applied to both the training data and the remaining unseen testing data. Feature scaling needs to be applied as the range of the individual features vary considerably across sequences. Table 2 shows the mean classification accuracy and standard deviation over all 240 combinations. While the SVM models are capable of performing moderately well on the data the model was trained on (i.e. validating the model) ($\mu = 68\%, \sigma = 6.4\%$), it can be seen that the models generalise very poorly ($LOI_{scaletotrain}$). Rerunning the LOI experiment ($LOI_{scaletoall}$) demonstrates that at least one source of variability is due to different scalings of the data across the videos. Applying a uniform scaling results in a jump in classification accuracy. This requires further investigation.

The third experiment used LOO. It can be seen from Table 2 (computed over all 16 test runs) that this reduces the effect of over-fitting. This suggests that there must be substantial variation in the characteristics of the audio-visual features across the videos. A useful direction would be to acquire appropriate annotated training data per speaker in order to develop better feature normalisation methods.

| Model | $\mu$ | $\sigma$ | Max | Min |
|---|---|---|---|---|
| $LOI_{scaletotrain}$ | 37.2% | 12.7% | 70% | 3.4% |
| $LOI_{scaletoall}$ | 48.8% | 11.3% | 70% | 19% |
| $LOO$ | 57.9% | 7.8% | 70% | 43.4% |

Table 2. Mean and standard deviation of classification accuracy for the two experiments. For $LOI_{scaletotrain}$ the training data is used to determine the scaling parameters. For $LOI_{scaletoall}$ the same scaling is applied to all data.

The results from the second, LOO experiment provide better results than the first experiment but there remains much scope for improvement. There are two main directions for future developments. The first is to extract and develop better features, for example higher level features such as expressions and gestures (video) and phonemes (audio), which are likely to be consistent over different videos and individuals. The second is that temporal dynamics are not currently used in the classification. Obviously dynamics play an important role and their inclusion should improve performance [3, 9].

## 5. Conclusion

This paper has described the first release of the CCDb (Cardiff Conversational Database), a database of natural dyadic conversations, designed to allow the detection, prediction and synthesis of facial backchannel expressions and gestures. The experimental analysis has provided a baseline classification which can be used to benchmark subsequent use of the database. The annotated corpus currently contains approximately 80 minutes of audio-visual conver-

sations between pairs of speakers. The data is annotated with a rich set of features, including speaker segmentation, facial expressions, facial gestures and a transcription of the audio. In addition the audio/visual features are provided. The corpus is publicly available, and can be obtained at www.cs.cf.ac.uk/CCDb. Further annotations of conversations from the corpus will be released over the next year, again including the same annotation labels as used here. We request the assistance of the wider community in annotating the data. We are also currently investigating dynamic modelling of the data for further benchmarking results.

## 6. Acknowledgements

The authors wish to thank the V&L Net and RIVIC (Research Institute of Visual Computing) for their support.

## References

[1] A. J. Aubrey, D. W. Cunningham, D. Marshall, P. L. Rosin, A. Shin, and C. Wallraven. The face speaks: Contextual and temporal sensitivity to backchannel responses. In *ACCV Workshop on Face analysis: The intersection of computer vision and human perception*, 2012. 2

[2] P. Boersma and D. Weenink. Praat: doing phonetics by computer [computer program]. *ACM Transactions on Intelligent Systems and Technology*, 2013. Software available at http://www.praat.org/. 5

[3] K. Bousmalis, L. Morency, and M. Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 746–752, March. 5

[4] R. Bridwhistell. *Kinesics and Context*. University of Pennsylvania Press, Philadelphia, 1970. 1

[5] R. E. Bull and G. Connelly. Body movement and emphasis in speech. *Journal of Nonverbal Behaviour*, 9:169 – 187, 1986. 4

[6] P. Carrera-Levillain and J. Fernandez-Dols. Neutral faces in context: Their emotional meaning and their function. *Journal of Nonverbal Behavior*, 18:281 – 299, 1994. 1

[7] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2, 5

[8] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models: their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, 1995. 4

[9] D. W. Cunningham and C. Wallraven. Dynamic information for the recognition of conversational expressions. *Journal of Vision*, 9(13), 2009. 5

[10] I. de Kok and D. Heylen. The multilis corpus - dealing with individual differences in nonverbal listening behavior. In A. Esposito, R. Martone, V. Müller, and G. Scarpetta, editors, *Third COST 2102 International Training School*, volume 6456 of *Lecture Notes in Computer Science*, pages 362–375, Berlin, 2011. Springer Verlag. ISBN=978-3-642-18184-9. 2

[11] J. de Ruiter, S. Rossignol, L. Vuurpijl, D. Cunningham, and W. J. Levelt. Slot: A research platform for investigating multimodal communication. *Behavior Research Methods, Instruments, & Computers*, 35(3):408–419, 2003. 1

[12] J. Fernandez-Dols, H. Wallbott, and F. Sanchez. Emotion category accessibility and the decoding of emotion from facial expression and context. *Journal of Nonverbal Behavior*, 15:107 – 124, 1991. 1

[13] M. Heldner, J. Edlund, and J. Hirschberg. Pitch similarity in the vicinity of backchannels. In *INTERSPEECH*, pages 3054–3057, 2010. 5

[14] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, Jan.-March. 2

[15] A. Mehrabian and S. Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31:248 – 252, 1967. 1

[16] L.-P. Morency, I. Kok, and J. Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *Proceedings of the 8th international conference on Intelligent Virtual Agents*, IVA '08, pages 176–190, Berlin, Heidelberg, 2008. Springer-Verlag. 2, 5

[17] M. Nusseck, D. W. Cunningham, C. Wallraven, and H. H. Bülthoff. The contribution of different facial regions to the recognition of conversational expressions. *Journal of Vision*, 8(8):1–23, 2008. 4

[18] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–4, Sept. 2

[19] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: a professional framework for multimodality research. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 2006. Software available at http://tla.mpi.nl/tools/tla-tools/elan/. 3

[20] D. Xudong. Listener response. In J.-O. O. Sigurd D'hondt and J. Versucheren, editors, *In The Pragmatics of Interaction*. John Benjamins, 2009. 1

[21] V. Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society.*, pages 567–578, 1970. 1

[22] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009. 1