Predicting Head Pose From Speech

David Greenwood

A thesis presented for the degree of
Doctor of Philosophy

School of Computer Science
University of East Anglia
United Kingdom
April 2018

Predicting Head Pose From Speech

David Greenwood

2018

Abstract

Speech animation, the process of animating a human-like model to give the impression it is talking, most commonly relies on the work of skilled animators, or performance capture. These approaches are time consuming, expensive, and lack the ability to scale. This thesis develops algorithms for content driven speech animation; models that learn visual actions from data without semantic labelling, to predict realistic speech animation from recorded audio.

We achieve these goals by first forming a multi-modal corpus that represents the *style* of speech we want to model; speech that is natural, expressive and prosodic. This allows us to train deep recurrent neural networks to predict compelling animation.

We first develop methods to predict the rigid head pose of a speaker. Predicting the head pose of a speaker from speech is not wholly deterministic, so our methods provide a large variety of plausible head pose trajectories from a single utterance. We then apply our methods to learn how to predict the head pose of the listener while in conversation, using only the voice of the speaker. Finally, we show how to predict the lip sync, facial expression, and rigid head pose of the speaker, simultaneously, solely from speech.

Predicting Head Pose From Speech

David Greenwood

Dedicated to Jayne.

I want to thank...

I first want to thank Barry-John Theobald. Quite simply, Barry allowed me to believe I could do this work, and set me off on the path that brought me to where I am now. When Barry's great talents were recognised elsewhere, Stephen Laycock and Andy Day provided the steady guidance and support I needed, and I warmly thank them both.

I especially want to thank my wife Jayne, and my children Lilly, Tom and Ben. You are the most important people in my world, thank you for giving me the things no one else could, that helped me so much along the way to my goal.

I must thank all the people who I met during the course of this work, and before, who gave insight and advice; you are so many I can't name you all.

Last, but definitely not least, I want to thank Iain Matthews. Without Iain, so much of this work would not have been possible.

Contents

List of Figures

List of Tables

1 Introduction

Speech animation involves transforming and deforming a human-like model, temporally synchronised to an audible utterance, to give the appearance that the model is speaking. The problem is challenging, as human viewers are very sensitive to natural human movement. Practical applications of speech animation, for example computer games and animated films, often rely on motion capture devices or laborious hand animation. Both of these approaches are expensive, time consuming and are unable to scale. Content-driven animation is rich and complex motion derived from easily obtained sparse input, offering appealing properties of lower costs, faster production, and scalability. However to date, there have been few convincing examples.

Human discourse essentially flows in two modes: the explicit mode of audible speech that carries the semantic meaning of some utterance, and a more supportive visual mode where non-verbal gestures complement and enhance the audible mode. Research suggests that speech and gesture may stem from the same internal process and share the same semantic meaning.

There are significant differences between these modes that are pertinent to the work in this thesis. Speech can be processed as a *language*. A language has grammatical rules and structure and has no ambiguity in meaning. Collecting speech data is relatively easy, and processing is fast and efficient on modern hardware.

Non verbal communication, on the other hand, is far less well understood. There have been many efforts to understand the meaning of gesture. Much of this work seeks a taxonomy, for example Birdwhistell [1952] proposes an equivalent to phonemes, dubbed 'kinemes' to describe such action. Ekman and Friesen [1978] develop the Facial Action Coding Sys-

tem(FACS), to account for the motion of the face. Although it appears, with the advent of pocket video cameras, as simple to capture visual speech as audio, the issues extend beyond mere hardware. Notwithstanding any discussion regarding the accuracy of the FACS analogy, to extract the coding from video is non trivial. In fact, any such taxonomy requires human experts to create an annotation, a clear limit to scalability.

The central aim of this thesis is to develop models that can *learn* visual actions from data without semantic labelling, and then, provide compelling speech animation from easily recorded sound.

1.1 Motivation and Research Objective

The pose of the head during speech has interesting properties that present unique modelling challenges. Some activity on the head, most obviously the motion of the lips, the jaw, generally the *orofacial* area have an obvious direct correspondence with speech. Lip accuracy is important, as mismatches between audio and visual can change what a viewer believes they have heard [McGurk and MacDonald, 1976]. Other activity, perhaps the upper facial areas, and notably, the rigid transformations of the skull appear to be independently controlled, yet have been shown to have correspondence with speech, and even contribute to speech comprehension [Munhall et al., 2004].

Increasingly, we find ourselves in a world with great demands on high quality animation, not only for the most obvious use cases within the motion picture and gaming industry, but also for projecting our presence remotely whether on screen, or indeed, immersive virtual reality environments. We also see increasing requirements for high quality animation for therapeutic use.

We are thus motivated by these observations to develop methods and algorithms to model the activity of the head, during speech, *from* speech. Our objectives are methods that can accommodate motion that not only has correspondence with speech, but also motion

that is less clearly connected; motion with non deterministic output that must still appear appropriate and plausible.

1.2 Contributions

Our key contributions to the field are three fold. First, we introduce a data-driven method to predict a diverse range of rigid head pose that dispenses with the idea that learning from speech data requires labelling or categorising the motion. We achieve this by developing a corpus that closely represents the emphatic *style* of speech we find in natural, unrestrained discourse. By recognising that gestures that accompany speech have a many to many relationship, we introduce the Conditional Variational Autoencoder (CVAE) to the task of modelling the rigid pose of the speaker's head. Not only does our method accommodate this difficult task, we gain scalability as our model differentiates the style of multiple speakers.

Secondly, the head pose of the listener in dyadic conversation provides important cues for the speaker. We show that we can directly learn from the voice of the speaker, without resorting to labelling moments of response, or following rule based algorithms, to synthesise these important listener responses.

Finally, the pose of the head is part of a complex bio-mechanical system that may not be best modelled by considering individual components. We develop an algorithm that allows us to predict lip sync, facial expression *and* rigid head pose, directly from speech. The end result is animation that is visually coherent, accurate, convincing and well received by viewers.

1.3 Publications

The following publications have resulted from the work in Chapters 6, 7 and 8:

- Predicting Head Pose from Speech with a Conditional Variational Autoencoder,
  Greenwood, D., Laycock, S., and Matthews, I. In *Proceedings of InterSpeech 2017*, pages 3991–3995.
- Predicting Head Pose in Dyadic Conversation,
  Greenwood, D., Laycock, S., and Matthews, I. In *International Conference on Intelligent Virtual Agents*, pages 160–169.
- Joint Learning of Facial Expression and Head Pose from Speech,
  Greenwood, D., Matthews, I., and Laycock, S. In *Proceedings of InterSpeech 2018*, pages 2484-2488.

1.4 Thesis Outline

This document is structured into three main parts:

In Chapter 2 we first discuss the history of co-speech activity of the head, and why it represents a difficult modelling task, best approached by learning from data.

Chapter 3 describes in detail the development of a multi-modal corpus that we use to learn appropriate speech animation. We describe our methods of speech feature extraction in Chapter 4 and in Chapter 5 we describe the tools we select and develop for modelling from our data.

Chapter 6 describes how we model the rigid head pose of the speaker learnt from data. We go on to describe the modelling of the listener's head pose in dyadic conversation in Chapter 7, and finally we describe how we can predict the full facial expression *and* rigid head pose of the speaker in Chapter 8.

2 Literature Review

This chapter reviews related work in modelling co-speech gesture and expression, focusing on the motion of the head, the expression of the face and the motion of the lips. We also touch on other co-speech activity involving manual gesture as some of the discussion has relevance for head motion. Finally, we close this section with views on the difficulty of modelling human likeness.

2.1 Speaker Head Motion

Speaker head motion is a rather intriguing aspect of visual speech. Head motion has been shown to contribute to speech comprehension, [Munhall et al., 2004] yet, unlike the articulators, it is under independent control. As the speech channel contains the most complete information stream in an utterance, it is a reasonable strategy to seek a mapping from within this stream that might enable plausible predictions of head pose. Indeed, many authors have been motivated by the significant measurable correlation between speech and head motion Deng et al. [2004]; Busso et al. [2005]; Hofer and Shimodaira [2007]; Busso et al. [2007].

When we speak, we encapsulate the semantics of our utterance in the words of our language. We have already stated that rigid head motion is strongly tied to speech, but consider how that occurs. For example, if we are expressing agreement, nodding the head is a common gestural supplement. However, just considering that simple gesture, speaking the same utterance at another time could well have the nodding action at a different phase or frequency. In considering just that simple case, we can appreciate that head pose should be considered as a one to many mapping. And yet there is more to it. When we speak naturally,

we do not issue a monotone dialogue, our voices are highly animated. We use expression, emphasis, intonation or *prosody* to make speech much more than merely words. With that in mind, we must consider that speech to head motion has a very diverse expectation.

It is interesting to consider the variety of approaches to synthesising the movement of the head during speech. Early approaches depend on hand labelling of audio or text input to form rule based systems, although some of these systems extend their output target to facial expression and manual gesture. Later work is inspired by Automatic Speech Recognition (ASR) techniques, particularly the Hidden Markov Model (HMM). More recent approaches use Artificial Neural Networks (ANNs), but we had to wait for the development of Graphics Processor Unit (GPU) processing for this idea to gain traction.

More than twenty years ago Cassell et al. [1994] presented a system claiming that it:

> "...automatically generates and animates conversations between multiple human-like agents with appropriate and synchronised speech, intonation, facial expressions, and hand gestures"

Perhaps we would not recognise the animation as being very lifelike today, but the importance of the synchronisation of speech with emotion, gesture, gaze, facial expression and all other aspects of visual prosody was central to its concept. Using a rule based approach, the system tied together audio speech synthesis, facial animation, based on Ekman's FACS, and a gesture generator using a look-up in a table of predefined gestures.

At the turn of the century, HMMs are the *de facto* standard in ASR applications [Young et al., 1997; Rabiner, 1989]. Many of the methods for ASR are recognised as applicable for the speaker head pose task by a number of authors. Busso et al. [2005] describe a method using a data-driven approach to synthesise appropriate head motion by sampling trained HMMs. Busso et al. [2005] concentrate on the rigid motion of the head citing the observations of Munhall et al. [2004], noting relatively little work has been done in this area to date.

They capture the close temporal relation between head motions and acoustic prosodic features using a bi-gram model trained from multi-modal data, similar to the language models used in speech recognition. The output from HMMs is not continuous, so the head motion must be represented discreetly. For this reason, the Linde-Buzo-Gray Vector Quantization (LBG-VQ) algorithm [Linde et al., 1980] is used to define k discrete head poses.

Busso et al. [2005] quote results using Canonical Correlation Analysis (CCA) of $r \approx 0.8$, although it is worth noting that the data they collect displays similar levels of prosodic correlation, which is regarded as highly dependent on context and speaker [Yehia et al., 2000]. Their approach involves considerable post processing as the model outputs a square waveform.

Hofer and Shimodaira [2007] maintain the HMM approach but add the refinement of modelling head motion from trajectories [Zen et al., 2007]. Of particular note are the correlation analysis results. Hofer and Shimodaira [2007] seek to verify the claims of Busso et al. [2005], but record the correlation within their own corpus as somewhat lower, $r \approx 0.08$. This represents further confirmation of the dependency on context and speaker.

More recently, Ben Youssef et al. [2013] proposed an improved clustering for motion. Whether clustering or quantisation, all of these approaches rely on a suitable labelling of the resulting motion units, either manually or automatically, which is a challenging problem in itself.

Kuratate et al. [1998] presented a paper describing a system that recorded facial motion using opto-electronic tracking to record a speaker's movements. Kuratate et al. [1998] made a sequence of laser scanned polygonal models of the speaker, capturing vowel visemes and non verbal facial poses during speech. Using PCA they decomposed the polygon data and created a parametric representation that could be controlled with the first five principal components. A linear estimator allowed the mesh to be controlled by the 18 tracked markers. Their model is effectively an Active Appearance Model (AAM).

Following on from this work, [Yehia et al., 2000] established figures showing the correlation between head pose and Fundamental Frequency (F0) speech features:

> " The experimental results show that about 80% of the variance observed for F0 can be determined from head motion."

It was noted however that the reverse mapping was far less: 25 to 50%, and further, these figures are highly speaker and utterance dependent.

Building on his earlier work, Kuratate et al. [1999] used the correlation between prosodic audio features and the motion of facial features to drive the speech animated model, the multilinear mapping between different modalities, (speech data and Electromyography (EMG) recordings of facial motion) was transformed using a small ANN of ten sigmoid neurons. Recent advances using the GPU [Bergstra et al., 2010; Jia et al., 2014] permit much larger and expressive models.

Li et al. [2013] argues that predicting head motion is better modelled as a regression problem noting that classification relies not only on the definitions of typical head motion patterns, but also the accurate recognition of these patterns. As well, the relationship between speech and head motion is regarded as a non-deterministic, many-to-many mapping problem.

Ding et al. [2014] propose a method that uses MLPs regression model to understand this relationship and predict Euler angles of nod, yaw and roll. They report advantages over the previous HMM based approaches and were able to avoid the problem of clustering the motion. They develop a corpus derived from news broadcast anchors tracked from video recordings in a studio environment. They make the observation that many face and head tracking methods involve placing markers and using special cameras, that might influence the behaviour of the subject. Ding et al. [2014] refer to the correlation analysis of Busso et al. [2005] but do not report on the correlation within their own corpus. Interestingly, they opt not to use audio features shown to relate to prosody (*e.g.* F0, Energy), instead training their

models using Mel-Frequency Cepstral Coefficients (MFCCs), fBank and Linear Predictor Coefficient (LPC).

Deep Bi-Directional Long Short Term Memory (BLSTM) models appear in Ding et al. [2015], where they report improvements over their own earlier work. More recently Haag and Shimodaira [2016] uses BLSTMs and Bottleneck features [Gehring et al., 2013].

2.2 Listener Head Pose

An avatar is a virtual representation of a human being. Strictly for the definition, an avatar is completely controlled by a human. Bailenson and Blascovich [2004] define an avatar as: "a perceptible digital representation whose behaviours reflect those executed, typically in real time, by a specific human being." Conversely, Embodied Conversational Agents (ECAs) have behaviour that is controlled by computer *algorithms*.

Many studies have shown that ECAs, elicit social behaviour in the human interlocutor [Bailenson et al., 2003; Cassell et al., 2002; Simons et al., 2007] , making ECAs a compelling argument for Human-Computer Interaction (HCI). ECAs allow interaction with machines using communication modalities with lifelong familiarity. These include speech, facial expression and gesture. Importantly, ECAs can also play an important role in Cognitive Behavioral Therapy (CBT) and behavioural study [Lisetti, 2008; Klinger et al., 2005; Dautenhahn and Werry, 2004; Hubal et al., 2008]. To succeed in these domains, ECAs must possess human-like behaviour while speaking and while *listening*.

In face to face communication, human interlocutors provide "back channels". Yngve [1970] introduced the term back channel to describe how "... both the person who has the turn and his partner are simultaneously engaged in both speaking and listening. This is because of the existence of what I call the back channel, over which the person who has the turn receives short messages such as 'yes' and 'uh-huh' without relinquishing the turn." The term implies there are two channels in conversation, the dominant channel of the speaker, and

the secondary channel of the listener. Back channels can be both verbal and non-verbal in nature. Although Yngve's description only concerns turn taking, later research [McCarthy, 2003; Allwood et al., 1992] suggested this linguistic feedback can also convey perception, comprehension, agreement, acceptance and empathy. Back channels include mimicry, that promotes engagement [Bevacqua et al., 2012].

Cassell et al. [2000], argues that the listener head nods could be driven by the speaker's raised voice. Ward and Tsukahara [2000] claimed back channels were, in part, invoked by the speaker's voice, when low pitched periods with specific intervals raised signals. They defined their model with the Algorithm 1.

Algorithm 1: Rule based back channel model

Upon detection of

P1: a region of pitch less than the 26th percentile pitch level and

P2: continuing for at least 100 milliseconds

P3: coming after at least 700 milliseconds of speech,

P4: providing you have not output back channel feedback within the preceding 800 milliseconds,

P5: after 700 milliseconds wait, you should produce back channel feedback.

Maatman et al. [2005] described a system that include posture features detected by a "tracker" as well as audio features from the speaker's voice. They were able to support the claims of Ward and Tsukahara [2000] but reported a need to relax their rules on interval timing.

Morency et al. [2008] introduced an interesting machine learning model using HMMs and Conditional Random Fields (CRFs). They made synchronised multi-modal recordings of a number of dyadic conversations with three video cameras. The videos were annotated by hand to identify visual features of the speaker, and audio features were extracted by

machine. The videos of the listener were also annotated by hand labelling the occurrence of back channels, for example listener nods. Most interestingly, they provide a quantitative evaluation of their own method and compare with the pitch and pause model of Ward and Tsukahara [2000]. They report an F1 score of 0.22 and 0.15 for their method and Ward's respectively.

Bevacqua et al. [2012] describe a sophisticated rule based system that includes a model with personality traits. The back channel timing adheres to the generally held views regarding pitch and timing of the speaker's voice. Their agent has rules that define nods and also facial expression based on FACS action units (AUs) [Hjortsjö, 1969; Ekman and Friesen, 1978]. The back channel type and frequency of the model were controlled by a Listener Intent Planner module. Interestingly, in their evaluation of the personality aspects of the model, they found that the personality type described as aggressive was far more easily identified than, for example, the trait described as pragmatic.

2.3 Facial Animation

The automatic production of realistic speech animation is a long held goal of many areas of graphics, speech and language research, and work extends deeply into the literature. The work often crosses domain boundaries, with authors working in 2D, 3D, capturing performance, and using linguistic and rule based models. The goals also differ, many works are concerned with ECAs that have a role in HCI and Human Machine Interface (HMI), and therapeutic psychology. The other significant research goal targets the entertainment industry, that places weighty demands on realistic high volume speech animation.

Many of the authors concerned with ECAs are interested in speech animation that not only gives the appearance the agent is talking, but that it should also convey personality and emotion. A number of authors pursue a rule based system to model this ambitious goal [Cohen and Massaro, 1993; Cassell et al., 1994] with a predefined set of face shapes. The

expressive, emotional aspects of speech are also modelled with rules, interpolating between categorically labelled emotional states [Cassell et al., 2001; Heylen et al., 2001].

Rule based systems have some advantages in orchestrating the output to suit specific domains, but arguably, data-driven methods have greater scope for modelling the subtle aspects of speech animation, facial expression, and emotion [Cao et al., 2005; Anderson et al., 2013; Li et al., 2016; Sadoughi and Busso, 2017].

Speech animation first appears in the graphics community. Lewis and Parke [1986] describe a lip syncing model based on LPCs to predict *visemes* [Fisher, 1968], the visual counterpart of phonemes. In this early work they acknowledge the importance of other aspects of speech animation, notably head pose, for fully expressive automated character animation. They also remark on the strong perceptual effects that we now refer to as the 'Uncanny Valley' (Section 2.5).

Linguistic based methods to produce plausible facial animation have been developed over several decades [Lewis, 1991; Mattheyses and Verhelst, 2015], either 3D mesh [Wang et al., 2011] or 2D video [Bregler et al., 1997] based. Their common requirement is some form of alignment of the phoneme content either as transcript or by prior processing with external tools [Taylor et al., 2017].

Taylor et al. [2017], described a method for animating the lips of a speaking character by training a Deep Neural Network (DNN) to predict AAM parameters, subsequently retargeted to an animators blend shape rig. The temporal aspect of the speech input utilises an overlapping sliding window processing the audio to a feature vector of phoneme labels. Taylor et al. [2017] claim the overlapping sliding window outperforms Long Short Term Memory (LSTM) networks, and earlier Decision Tree methods of Kim et al. [2015]. They use an interesting feature vector format that includes transition labelling in addition to the phone label.

The complex relationship between co-articulated phonemes and visemes is defined as a many-to-many mapping in the work of Taylor et al. [2012], superseding the static shape-to-shape model of Fisher [1968]. Lip sync is a subset of facial animation that has especially high demands on accuracy, with incorrect lip motion leading to viewer distraction and confusion. It has been shown that mismatch between visual and audio speech can change perceived hearing McGurk and MacDonald [1976].

Data-driven, or machine learning based models that rely only on the input of audio have a similarly lengthy history. Voice Puppetry [Brand, 1999] is a notable example that uses HMMs for trajectory sampling. Most recently Suwajanakorn et al. [2017] use a regression model of LSTM [Hochreiter and Schmidhuber, 1997] networks to produce highly plausible 2D lip animation. Karras et al. [2017] employ a deep neural network combining fully connected layers and Convolutional Neural Networks (CNNs) to model facial animation with emotional content.

The recent work by Taylor et al. [2017], Suwajanakorn et al. [2017] and Karras et al. [2017], arguably represent the state of the art for data driven facial animation, and these three works appeared in the literature at the same time. These three works all used a deep learning approach, but remarkably diverse architecture. The 2.5D video model of Suwajanakorn et al. [2017] uses a deep LSTM to accurately replace the lip motion of Barack Obama, Karras et al. [2017] use CNN to drive a canonical mesh, and Taylor et al. [2017] use a deep MLP to drive an AAM. Interestingly, the latter two 3D methods had hand animated head pose applied to reduce perceptual dissonance when presented.

2.4 Gesture

What do we mean by gesture? In the context of this work we are interested in the function of gesture in parallel with speech. That is gesture that occurs at the same time as speech, rather than any gesture that is intended to replace speech such as the emblematic gesture

described by Ekman and Friesen [1969]. In the same way, we exclude gestures that occur when a speaker is performing lexical search, and gestures that occur when speech fails [Butterworth and Beattie, 1978], as clearly, to synthesise animation to accompany speech, moments of silence present additional challenges. Of course, we also remove from consideration language transmitted as hand signals, for example sign language used by deaf people or descriptive gestures used by people in noisy environments.

2.4.1 The Function of Gesture

People gesture spontaneously during speech, and a considerable body of evidence shows that gesture seems to support and expand the audible mode and is closely related to many aspects of language. There is far less agreement of the *function* of gesture. Some research claims that gesture is a post speech process, a translation of speech, in language production [Butterworth and Hadar, 1989]. Other research [Kendon, 1972; McNeill, 1992], suggests both gesture and speech stem from the same underlying propositional representation that has both visual and linguistic aspects. The claim is that speech and gesture work together to convey meaning.

Goldin-Meadow et al. [1996] discusses the role of gesture as a language substitute. They observe that when gesture exists as the sole modality, it assumes grammatical properties of human language, particularly segmentation and hierarchical combination. During co-speech gesturing, gesture and speech synchronise grammatically forming a unified linguistic system.

2.4.2 Gesture Taxonomy

Ekman and others [Ekman and Friesen, 1969; Kendon, 1983], categorised gestures covering a wide range of phenomena. Ekman's categories included many of the gesture types outside of the scope of this work, and his categories bridged descriptions by other authors. We

are interested in co-speech gestures that may serve a function similar to speech or support speech. McNeill [1992] described four types of gesture shown to occur during narrative discourse.

Iconic Gestures

Iconic gestures are closely related to speech and support an action or event that is being described; for example "he climbed the ladder" accompanied by the hand rising upwards.

Metaphoric Gestures

Metaphoric gestures also illustrate actions or events, but explain abstract concepts that may not have a physical form. Such gestures describe for example, the passing of time, or symbolise complexity.

Deictic Gestures

Deictic gestures describe physical space and direction, in relation to the speaker. An example might be; "Alice looked at Bob, and he looked back...", with a hand pointing first left then right. McCullough [1992] claims deictic gestures accurately describe space and direction.

Beat Gestures

Beat gestures are small rhythmic gestures that do not change in form with the content of the speech. Beats can be a single staccato strike, or a repetition maintained for as long as necessary to make a particular point. Beats are used to add emphasis and stress to the spoken utterance.

2.5 The Uncanny Valley

The Uncanny Valley is a hypothesis that suggests that as human models look and behave closer to real human beings, human observers have increasing feelings of revulsion. The level of acceptance dips considerably as realism within the model increases, before returning to high levels when the model is completely lifelike. Figure 2.1 illustrates the phenomenon.

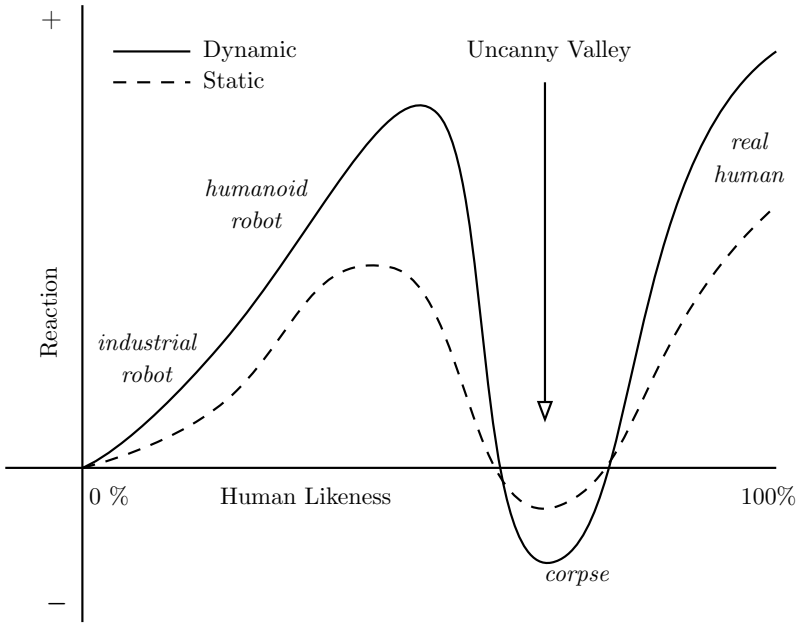


Figure 2.1: Mori's original hypothesis states: "That as the appearance of a robot is made more human, a human observer's emotional response to the robot will become increasingly positive and empathic, until a point is reached beyond which the response quickly becomes that of strong revulsion. However, as the robot's appearance continues to become less distinguishable from that of a human being, the emotional response becomes positive once more and approaches human-to-human empathy levels." [Mori, 1970]

The term was first used by the robotics professor Masahiro Mori [Mori, 1970], although a 1906 work by Ernst Jentsch [Jentsch, 1997] referred to the concept, and in an essay on the nature of the 'uncanny' from 1919, Freud [1955] describes his extreme discomfort at seeing someone wearing a prosthetic limb.

There is considerable anecdotal evidence for the Uncanny Valley from animation, robotics and art works, but this does not in itself support the valley model. Brenton points out that, in the Graph 2.1, an asymptote goes through the first third and another through the last third. The middle section is referred to as a valley, but only because it has been drawn that way. It could also be represented as a stepped discontinuity [Brenton et al., 2005].

Gathering more empirical evidence, Saygin et al. [2010] performed a Functional Magnetic Resonance Imaging (fMRI) study of the perception of human and artificial agents. Participants were shown videos of body movements performed by a realistic android, the same movements performed by the human actor the android was modelled on, and finally the android again but with its synthetic skin removed exposing its mechanism. Considerable brain activity was recorded in areas sensitive to body movements (anterior intraparietal cortex).

> "We interpret these results within the framework of predictive coding and suggest that the uncanny valley phenomenon may have its roots in processing conflicts within the brain's action perception system."

Numerous highly detailed animations have been achieved e.g. [Borshukov et al., 2005; Alexander et al., 2009], often using motion tracked data from multiple cameras. Claims have been made suggesting some of these examples are indistinguishable from human performances, and although impressive, many people are not convinced of those claims. Certainly, not all animation aims to be realistic in the sense of human characters. Many animations are of non-human characters. Often greatly exaggerated expressions are performed, but are accepted by viewers because the rules governing their movement is based on realistic sub-structure.

One interesting report from Moore [2012] offers a "Bayesian Explanation" of the Uncanny Valley. He argues the disparity of findings within the literature may in part be related to the semantics of the original Japanese terminology. He presents a Bayesian model of categorical perception, an extension to Feldman's model [Feldman et al., 2009], to account for differential perceptual distortion across multiple cues. He claims the model is the first quantitative explanation of the Uncanny Valley.

3 Corpus

We believe clean, unbiased data is an essential part of supervised learning and, in the absence of readily available multi-media corpora, we develop our own corpus. In this chapter we describe the data collection process, from the recording of audio and video, tracking, and the process of converting raw audio and video footage into useful data for training subsequent models.

3.1 Actors

We hired two actors, one female (Subject A, Amanda, shown in Figure 3.1a), one male (Subject B, Joshua, shown in Figure 3.1b) to recite from a scripted set of short conversational scenarios. The actors were encouraged to speak emotively and emphatically in order to provide natural, expressive and prosodic speech.

3.2 Video

The cameras, Sony®PMW-EX3 XDCAM HD422, were arranged in a radial pattern (Figure 3.2), such that three cameras were aimed at each individual actor. A central camera gave the frontal view, and a left and right camera at approximately 45 degrees off the centre axis provided generous image and landmark correspondence for both the left-centre and right-centre stereo pairs. The focal plane of the cameras were approximately two metres from the subject, giving a natural perspective, and minimising distortion.

The recordings were made simultaneously for both actors, so in all six cameras were synchronised. The actors were seated, back to back, and a beam splitter for prompts also

(a) Subject A, Amanda **(b)** Subject B, Joshua

Figure 3.1: Our two actors, showing the arrangement of marked landmark sites on their faces that were tracked, in combination with the natural sites around the lips, eyes, brows and nose, using AAMs.

displayed their interlocutor. Figure 3.2 illustrates this arrangement. The cameras were oriented to portrait to maximise the recording area. The video resolution is 1280 × 720, with a sample rate of 59.94 Frames per Second (FPS). We used SMPTE time code to maintain synchronisation during subsequent edits of the footage. Figures 3.3 and 3.4 illustrate the view of each camera triplet.

3.3 Audio

Audio was recorded with two Audio-Technica®AT899 Subminiature Omnidirectional Condenser Lavalier Microphones. The recording sample rate was 44.1 KHz/16 bit, and was later down sampled to 16 KHz. The participants were invited to recite a short conversational scenario. For example, Speaker A would make the utterance:

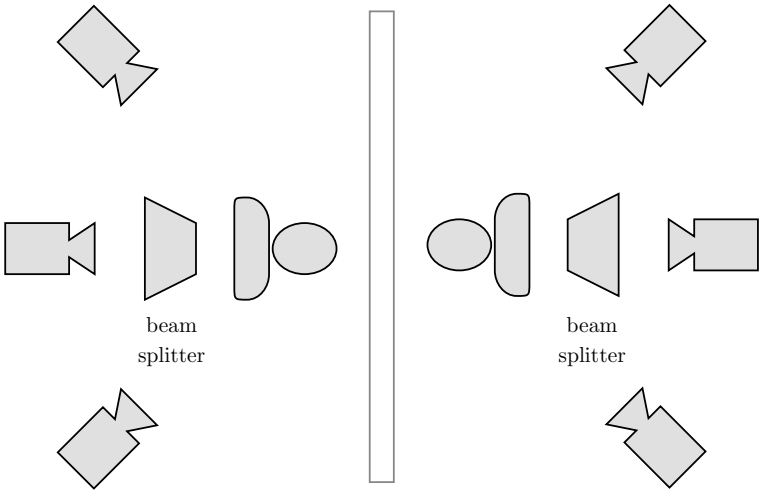


Figure 3.2: Cameras are arranged in a radial pattern to give generous correspondence across three images. Distance from subject is around two metres to give a natural perspective. The two subjects are seated back to back but separated by a green curtain. They can see each other by virtue of the centre camera being projected to the counterpart beam splitter, that also displays the text prompts.

Figure 3.3: Three cameras synchronised to give multi-view stereo. Here we show a combined frame from the three cameras, we can see the degree of correspondence in the landmarks across each view for Subject A.

Figure 3.4: Subject B has three synchronised cameras. When we show the three images together, we can see the correspondence across the three views.

"You have won a set of your very own cupcake tins!"

Speaker B would respond with:

"I am the happiest baker on the planet right now!"

Then, a further utterance by A:

"And I am thrilled to announce that as a bonus prize, you will also receive a year's worth of batter."

Finally, the response from B:

"I will make the most amazing red velvet cupcakes the world has ever seen!"

This A to B, A to B exchange represents a complete *vignette*, of which 314 were recorded. Each vignette was enacted, sometimes twice, with three speaking styles, and the actors exchanged A and B roles. In all, around 3600 utterances were captured, giving a total of near six hours of speech. Care was taken that the dialogue accurately reflected the script, to avoid the necessity to annotate to word level by hand. In addition to the expressive speech, a set of neutral statements was recorded, made by each actor.

3.4 Tracking

We tracked the landmarks in all the camera views using AAMs [Cootes et al., 2001], trained on a selected set of extreme poses. The training data was hand annotated by marking each landmark position, on each selected frame, for each camera view, for each actor. To be clear, we created a unique model for every video stream.

3.4.1 Active Appearance Models

AAMs may refer to both models, or, models along with their fitting algorithms. They are a deformable tracking method for modelling photo-realistic appearance, used successfully for whole or part faces, medical imaging and other diverse applications [Edwards et al., 1998; Cootes and Taylor, 2001; Baker and Matthews, 2001; Matthews and Baker, 2004; Trutoiu et al., 2011; van der Maaten and Hendriks, 2012]. Building an AAM requires a set of n annotated images that have m *corresponding* landmarks identified on each image:

$$\boldsymbol{l}_n = (x_{n,1}, y_{n,1}, \ \ldots, \ x_{n,m}, y_{n,m}) \tag{3.1}$$

In our own example, we combined natural landmarks with marked sites on the faces of our subjects. The points in the training data are aligned using Procrustes Analysis, to give shapes:

$$\boldsymbol{s}_1, \ \ldots, \ \boldsymbol{s}_n \tag{3.2}$$

The variation in the shapes is parametrised by forming a matrix S of all the shapes less the *mean* shape and performing PCA:

$$\begin{aligned} \bar{\boldsymbol{s}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_i \\ \boldsymbol{S} &= (\boldsymbol{s}_1 - \bar{\boldsymbol{s}}, \ \ldots, \ \boldsymbol{s}_n - \bar{\boldsymbol{s}}) \end{aligned} \tag{3.3}$$

Now any shape $\boldsymbol{s}$ can be represented by a set of parameters $\boldsymbol{b}_s$, the shape eigenvectors P_s and the mean shape $\bar{\boldsymbol{s}}$ with:

$$\boldsymbol{s} = \bar{\boldsymbol{s}} + \boldsymbol{P}_s \boldsymbol{b}_s \tag{3.4}$$

We now need to parametrise the pixel data. Each image within the triangulated bounds of the shape model is warped to the mean shape, producing a set of image vectors that vary in *appearance*, but not shape.

$$\boldsymbol{a}_1, \ \ldots, \ \boldsymbol{a}_n \tag{3.5}$$

In a similar way to the shape, appearance variation is parametrised using PCA. Again, a matrix $\boldsymbol{A}$ is formed by concatenating all the vectorised appearance vectors, less the mean appearance.

$$\begin{aligned} \bar{\boldsymbol{a}} &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{a}_i \\ \boldsymbol{A} &= (\boldsymbol{a}_1 - \bar{\boldsymbol{a}}, \ \ldots, \ \boldsymbol{a}_n - \bar{\boldsymbol{a}}) \end{aligned} \tag{3.6}$$

Now an appearance $\boldsymbol{a}$ can be formed with:

$$\boldsymbol{a} = \bar{\boldsymbol{a}} + \boldsymbol{P}_a \boldsymbol{b}_a \tag{3.7}$$

where $\boldsymbol{P}_a$ are the appearance eigenvectors and $\boldsymbol{b}_a$ are appearance parameters, and $\bar{\boldsymbol{a}}$ is the mean appearance. Our model retained independent shape and appearance, and used the *project out*, inverse composition, and gradient descent for fitting. Our tracking model is described in detail by Matthews and Baker [2004] and dubbed a Flexible Appearance Model (FAM).

3.4.2 Model Fitting

With our trained AAMs, we made a first pass at fitting the model to every frame of video. Typically, we would find frames with poor fitting, correct these by hand, add them to the

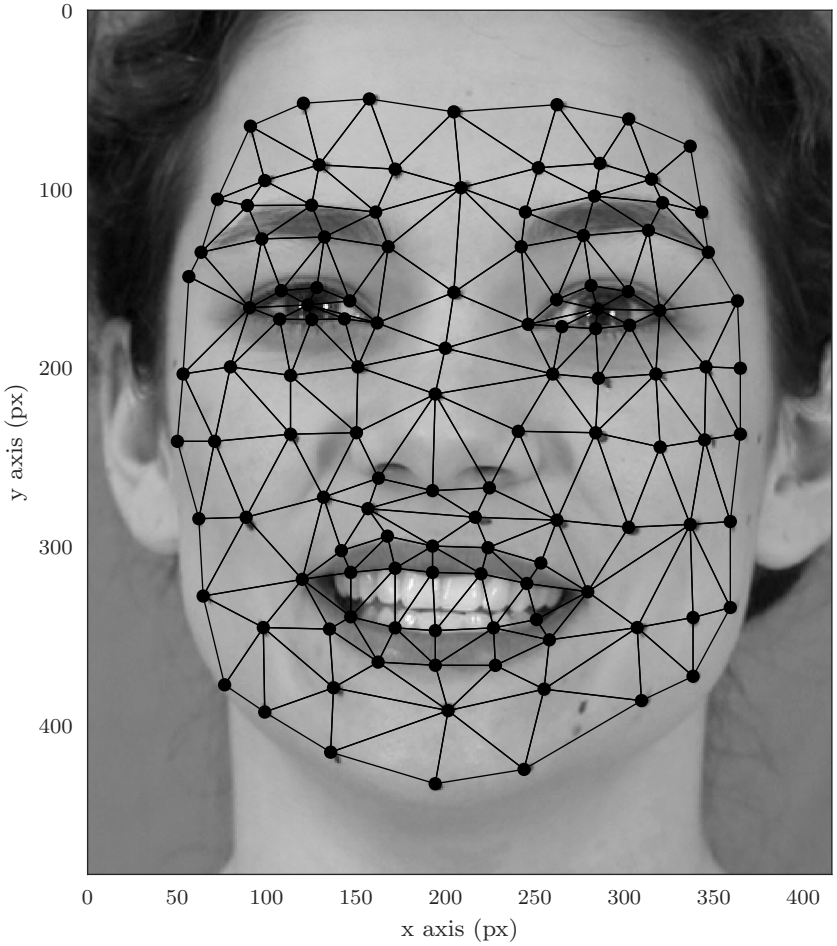


Figure 3.5: PDM for Speaker A, centre camera. The PDM is used, along with the pixel appearance, for our AAM face tracking. Here we show the landmarks and the triangulation that defines the area of the image used for the appearance model.

training data and rebuild the model. After a handful of iterations, we became confident that the model was adequately trained. We took care to ensure that the model tracked the facial features for as long as possible, but inevitably, at moments when the subject's face was occluded, or left the frame, tracking would fail. At this point the tracking was stopped, returned to a point where the model could converge, and resumed. The model fitting rate was in the order of 10 FPS, and unfortunately required monitoring due to the aforementioned tracking failures. A future improvement would be a face detection algorithm, allowing automatic skipping of occluded or missing faces, with the hope that the process could proceed entirely automatically.

3.5 Camera Calibration

Camera calibration estimates parameters of a camera lens and image system. These parameters include intrinsic, extrinsic and distortion coefficients. We used the method proposed by Jean-Yves Bouguet [Bouguet, 2002]. The method is a *photogrammetric* calibration, where calibration is performed by recording images of a calibration object [Zhang, 1999].

3.5.1 Pinhole Camera Model

The 3×4 camera projection matrix $\boldsymbol{P}$ describes a mapping of 3D world points to 2D points in the image plane. If the world and image points are expressed in homogeneous coordinates, the perspective projection can be shown as a matrix multiplication:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{3.8}$$

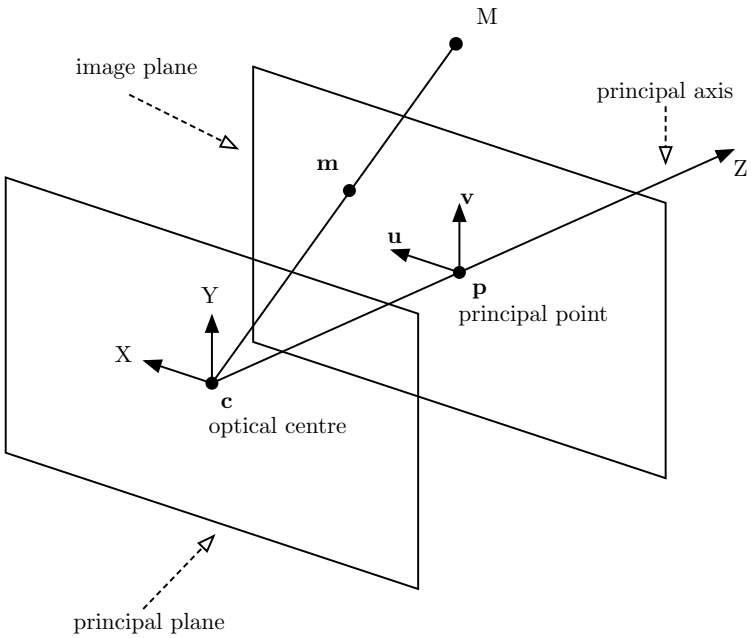


Figure 3.6: Illustration of an ideal pinhole camera.

This can be written compactly:

$$\boldsymbol{m} = \boldsymbol{P}\boldsymbol{M} \tag{3.9}$$

This equation holds only for the special case of the camera at the world origin and only has information about the focal length. More generally the camera will be transformed by some rotation and translation, and we need to take into account pixel aspect, skew and optical centre offsets. We can consider the projection matrix $\boldsymbol{P}$, as the product of the *intrinsic* matrix, $\boldsymbol{K}$ multiplied by the augmented matrix $[\boldsymbol{R} \mid \boldsymbol{t}]$; the *extrinsic* parameters.

$$\boldsymbol{P} = \boldsymbol{K}[\boldsymbol{R} \mid \boldsymbol{t}] \tag{3.10}$$

$$\boldsymbol{K} = \begin{bmatrix} \alpha_u & s & u_0 \\ 0 & \alpha_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.11}$$

with the five intrinsic parameters:

- α_u is the scale factor in the u-coordinate direction
- α_v is the scale factor in the v-coordinate direction
- s is the skew
- $(u_0, v_0)^T$ are the coordinates of the principal point.

and where the extrinsic parameters:

- $\boldsymbol{R}$ is a rotation matrix
- $\boldsymbol{t}$ is a column vector, translating the optical centre.

A detailed explanation of the perspective camera can be found in [Hartley and Zisserman, 2004, chap. 6].

3.5.2 Distortion Parameters

The pinhole camera is an ideal model. Real world examples of cameras usually have some amount of *radial* (Equation 3.12) and *tangential* (Equation 3.13) distortion giving the coefficients shown in Equation 3.14.

$$\begin{aligned} x_{distorted} &= x(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \\ y_{distorted} &= y(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) \end{aligned} \tag{3.12}$$

$$\begin{aligned} x_{distorted} &= x + [2p_1 xy + p_2(r^2 + 2x^2)] \\ y_{distorted} &= y + [p_1(r^2 + 2y^2) + 2p_2 xy] \end{aligned} \tag{3.13}$$

$$\text{Distortion coefficients :} \quad k_1 \quad k_2 \quad p_1 \quad p_2 \quad k_3 \tag{3.14}$$

We used the Camera Calibration Toolkit that is available in MATLAB® and also in Open Source Computer Vision Library (OpenCV) [Bradski, 2000]. This provides functions for calculating these coefficients, along with the intrinsic and extrinsic matrices. Figure 3.7 shows our camera calibration target, a chequerboard of 10×6 20mm squares on a rigid

plane. We recorded a number of frames of video, while moving the calibration target within the bounds of the camera field of view. We made sure to include in plane rotations of the target, as well as the other two axes, while taking care that the target was wholly visible in all three cameras at once. Calibration of an individual camera does not require calibration images across multiple views, we need this correspondence for stereo triangulation, described in Section 3.6. We note that a considerable excess of target images were required, as depth of field and motion blur reduced sharpness of some images, requiring rejection. Another artefact we did not anticipate was the degree of specular reflection of the black squares at certain angles of the target, and in these cases the Camera Calibration Toolkit would fail to find the target corners.

3.6 Stereo Triangulation

Epipolar geometry is a property of two views that has no dependence on scene structure and only depends on the intrinsic and extrinsic properties of the cameras [Hartley and Zisserman, 2004, chap. 9]. An image point x back-projects to a ray l in world space defined by the first camera centre, C, and x. This ray is a line, the *epipolar* line l', in the second image plane. Therefore X must lie on l' (Figure 3.8) [Hartley and Zisserman, 2004, chap. 9].

The Fundamental Matrix $\boldsymbol{F}$ maps a point in one image to a point in the other image and satisfies :

$$\boldsymbol{x'}^T \boldsymbol{F} \boldsymbol{x} \tag{3.15}$$

The Essential Matrix $\boldsymbol{E}$ describes the location of the second camera relative to the first in global coordinates.

We have a number of corresponding points in our calibration images. With the cameras previously calibrated, we again use the MATLAB® Camera Calibration Toolkit to form a stereo pair for each of our left-centre and centre-right pairs.

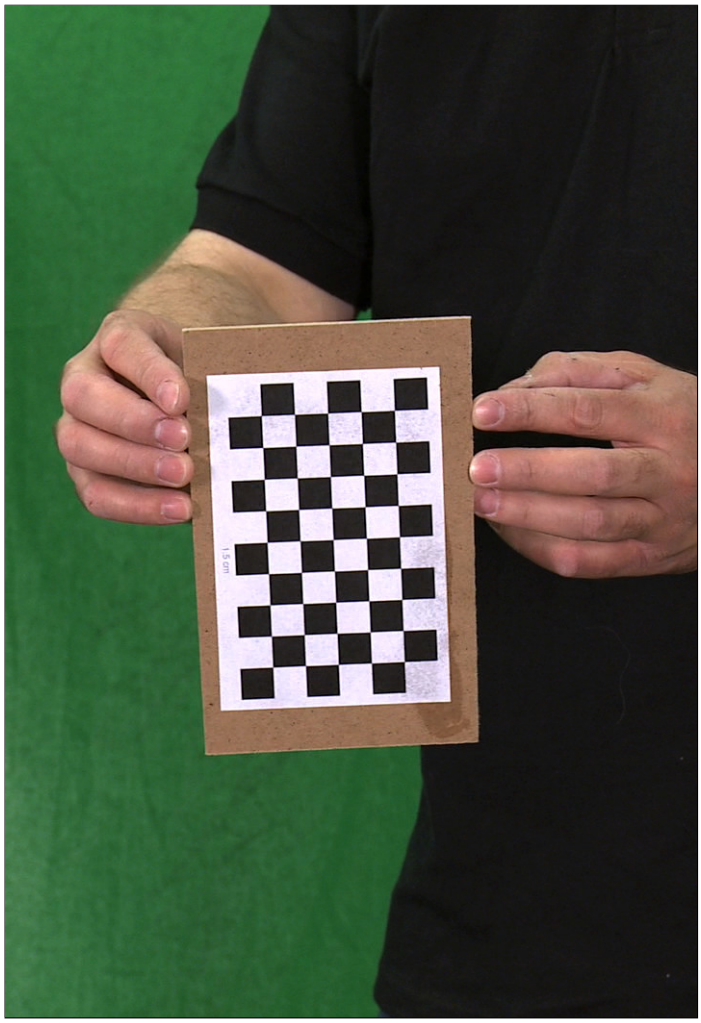


Figure 3.7: Cameras were calibrated using a chequerboard pattern, visible in all three cameras simultaneously. We recorded several seconds of video, while moving the calibration target within the bounds of the camera field of view.

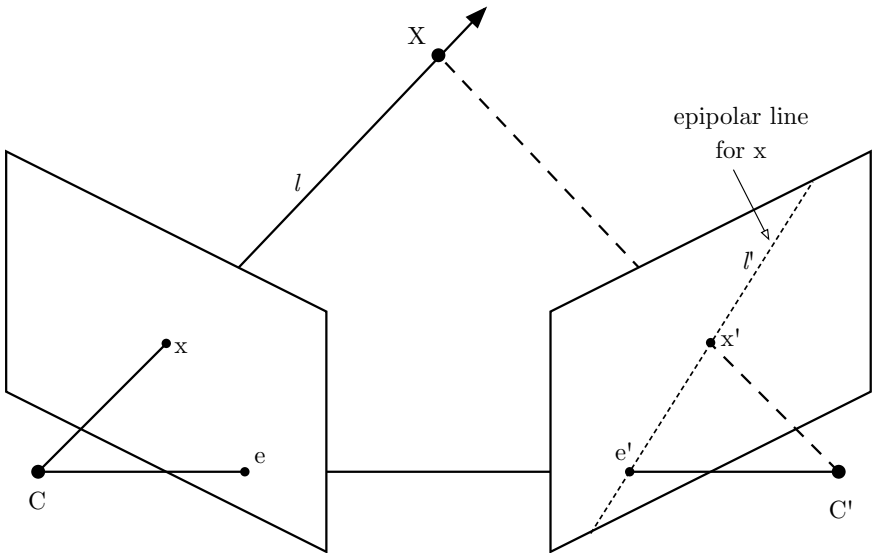


Figure 3.8: Epipolar geometry. An image point x back-projects to a ray l in world space defined by the first camera centre, C, and x. This ray is a line, l' in the second image plane, so X must lie on l' as image point x'.

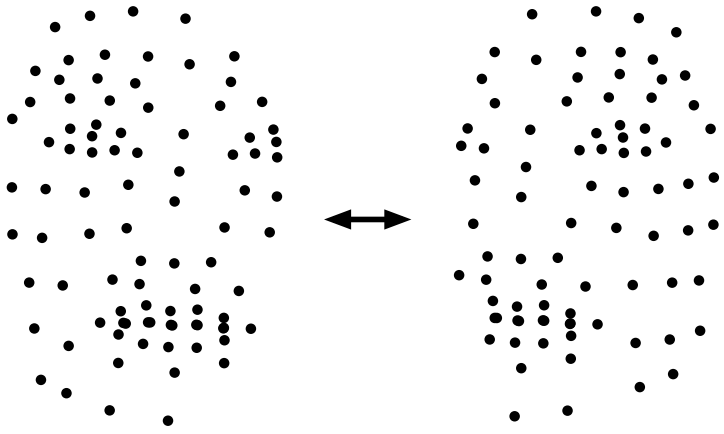


Figure 3.9: Two stereo pairs allow the triangulation of two sets of 3D points. The two sets of points are merged to one to form our complete shape model.

3.7 Data Processing

Now that we have recorded our data set, there remains a process of annotation and editing before we can use it to train predictive models. We will discuss parametrisation in the next chapter, so for now we can concentrate on removing noisy data, and preprocessing the tracking data.

3.7.1 Annotation

The audio was annotated using the PRAAT software package [Boersma and Weenik, 1996]. Each utterance was isolated, and any incorrect statements, social chat, cross talk or involuntary sounds marked for removal. The timing of the trimmed sections of audio were used to define the in-out points of the tracked visual features. This was important, as moments of social chat and so on often had high levels of visual occlusion and subsequently experienced tracking failures. Excluding such extraneous material minimised the number of sequences that had to be manually removed after segmentation due to tracking inaccuracies.

3.7.2 Shape Model

For each subject we have defined a shape model $\boldsymbol{X}$ described by a set of m three dimensional points shown in Equation 3.16.

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & x_{m3} \end{bmatrix} \tag{3.16}$$

3.7.3 Separation of Deformation and Transformation

The separation of rigid and non-rigid motion is an active research area itself, Nonrigid Structure from Motion (NRSfM) [Black and Yacoob, 1995; Dai et al., 2014; Ramakrishna et al., 2012]. However we can take advantage of some observations made of our data. Although all the landmark sites are free to deform, it is clear some sites deform far less than others. Areas around the mouth, for example, are highly deforming. Much less so, are the sites at the inner eye and bridge of the nose. Given this observation we are able to reduce the problem to one of Structure from Motion (SfM), a very well understood problem.

We examine the video sequences and select t consecutive tracked frames of our vectorised shape model, when the facial expression is neutral. We take the mean over time to eliminate any minor fluctuations (Equation 3.17).

$$\boldsymbol{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{t1} & s_{t2} & \dots & s_{tn} \end{bmatrix} \tag{3.17}$$

$$\bar{\boldsymbol{S}} = \frac{1}{t} \sum_{i=1}^{t} s_{ij}$$

We then take the mean of each 3D landmark, subtract from the mean shape, to place our model at the origin. We now designate this as a neutral expression.

$$\begin{aligned} \bar{\boldsymbol{X}} &= \text{vec}(\bar{\boldsymbol{S}})^{-1}_{m,3} \\ \boldsymbol{X}_{neutral} &= \bar{\boldsymbol{X}} - \frac{1}{m} \sum_{i=1}^{m} \bar{\boldsymbol{X}}_{ij} \end{aligned} \tag{3.18}$$

The intention here is to create a reference pose and, by using Ordinary Procrustes Analysis (OPA), we can find the translation and rotation of all recorded head pose frames relative

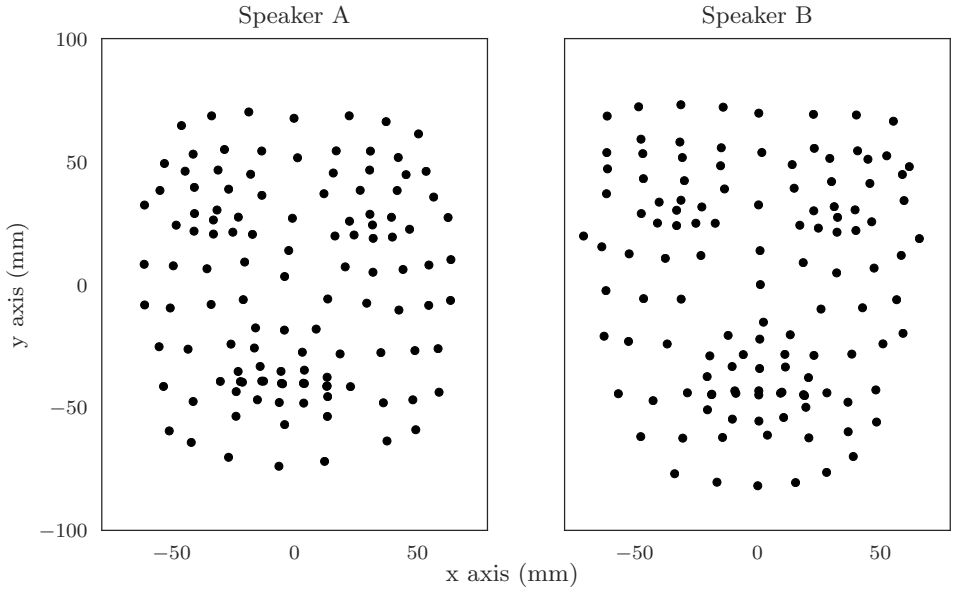


Figure 3.10: A neutral pose for each speaker, derived from the temporal mean of 10 frames of tracked video. Note, this figure shows the 3D model projected to 2D on the z axis.

to this reference. Note, this is not Generalized Procrustes Analysis (GPA), Gower [1975], which seeks to find the optimal alignment of a population of $n \times m$ data. To perform OPA, we first translate all the shapes to the origin. The Procrustes algorithm usually considers scale, but we know that in our case scale does not change, so we can ignore this step.

The rotation matrix uniquely describes a rotation in $\mathbb{R}^3$. Due to the anatomical and environmental limits on head pose in our data set, we choose to use Euler angles to describe head pose, principally because the change of Euler angle over time will be a differentiable value. Given 3 Euler angles ψ, θ, ϕ, the rotation matrix $\boldsymbol{R}$ is the product of the rotation matrix about each axis, shown in Equations (3.19, 3.20).

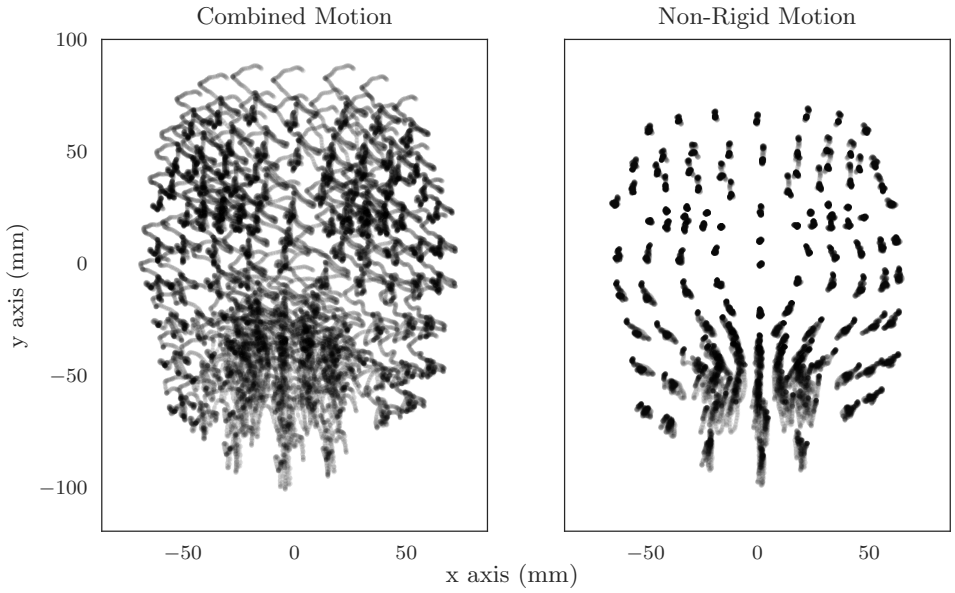


Figure 3.11: Using Procrustes Analysis, we separate the rigid and non-rigid motion of Speaker A. Note the stability of the landmarks in the vicinity of the inner eye and at the bridge of the nose. These were chosen as the least deforming points for alignment process.

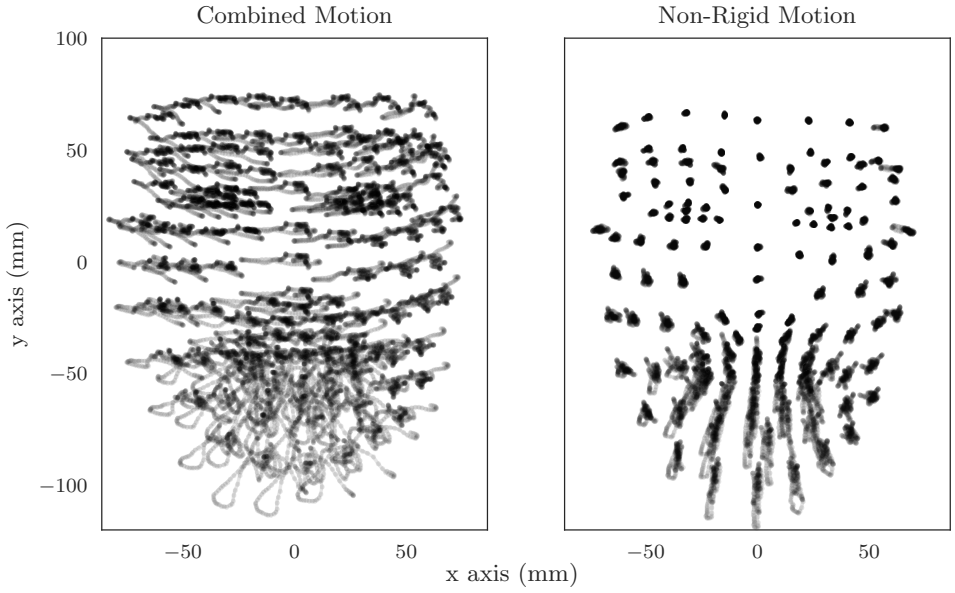


Figure 3.12: Here we plot the motion of the landmarks on Speaker B. The left hand figure shows the motion before we separate the rigid transformations from the non rigid deformations.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi & \cos\psi \end{bmatrix}$$

$$\boldsymbol{Y} = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \tag{3.19}$$

$$\boldsymbol{Z} = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\boldsymbol{R} = \boldsymbol{XYZ} \tag{3.20}$$

The inverse of the rotation matrix is ambiguous, but in our case we know we will not experience gimbal lock, or rotation angles greater than 90 degrees, so can be derived by Equation 3.21. We have now separated *deformations* and *transformations* in our shape model.

$$\begin{aligned} \psi &= \quad arctan(r_{32}/r_{33}) \\ \theta &= -arcsin(r_{31}) \\ \phi &= \quad arctan(r_{21}/r_{11}) \end{aligned} \tag{3.21}$$

3.8 Remarks on Data Collection

In this final section of this chapter, it is worth making some remarks about the data collection process. It can not be over stated how difficult it is to collect clean, unbiased multi-modal data. Great effort was taken to achieve this aim, yet, if the task were to be presented again, some mistakes could be corrected. Some of these errors, and there may be more as we continue to explore the data, are mentioned now.

Take One

We had an earlier data collection, collecting full body motion using a Microsoft Kinect® structured light depth camera and head pose with a Sony PS3® motion controller. Unfortunately the data was simply too noisy to train meaningful models, further highlighting the difficulty in collecting such data.

Silence Model

As we collected the data, the goal was to explore head pose variation during speech. We did not fully consider the starting pose, and how we arrive at the starting pose from silence, or another mode such as listening. We would very much like to include this and what one might describe as ambient motion in any future data collection.

Uniform Landmarking

Although the marked face pattern was intended to be the same, when tracing Subject A's eyebrows 7 points seemed adequate, but once we processed Subject B, we needed 8 points to accommodate the more angular brow pattern. Also, different levels of occlusion across our two subjects while recording, resulted in the final combined 3D point model having dissimilar numbers of points. We did not realise the significance of this difference while

developing models of rigid head pose prediction, but parametrising the face motion (in Chapter 8) required each subject to be considered separately.

3.8.1 FutureWork

A probable future direction, even for the data collected for this corpus, is to choose a different tracking method. Recent improvements to mesh fitting algorithms [McDonagh et al., 2016; Laine et al., 2017] for performance capture, present a convincing argument for doing this, and may permit tracking of the off script, social chat regions of our capture.

4 Feature Extraction

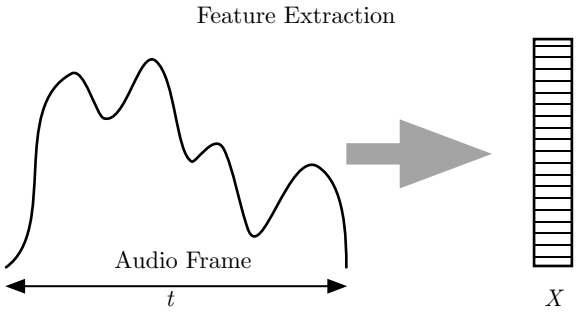


Figure 4.1: Feature extraction. Speech features are extracted by taking a short duration $t = 2/59.94$ s of audio and transforming to a feature vector X.

The key aim of the work is to synthesise speech animation from the recorded speech signal. As such, extracting useful information from the speech signal is a primary requirement. On hearing speech, considerable information is available to the listener. It is possible to distinguish gender, age or identify an individual. Depending on context, one can gather insight into emotion and truth. Of course, if we hear a language we understand, we gain the semantics of the spoken word. There is a plethora of literature focused on analysing speech. This chapter covers methods for extracting information from the speech in our corpus, evaluation of effective features and justification of feature choice.

4.1 Audio Features

For most people the first impression of speech is from hearing. Sound waves propagating through the fluid medium of the air around us reach our ears. The human ear, or more

specifically the cochlea, is in essence a frequency analysing device [Gold and Pumphrey, 1948]. Perhaps the most mature field of computational speech processing is Automatic Speech Recognition (ASR), so we must consider audio features employed in that domain. Many approaches to ASR respect the function of the cochlea, but also other aspects of speech production and hearing. We consider a number of audio features, particularly features in the frequency domain, but also some energy features.

4.1.1 Mel Frequency Cepstral Coefficients

The first audio feature to discuss is greatly influenced by the function of the cochlea and is arguably the *de facto* standard for ASR, the MFCC. MFCCs have also found application in speaker recognition [Tiwari, 2010], and in Music Information Retrieval (MIR) [Casey et al., 2008; Logan et al., 2000]. MFCCs are a frequency transformation, respecting the perceptions of human hearing.

MFCC Extraction

Audio speech signals are continually changing over time, so it is often convenient to consider short time frames (Figure 4.1). Within a short time frame, the speech signal is statistically stationary. Convention in the ASR community is to select a time frame between 20 to 40 ms. We maintain this convention for our corpus, as we can align our audio framing with our motion capture frequency while adhering to this standard. We frame our audio with a window duration of 2/59.94 s, and overlap each frame at 1/59.94, giving us a frequency stationary audio frame for every frame of video we recorded (Figure 4.2).

The short time frame is windowed with a Hamming function to reduce the energies at the edge of the frames [Heinzel et al., 2002].

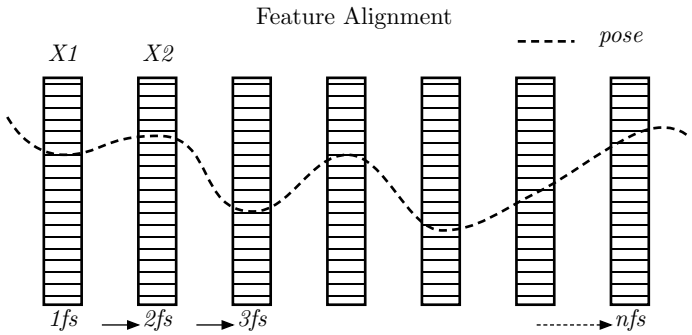


Figure 4.2: Feature alignment. Speech features are aligned to the sampling frequency of the head pose trajectory.

Motivated by observation of the cochlea as a frequency analyser, the periodogram [Schuster, 1898], or power spectrum, of each short time frame is calculated, by taking the squared absolute Discreet Fourier Transform (DFT).

The cochlea can not discriminate closely spaced frequencies, and as frequency increases, this limitation increases. Therefore, we group periodogram bins by convolving increasingly wider triangular filters as frequency increases. This gives an estimate of energy at different frequency regions.

$$m = 2595 \log_{10} \left(1 + \frac{f}{700}\right) \tag{4.1}$$

$$f = 700 \left(\exp^{m/1127} - 1\right) \tag{4.2}$$

The Mel scale relates perceived frequency to its actual measured frequency [Stevens et al., 1937]. By spacing the filters along the Mel scale, the energies more closely follow human hearing perception. Figure 4.3 shows a bank of forty triangular overlapping filters. Frequency is converted to Mel scale in Equation 4.1, and the inverse is Equation 4.2. A sound perceived as twice as loud requires approximately 8 times the energy. The logarithm of the filter bank is a compression operation that again is in response to human perception of sound. We show an illustration of the Log Filter Bank (LogfBank) in Figure 4.4. We will discuss stopping at this stage later, but for now MFCC extraction has one last step.

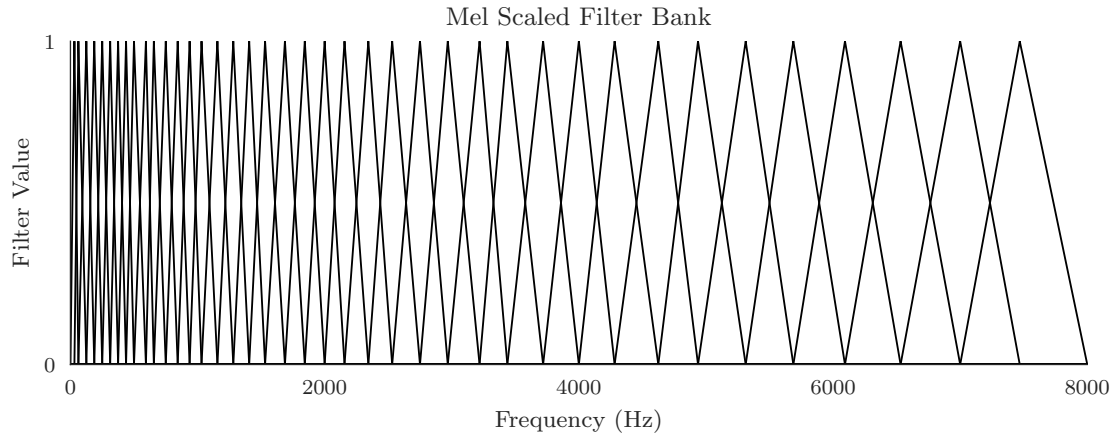


Figure 4.3: Mel Scaled filter banks. We illustrate Mel scaled fBanks by showing our bank of 40 triangular filters, the parameters that we use for all our modelling.

Finally, the Discreet Cosine Transform (DCT) (Equation 4.3) is taken of the log filter bank. As the filters are all overlapping, there is a high degree of correlation in the energies. The DCT decorrelates those energies.

For ASR, the final step is to truncate the MFCCs, typically discarding coefficients above 12 or 13. We are not trying to perform speech recognition, rather we are using the MFCC as an established, perceptually motivated high compression of the speech signal. We opt not to discard the high frequency coefficients and instead allow our model to use them if necessary.

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)k\right] \quad k = 0, \ldots, N-1 \tag{4.3}$$

In summary, our implementation steps of MFCC extraction are:

- Frame the audio signal into 2/59.94 s duration frames with an overlap of 1/59.94 s.
- For each frame, calculate the absolute magnitude of the DFT.
- Convolution of the mel spaced filter bank and the periodogram.
- Take the logarithm of all fBank energies.
- Take the DCT of the LogfBank.

- For ASR, DCT coefficients are truncated, but for our task we retain all of them.

4.1.2 Vocal Tract Model

As MFCCs are motivated by the anatomy of the human ear and how we hear sounds, the Linear Predictor Coefficient (LPC) model is based on a simplified view of the vocal tract as a tube of varying diameter [Makhoul, 1975; Markel and Gray, 1982].

Linear Predictor Coefficients

Linear Predictive Coding is defined as a method for encoding a signal in which a future value is predicted by a linear function of the past values of the signal. At a time t, the speech sample $s(t)$ is the linear function that is the weighted sum of k previous samples. The coefficients of the function characterise the shape of the vocal tract and offer a very compact representation.

Line Spectral Frequencies

Line spectrum pair (LSP) decomposition is a method developed for robust representation of the coefficients of linear predictive models [Itakura, 1975]. The angles of LSP polynomial roots are termed Line Spectral Frequencies (LSF) and they provide an unambiguous representation of the LP model. [Bäckström and Magi, 2006; Schüssler, 1976]

Early Experiments

Other researchers have experimented with vocal tract features, [Ding et al., 2014] to predict head pose, and report significantly lower performance compared with spectral features. Our early experiments using vocal tract modelling to predict head pose did not produce useful results so this particular feature type was abandoned early on.

4.1.3 Energy

The energy is the root mean square of each short time frame, shown in Equation 4.4. This feature vector is a useful guide to prosody and voice activity. A threshold of the energy performs as a simple Voice Activity Detection (VAD). Our early experiments included concatenation of an energy term to spectral features, later as we standardised on LogfBank features, we removed this step as it offered no advantage.

$$\text{energy} = \sqrt{\frac{\sum_{i=0}^{N-1} x(i)^2}{N}} \tag{4.4}$$

4.1.4 Pitch

Pitch is defined as "that auditory attribute of sound according to which sounds can be ordered on a scale from low to high." Whereas that definition is from a psycho-acoustical terminology [of America Standards Institute, 1973], for practical purposes, pitch can be considered as the Fundamental Frequency (F0) of a harmonic signal. Pitch is a useful indicator of prosody, identifying voiced vowel sounds.

Time Domain

A simple technique for estimating pitch is to count the number of times that the signal crosses the 0 level reference. Although easy to calculate, Zero Crossing Rate (ZCR) lacks accuracy with noisy signals, or those where the partials are stronger than the fundamental.

$$f(k) = \sum_{i=0}^{N-k-1} x_i x_{i+k} \tag{4.5}$$

The Autocorrelation Function (ACF), shown in Equation 4.5 is periodic if the signal is periodic. Fundamental frequency is estimated by choosing the highest non-zero-lag peak by

searching within a range of lags. De Cheveigné and Kawahara [2002] use the ACF as the basis of their method.

Frequency Domain

The fundamental frequency can be estimated by measuring the frequencies of the higher harmonic components and finding the Greatest Common Divisor (GCD) of these harmonics [Schroeder, 1968]. Another technique notes the cepstrum has a strong peak corresponding to the pitch period of voiced-speech segments [Noll, 1967]. More recently combinations of these methods have been exploited to achieve more accurate estimates [Kasi and Zahorian, 2002].

Early Experiments

There has been other work, [Kuratate et al., 1999], using F0 to predict head pose. In our own experiments, we don't gain any useful learning with F0 as a single feature. In a similar way to energy, we used F0 as a concatenation feature, but later discarded it when we standardised on LogfBank.

4.1.5 Filter Bank Features

Although we began head pose prediction experiments using MFCCs, motivated by their prominent position in speech processing literature, we quickly standardised on Filter Bank features, specifically Log Filter Bank (LogfBank). In the paper "Recent Advances in Deep Learning for Speech Research at Microsoft", Deng et al. [2013b] conduct a comprehensive comparison of cosine transformed audio features, specifically MFCCs, with primitive spectral features. They were able to demonstrate significant improvement for word error rate for LogfBank over MFCC features. This is in the case of the increasingly effective deep learning techniques for ASR which do not require decorrelation of the feature vector. Although

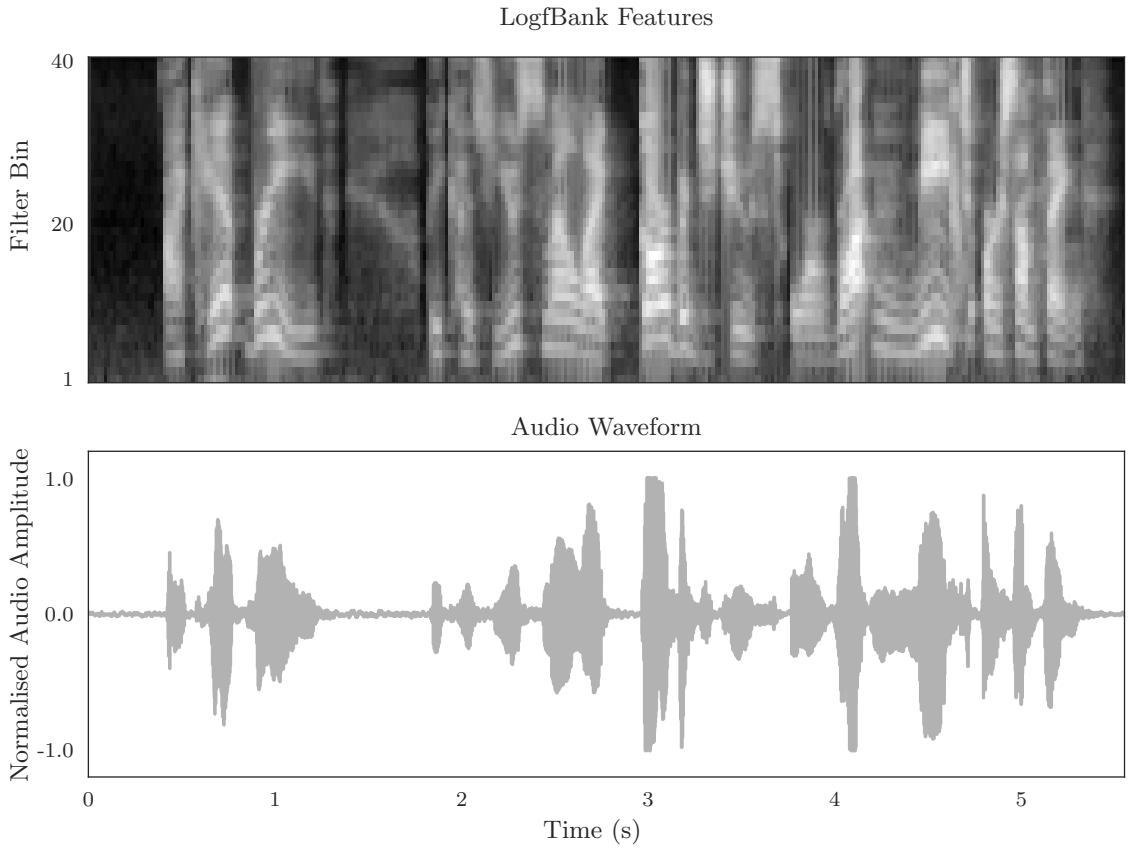


Figure 4.4: Our standardised Audio Feature, the log of the raw filter bank values: Log Filter Bank (LogfBank)

our task is not ASR, we are exclusively employing deep learning techniques to develop our data driven models. For the very practical reasons of available processing resources, it was necessary to standardise our audio feature at an early stage of our research. From our own early experiments showing LogfBank as useful for predicting speaker head pose, to the work in Deng et al. [2013b], we standardise our audio feature as Log Filter Bank (LogfBank). We show our standard audio feature, LogfBank, in Figure 4.4. The process of extracting our features is exactly as described in Section 4.1.1, but importantly, we do not use the final DCT step as decorrelation is not desirable for deep learning.

4.2 Trained Audio Features

A number of authors have examined the relationship that exists between rigid head movement and speech by measuring correlation between extracted audio features and head pose angles. The hypothesis being, if the correlates are significant, the feature is a good choice for modelling head pose.

Kuratate et al. [1999] described the correlation between head motion and F0, a basic component of prosody. The mean correlation to each of pitch, yaw and roll (X, Y, Z rotations in a Y up right handed coordinate system) channels was reported as correlation coefficient, $r = 0.83$. They noted this high value as being "sensitive to the absolute values, rather than the spatio-temporal patterning, of head posture".

Hofer and Shimodaira [2007] examined the correlation between Euler angles of the rigid motion channels and F0, Root Mean Square (RMS) energy and their derivatives. On finding no significant correlation with each of these single dimension features, they used CCA to measure correlation against a combined audio feature vector consisting of the first 12 MFCCs, F0, energy, and their respective first and second derivatives. Significantly, they recorded a best correlation of only 0.08, within utterance and global level speech.

Busso et al. [2007] investigated the rigid head motion correlation with a speech feature vector of F0 combined with RMS energy and respective first and second derivatives. The correlation was measured across a range of emotions: neutral, sad, happy and angry. In this study moderately high values of $r = 0.69$ to $r = 0.74$ were reported.

Given the previous observations regarding the relationship of speech to head pose, and the recent surge in the use of DNNs for many aspects of speech and language modelling, it is prudent to consider the appropriate DNN architecture. LSTM networks have found application in many areas of language modelling including ASR, translation and speech prediction [Graves and Jaitly, 2014; Sutskever et al., 2014; Cho et al., 2014; Graves, 2013]. Other authors have found CNNs effective for these and other diverse applications, sometimes

in combination with LSTMs or some related variant [Sainath et al., 2014, 2013; Trigeorgis et al., 2016; Han et al., 2014].

We propose an experiment to use a CNN front end, feeding to a deep BLSTM back end to produce purpose trained audio features for head pose modelling.

4.2.1 Correlation

We replicated experiments from the literature measuring correlation of head pose with audio features derived from our own corpus. Table 4.1 shows Pearson's r correlation values for the 1D features RMS energy and F0 with each of the nod (x), yaw (y) and roll (z) axes of rotation. We show the mean value for each channel as the mean correlation of all individual utterances as described in Kuratate et al. [1999]. Clearly, these values do not reflect the findings of Kuratate et al. [1999], so we also show the *best* correlation for an individual utterance. It is the case, certainly within our data, that a speaker is equally likely to move their head from left to right during speech, as right to left. This action creates both positive and negative correlation, and so we assume this is the meaning of Kuratate's comment on absolute values. With this assumption, we find similar values to Kuratate et al. [1999].

Table 4.1: Speaker A and B correlation coefficients for single dimension features and each head pose channel.

Feature	Nod	Yaw	Roll
Energy mean A	-0.083	0.071	-0.015
Energy best A	0.612	0.722	0.549
F0 mean A	-0.160	0.052	-0.032
F0 best A	0.625	0.751	0.822
Energy mean B	-0.097	-0.003	-0.034
Energy best B	0.490	0.563	0.590
F0 mean B	-0.161	0.099	-0.011
F0 best B	0.838	0.771	0.903

Table 4.2 shows the correlation values for the features employed by Hofer and Shimodaira [2007] and Busso et al. [2007]. We used CCA to find the best correlation for the features they

described and the combined x, y, z rotations of the head. We find that our data produces values somewhat similar, but slightly lower, to Busso *et al.* when we take the mean value for each individual utterance. When compared to Hofer *et al.*, we find much higher values. To offer an explanation, we also report on the correlation to the concatenation of all our utterances; as if one long continuous utterance. Here, the somewhat ambiguous mapping of speech to head pose reduces correlation values significantly, though not as low as Hofer *et al.* recorded. We suggest the context of data collection is responsible for this remaining difference, with our rather more expressive speech being generally more correlated with head pose.

Table 4.2: Using CCA to show correlation with higher dimension features for speaker A and B, as described by Hofer and Shimodaira [2007] and Busso et al. [2007]. The "mean" column shows the mean of all the individual utterance correlations, the "conc" column shows the single correlation for the concatenation of all utterances.

Feature	mean	conc
Hofer A	0.884	0.230
Busso A	0.584	0.091
Hofer B	0.889	0.213
Busso B	0.569	0.144

The spectral features used in these comparisons rely on the standard Mel-scaled filter bank, we examined the correlation of raw fBank values and also MFCC features with our head pose data. In this case we used 40 filter banks and did not truncate the MFCCs. We scaled our filter banks with the commonly used formula shown in Equation 4.1. Table 4.3 shows that the correlates are broadly similar, with MFCCs having a small advantage for the arguably more general case of the concatenated utterances.

4.2.2 Model Topology

We built two network variants to learn the required head pose to acoustic features. The first will allow us to extract a filter bank given magnitude spectrum input. This concept

Table 4.3: Spectral features for speaker A and B, with 40 coefficients in each of fBank and MFCC.

Feature	mean	conc
fBank A	0.778	0.178
MFCC A	0.794	0.185
fBank B	0.775	0.098
MFCC B	0.794	0.182

is similar to that proposed by Sainath et al. [2014] for the ASR task, but we use Rectified Linear Unit (ReLU) as our non-linearity, feed into BLSTM and assess performance with head pose correlation. Our second network accepts raw waveform input directly, with two convolutional layers trained as our feature extractor.

Deng et al. [2013b] effectively argue against MFCCs for deep learning applications. Furthermore, as there are varying opinions on what exactly the Mel-scale is [Tobias, 2012], our aim is to build a CNN front end that we hope will achieve an improvement over our raw Mel-scaled filter bank.

4.2.3 Network 1

The first network has a CNN front end of $k = 40$ filters of length 1, and perform a 1D convolution over t time frames with n channels. The $k \times t$ result feeds into a deep BLSTM with output to the real values of the angular head motion. The network topology is illustrated in Figure 4.5.

We used a sliding frame over the time domain audio signal of $2/59.94\,\mathrm{s}$ with an overlap of $1/59.94\,\mathrm{s}$, matching the sampling rate of our motion data. Following convention, each frame was multiplied by a Hamming window. Our network input in this case is the magnitude of the DFT with $n = 268$ frequency domain coefficients as the number of channels and $t = 59$ time frames. The coefficients were individually normalised to have zero mean and standard deviation of one over the data set.

Depending on the size of the input, the total number of parameters for the network is in the order of around a million, with the vast majority in the deep BLSTM. The number of parameters for the CNN filter bank are approximately ten thousand; a small proportion of the training workload. This network is regularised using a dropout value of 0.25 between the BLSTM layers. We do not use dropout on input.

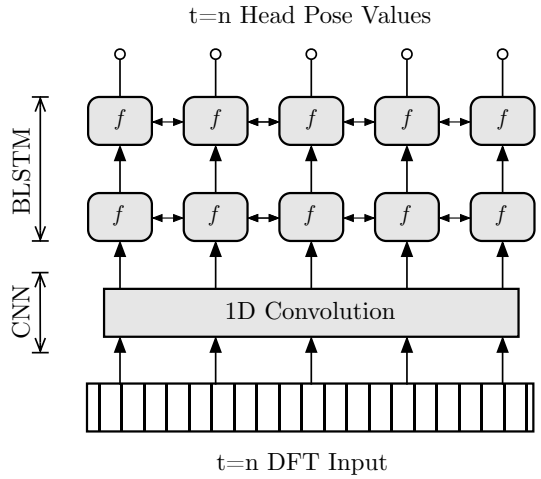


Figure 4.5: Network 1 accepts absolute DFT (magnitude spectrum) input. A bank of k CNN filters feed into deep BLSTM layers.

4.2.4 Network 2

For comparison, we trained an end to end solution that presents the raw waveform as input to our network. Figure 4.6 shows the topology of the network. We use two CNN layers, the first uses $n = 268$ filters of length equal to our audio frame used in 4.2.3 (Equation 4.6). We perform 1D convolution over time with 1 channel.

$$\lceil 2/59.94 * 16000 \rceil = 534 \tag{4.6}$$

We do not use max pooling, instead setting stride equal to the step size in our first network (Equation 4.7).

$$\lceil 1/59.94 * 16000 \rceil = 267 \tag{4.7}$$

The second convolutional filter bank has $k = 40$ filters of length 36, and we perform a 1D convolution over t time frames with $n = 268$ channels, from layer 1. In this network, the CNN front end bears a more significant training load, with an order of magnitude increase in the number of parameters. To equalise the number of parameters in the two networks we reduce the number of hidden units downstream. This network also uses dropout for regularisation between the BLSTM layers, with a value of 0.25. Again, no dropout on input is used.

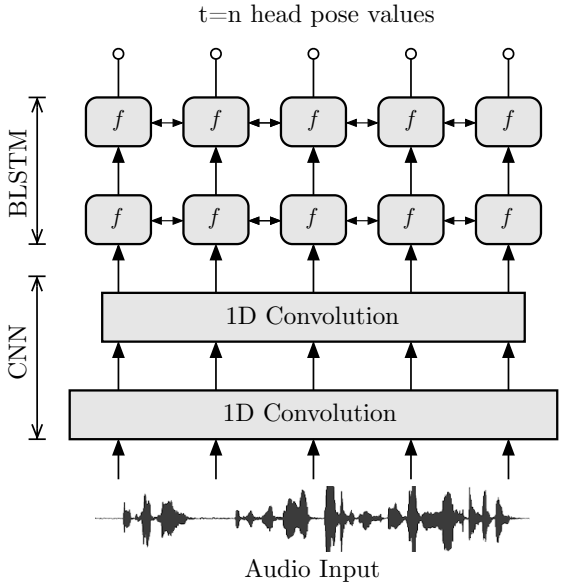


Figure 4.6: Network 2 accepts raw waveform input. Two banks of CNN filters feed into deep BLSTM layers.

4.2.5 Trained Features Results

We trained the networks on our entire data set, split 90% for training and 10% for validation. Our objective function is Mean Squared Error (MSE) against the real values of head pose.

Training continues until no further improvement on the validation set, with a patience of 5 epochs. Model weights are saved at each epoch, and we then select the model with the lowest overall validation error.

For each network, we intercept the output of the model after the CNN front end. This gives us a 40 dimensional set of feature vectors from either the magnitude spectrum input of network 1 or the raw waveform input of network 2. We perform the same CCA correlation comparison over our data described in Section 4.2.1, the results are shown in Table 4.4.

We report correlation for each speaker individually, where the *mean* value is for correlation calculated for each individual utterance, and the *conc* value is the concatenation of all utterances in a randomly selected set of 500. In addition, we make a further random selection of 500 utterances combining both speakers. *fBank* shows the CCA correlation values of Mel-scaled filter bank audio features with our head motion data, *cnn1* are the correlation values for the same utterances from network 1 and *cnn2* are the correlates from network 2.

We note that the features extracted from network 1 show improved correlation, both for individual speakers and the two speakers combined, with the best improvement for combined speakers. Network 2 shows further improvement, although small for individual speakers. The combined speakers result is somewhat better than fBank and a more significant gain over network 1 for this measure. For this network the raw waveform certainly contains more information than the magnitude spectrum of network 1, but we do not entirely attribute the improvement to that observation, as likely some of this extra information will be redundant or irrelevant. We suggest that the filter width in the second CNN layer, which has a large temporal span, is the most significant factor for the head pose task. There are also many more CNN parameters in network 2 so we expect larger quantities of training data would show further improvement.

Table 4.4: Results for each network showing correlation to head pose, with fBank features for comparison.

Feature	mean	conc
fBank speaker A	0.778	0.178
fBank speaker B	0.775	0.098
fBank speaker A+B	0.785	0.283
CNN 1 speaker A	0.842	0.181
CNN 1 speaker B	0.847	0.171
CNN 1 speaker A+B	0.857	0.449
CNN 2 speaker A	0.876	0.228
CNN 2 speaker B	0.895	0.173
CNN 2 speaker A+B	0.896	0.568

4.2.6 Trained Features Discussion

Our experiments have shown that features extracted from trained CNNs out perform hand crafted features when we measure correlation to head pose. Especially encouraging is the improvement for our combined speakers over raw fBank audio features, which is possibly more representative of the general case. Unfortunately, when we try to use learned audio features as the input to a regression model to predict head pose, we get poor results. It is probable that substantially larger data quantities may change that outcome, but we leave audio feature extraction to the hand engineered features we described earlier in this chapter.

Although this experiment did not prove fruitful at the time it was conducted, increasingly available data makes the prospect of repeating the work viable. In the computer vision community, there are a number of pre-trained network front ends available for researchers to tune to more specific requirements. At the time we conducted these experiments, no such speech related networks were available, but may well be soon. Further, our own experience on encouraging learning via an intermediate space (Chapter 8) motivates a rethink of this idea.

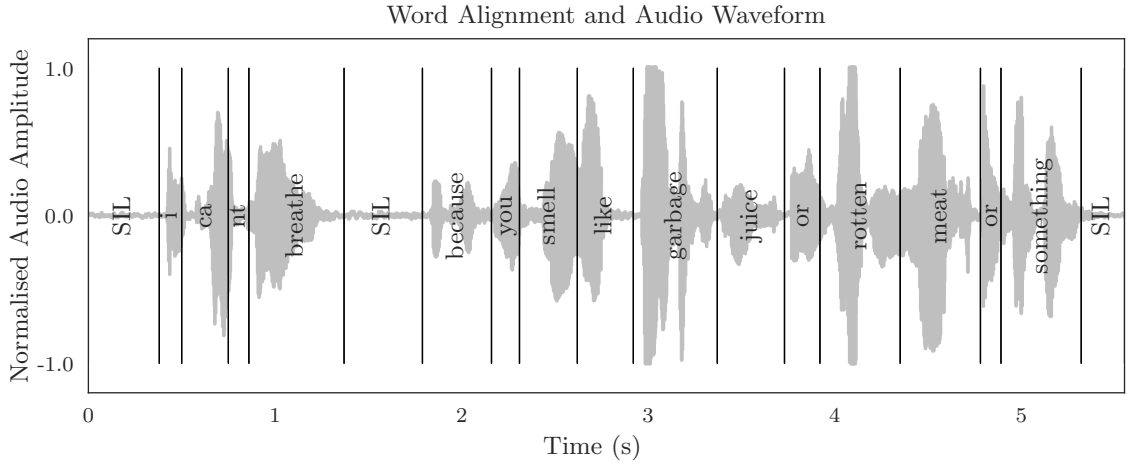


Figure 4.7: Words are force aligned to the waveform.

4.3 Forced Alignment

As a by-product of our data collection process, we had a reasonably accurate transcript of all the utterances in our corpus. However, aligning the transcript, first to word, then to phoneme level required a lot of manual editing. We applied an iterative process to correct misspelled, missing or repeated words in the transcript. We also needed to find mispronunciations in the audio that could be permitted such as "cannot" to "can't" and exclude any that were not well pronounced. The Montreal Forced Aligner (MFA) [McAuliffe et al., 2017] and Gentle [Ochshorn and Hawkins, 2017] forced aligners (both based on the Kaldi Neural Network Model, Povey et al. [2011]) were used to identify out of dictionary words and contractions, with subsequent editing of transcripts and word dictionary, until we had satisfactory alignment of every utterance in the corpus.

4.4 Phonemes

We used the Carnegie Mellon University Pronouncing Dictionary (CMUdict), an open-source pronunciation dictionary for North American English, that is actively maintained

and periodically updated. CMUdict has mappings from over 130,000 words to their pronunciations in the ARPAbet phoneme set, a common standard for English pronunciation. Strictly, the dictionary maps *Phones*; distinct sounds without association with the meaning of the spoken word, as opposed to Phonemes which if incorrectly labelled change the meaning of the word. This allows us to consider modelling speech outside of our training language, or singing possibly.

In addition to the 39 phones in the dictionary, there are lexical stress markers on the vowels:

0 No stress.

1 Primary stress.

2 Secondary stress.

Ultimately, we arrive at a total of 71 different phone labels, including the stress markers for vowels, which we embed using One Hot Encoding (OHE) to remove categorical ranking. Using the very same tools we used to force align words, we align the phone labels to the audio. Figure 4.8 shows this alignment for the example word "something". We need to consider how we can use this alignment to train our models, as it appears we have introduced an arbitrary interval for this feature.

To make use of the phone as a feature vector for model training, we repeat the phone label for every sample of motion at our sample frequency of 1/59.94 s, changing the phone emission as the alignment timing updates. This concept is illustrated in Figure 4.9, in which we show the alignment of the word "something" at a sample rate equal to the motion sample rate. Note, to avoid printing the labels too small, we omit the lexical markers on the vowels as previously shown in Figure 4.8.

Table 4.5: CMU Phonemes Table. Here we reproduce the examples found with the CMU Pronunciation Dictionary for American English.

Phoneme	Example	Translation
AA	odd	AA DH
AE	at	AE T
AH	hut	HH AH T
AO	ought	AO T
AW	cow	K AW
AY	hide	HH AY D
B	be	B IY
CH	cheese	CH IY Z
D	dee	D IY
DH	thee	DH IY
EH	Ed	EH D
ER	hurt	HH ER T
EY	ate	EY T
F	fee	F IY
G	green	G R IY N
HH	he	HH IY
IH	it	IH T
IY	eat	IY T
JH	gee	JH IY
K	key	K IY
L	lee	L IY
M	me	M IY
N	knee	N IY
NG	ping	P IH NG
OW	oat	OW T
OY	toy	T OY
P	pee	P IY
R	read	R IY D
S	sea	S IY
SH	she	SH IY
T	tea	T IY
TH	theta	TH EY T AH
UH	hood	HH UH D
UW	two	T UW
V	vee	V IY
W	we	W IY
Y	yield	Y IY L D
Z	zee	Z IY
ZH	seizure	S IY ZH ER

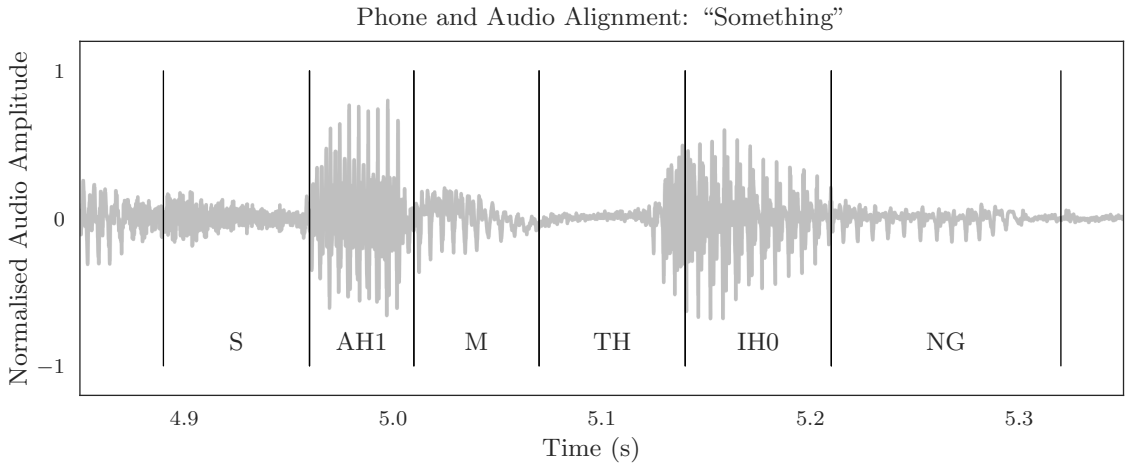


Figure 4.8: Phones are force aligned to the waveform. Here Subject B says “something”. We show the temporal alignment of each phone, with the stress marker on the vowels.

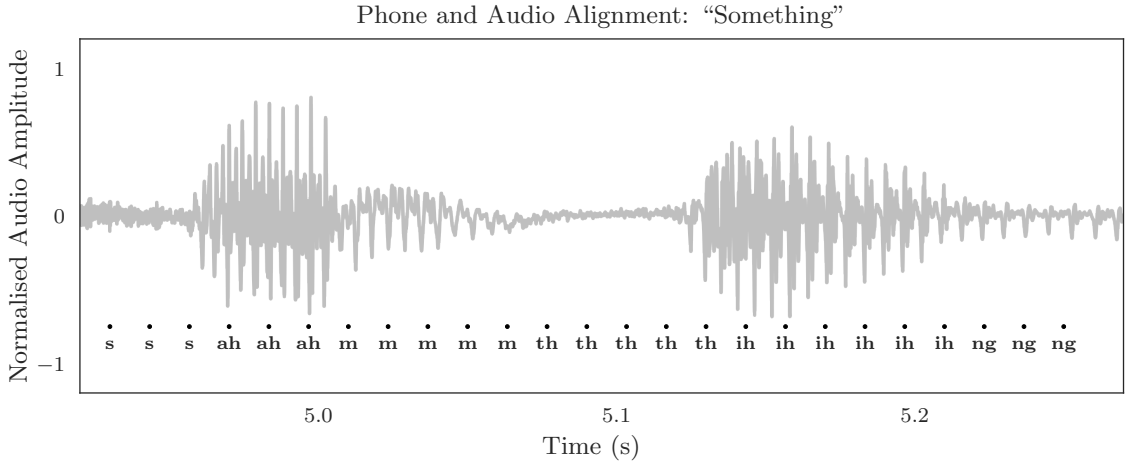


Figure 4.9: A phone is emitted at the sample frequency of the motion. For the word “something”, we show the phone label at each motion sample. We do not include stress markers in this illustration simply for clarity.

4.5 Text Parsing

We believe sentence level and word level decomposition of speech offers a view on action planning [Damasio et al., 1996; Kendon, 1994] that may not be available on a frame wise basis. Conditional generators that we use on a frame wise basis in Chapter 6 can be conditioned at a lower temporal level. The number of possible outcomes increases considerably, and our data does not contain sufficient examples to gain traction with this idea, but we show here examples of feature spaces we have explored. We used both SyntaxNet [Andor et al., 2016] and spaCy [Honnibal and Johnson, 2015]. Both of these syntactic parsers claim human level performance, and comparison on our own corpus vocabulary confirms near identical results for both models. One use case is the embodied agent. With current state of the art text to speech generators, and a model capable of generating convincing animation at the sentence level, we would have a very compelling avatar.

Table 4.6: The parse table for our example utterance: "I am the happiest baker on the planet right now!"

	PoS	Tag	Dep
I	PRON	PRP	nsubj
am	VERB	VBP	ROOT
the	DET	DT	det
happiest	ADJ	JJS	amod
baker	NOUN	NN	attr
on	ADP	IN	prep
the	DET	DT	det
planet	NOUN	NN	pobj
right	ADV	RB	advmod
now	ADV	RB	advmod
!	PUNCT	.	punct

Table 4.6 shows an example sentence from our corpus: "I am the happiest baker on the planet right now!" The Parts of Speech (PoS) are word types assigned to tokens using the Universal Dependencies scheme [Silveira et al., 2014], like verb or noun. Dependencies (Dep) are syntactic labels, describing the relations between individual tokens, like subject

or object. Tags are a similar, but finer grained label based on the ClearNLP project [Choi, 2016].

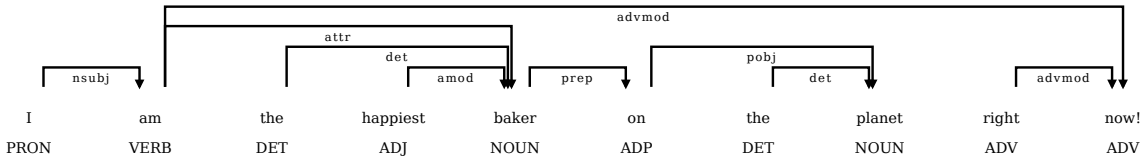


Figure 4.10: The parse tree for our example utterance: "I am the happiest baker on the planet right now!"

Figure 4.10 shows an alternative tree view of the same data. This view gives some insight to the degree of emphasis a speaker might apply to the utterance, and also the potential for a model to learn from this global sentence view.

4.6 Word Embedding

With a granularity sitting between the phoneme level and the sentence level, word embeddings are another feature we developed from our data. Word2Vec [Mikolov et al., 2013] and Global Vectors for Word Representation (GloVe) [Pennington et al., 2014] are methods for embedding a large dictionary in a compact vector space, such that words that are semantically similar are near to each other in that space. The embeddings are created by training a single layer Feed-Forward Neural Network (FFN) to predict the probability of adjacent words. When training has converged, the weights of the layer are used as a base to project the One Hot Encoding (OHE) word to the vector representation. The concept of training the model for one task, to use its weights as a filter for another task provided the inspiration for our work in 4.2.

Figure 4.11: Corpus vocabulary embedded in GloVe. The 2823 unique tokens in our data set plotted using t-Distributed Stochastic Neighbour Embedding (t-SNE) for 2D space conversion. One can see how words with similar meanings are close in the vector space.

4.7 Discussion

We have discussed a number of speech features, both audio and text based, that we have experimented with and found successful for predicting head pose from speech. Along the way we have discarded many feature types that have appeared promising, but ultimately did not provide meaningful learning for our tasks. At the head of this chapter, we stated that extracting useful information from the speech signal was essential to the task of making predictions of head pose. While this is true, the use of hand engineered features, MFCC, fBank etc., leaves us with the strong belief that, with a little better model engineering, and perhaps a lot more data, we can train models directly from raw audio input.

5 Neural Networks

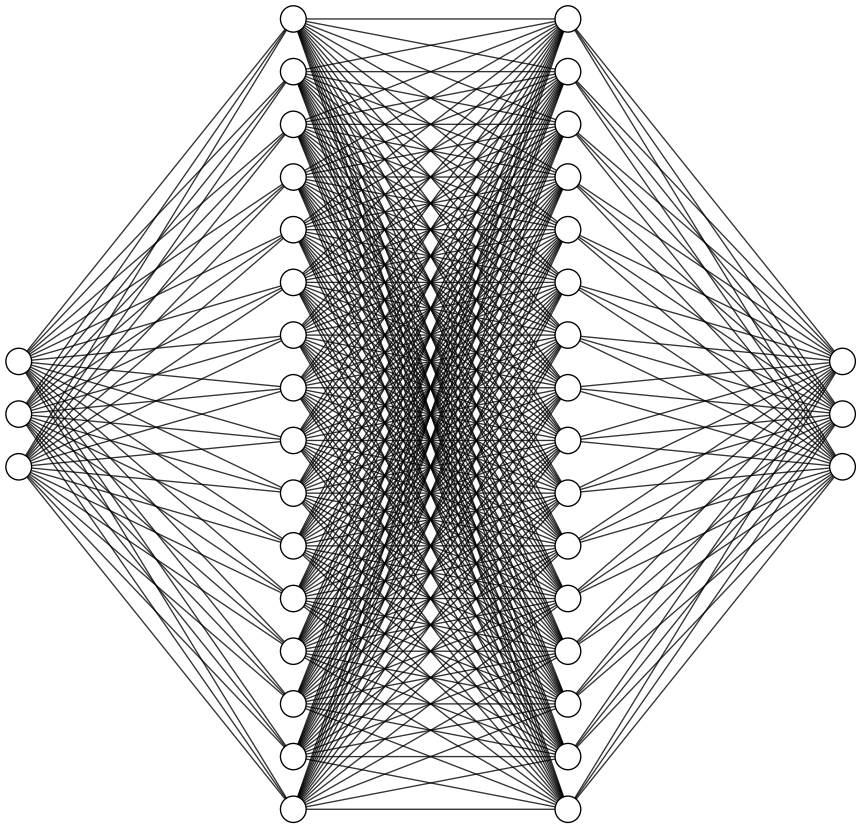


Figure 5.1: Diagram of a Neural Network. We can illustrate the weights of a neural network as the connections on nodes in a graph. Each node of the graph is referred to as a hidden unit, and holds our *activation* function.

Neural networks offer a powerful ability to model highly complex functions. Deep learning, using neural networks with many hidden layers, provides great utility for supervised learning tasks and, for a number of medium to large data driven tasks is the state of the art. Although the thesis does not claim to advance the development of neural machine learning, we make extensive use of this capacity. In this chapter we provide a brief introduction to neural networks with respect to our task, and provide justification for choosing these methods. We also discuss the *topology* of the networks we construct to model the pose of the head of a human speaker. We formalise each of our networks with the appropriate equations, and show diagrams that inform the reader of the flow of data for each variation of network we implement.

5.1 Multi-Layer Perceptron

Multi-Layer Perceptrons (MLPs), also called Feed-Forward Neural Networks (FFNs) or ANNs [Rosenblatt, 1958, 1961], are at the foundation of deep learning. A MLP aims to learn a function f that maps input $\boldsymbol{x}$ to output $\boldsymbol{y}$, by learning parameters θ, so $\boldsymbol{y} = f(\boldsymbol{x}; \theta)$. Specifically, we do not use MLPs to build our models for our work, but they conceptually underpin all the networks we do build and are an essential introduction.

$$\begin{aligned} \boldsymbol{x} &= [x_1 \ x_2] \\ \boldsymbol{w} &= [w_1 \ w_2] \\ \theta &= \boldsymbol{w}, b \\ \sigma(z) &= \frac{1}{1 + e^{-z}} \\ f(\boldsymbol{x}) &= \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b) \end{aligned} \tag{5.1}$$

We show a simple example of an MLP in Equation 5.1. Here $\boldsymbol{x}$ is a feature vector of two features, and θ is the learned weights, $\boldsymbol{w}$ and bias b. The non-linearity σ is a simple sigmoid function. We also show the same small network in Figure 5.2. We do not show the bias

in the Figure, assuming it is always present, and the weights are represented by the lines connecting the inputs to the layer of one hidden unit in the centre, which illustrates the function f we have learnt, connected to the output y. Strictly, the diagram form infers a weight connecting f to y, this is always unity. Illustrating neural networks in this way can

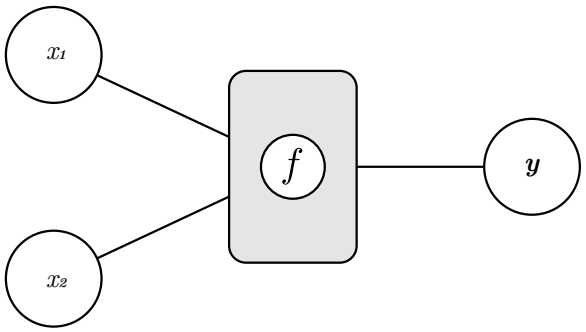


Figure 5.2: The MLP is the basic building block of the ANNs used throughout this work. In this figure we show an example for two dimensional input and single dimension output. The lines connecting the nodes represent weights in the neural network. The network has one layer, where the weights and bias are linearly combined and attenuated with an *activation* function.

become very complex quickly (Figure 5.1). So we generally prefer to compress the view of individual weights and hidden units to a more compact graphical form that we use in Figure 5.3. Although we define the activation function as a sigmoid in this example, we actually use a number of non linear functions for different applications throughout this work. We plot the functions we use in Figure 5.4.

5.2 Recurrent Neural Networks

Our task is modelling a time series; an ordered sequence of values of a variable at equally spaced time intervals [Natrella, 2010, 6.4.1]. Although powerful, the MLP does not naturally

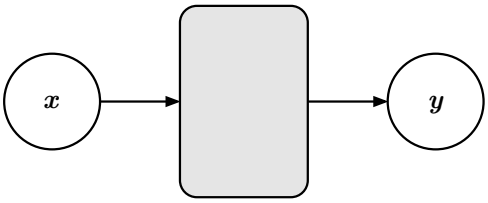


Figure 5.3: Here we show a compact diagram for a MLP with one hidden layer. We no longer draw a line to represent individual weights, which can become very dense in the illustration. Instead, we assume the weights are determined by the number of inputs and outputs at each layer.

lend itself to modelling sequences and series. In the previous section, MLPs do not form cycles, simply mapping input x to output y. If we allow cyclic connections, we arrive at the Recurrent Neural Network (RNN). RNNs are specialised networks for modelling data of the form $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$ or compactly, $\boldsymbol{x}^{(t)}$ with the time step t in the interval $[1, n]$. By permitting a cycle, RNNs, can potentially map the complete history of previous inputs to each output. Many varieties have been proposed in the literature (e.g., Elman [1990], Jordan [1986]). We will show a simple example with one hidden layer in Figure 5.5, and equivalently in Equation 5.2, with h symbolising the hidden state the cycle passes through.

$$\begin{aligned} \boldsymbol{h}_t &= \sigma_h(\boldsymbol{W}_h \boldsymbol{x}_t + \boldsymbol{U}_h \boldsymbol{h}_{t-1} + \boldsymbol{b}_h) \\ \boldsymbol{y}_t &= \sigma_y(\boldsymbol{W}_y \boldsymbol{h}_t + \boldsymbol{b}_h) \end{aligned} \tag{5.2}$$

For a finite number of time steps we can unroll the graph. We illustrate this idea in Figure 5.6. This concept is fundamental to our purpose of time series modelling. The important detail of this illustration is that each time step shares the same function f and weights

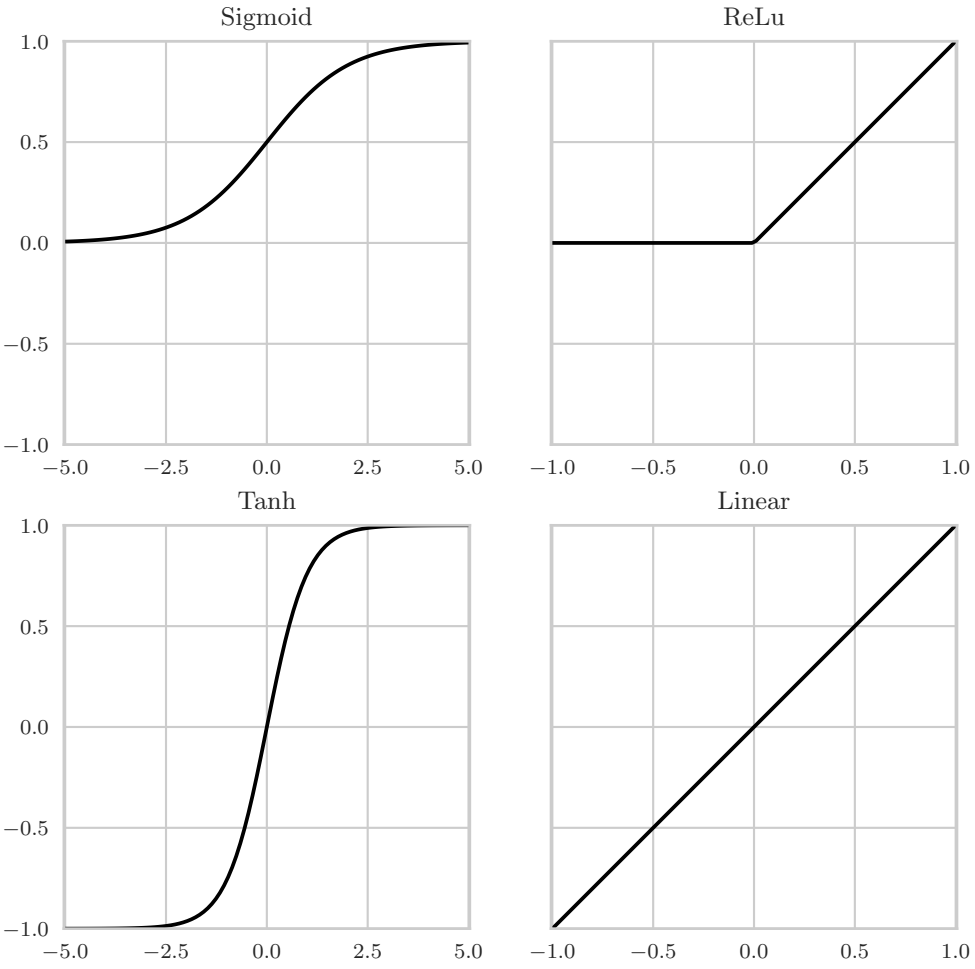


Figure 5.4: A number of activation functions are used in the ANNs in this thesis. Each has useful properties for a given application.

$\boldsymbol{W}$ and $\boldsymbol{U}$. This allows learning of arbitrary sequence lengths and predictions of sequence lengths that did not appear in the training phase. We can take advantage of that property by building RNNs that:

- produce output at each time step, with recurrent connections between each time step, shown in Figure 5.7.
- have recurrent connections between each time step, and have dissimilar input and output sequence lengths, shown in Figure 5.8.

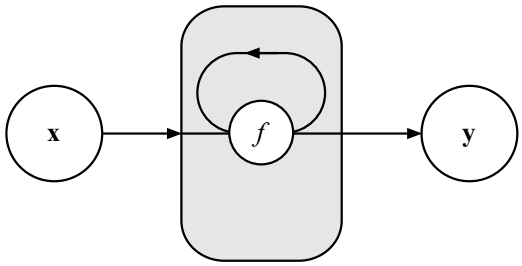


Figure 5.5: An RNN has a cyclic feedback loop allowing events in the past to influence output in the present.

- have recurrent connections between each time step, but produce output at only the last time step, shown in Figure 5.9.
- have recurrent connections between each time step, have input at only the first time step, but produce output at all time steps, shown in Figure 5.10.

5.3 Long Short Term Memory

The RNN described in the previous section has the very appealing property of making use of the data at earlier time steps to influence the prediction at the present time step. In practice, this ability is limited to influence from just the recent past due to *vanishing gradients* [Bengio et al., 1994]. For our task of modelling head pose during speech, we can see event durations in excess of 500 ms, or greater than 30 motion time steps. This limitation of the RNN renders them unsuitable for our task.

The Long Short Term Memory (LSTM), introduced by Hochreiter and Schmidhuber [1997] is an advance of the simple RNN that specifically addresses this limitation. The LSTM has

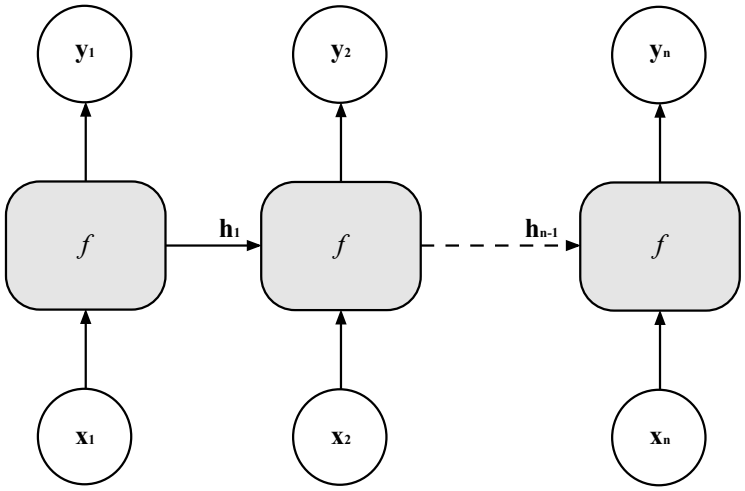


Figure 5.6: The concept of *unrolling* RNNs helps to illustrate the notion of modelling events over time.

a 'Cell' ($\boldsymbol{C}$ in Equations (5.5), (5.6)) that controls how much of the past is relevant to the present. The Cell is regulated by three gates, the input, forget, and output gate. Each of these gates outputs values in the interval $[0, 1]$.

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{5.3}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{5.4}$$

$$\tilde{\boldsymbol{C}}_t = tanh(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \tag{5.5}$$

$$\boldsymbol{C}_t = \boldsymbol{i}_t * \tilde{\boldsymbol{C}}_t + \boldsymbol{f}_t * \boldsymbol{C}_{t-1} \tag{5.6}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \tag{5.7}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t * tanh(\boldsymbol{C}_t) \tag{5.8}$$

All of our models feature LSTM networks. There are many variations to consider; a study by Greff et al. [2017] investigating a number of varieties (with 15 years of total processing time), concluded that no variations have improved on the original design. Melis et al. [2018] demonstrate that the LSTM offers state of the art on many language modelling

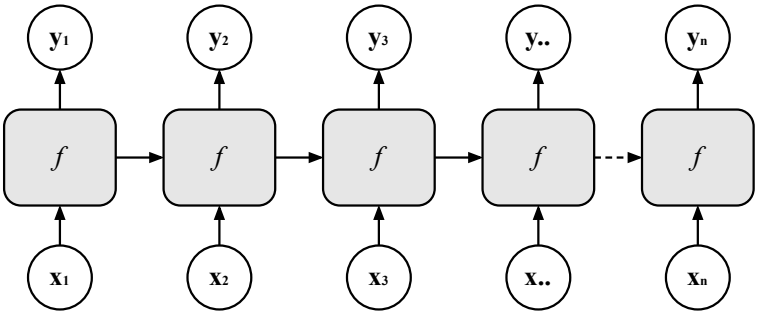


Figure 5.7: The RNN can model data with many inputs and many outputs.

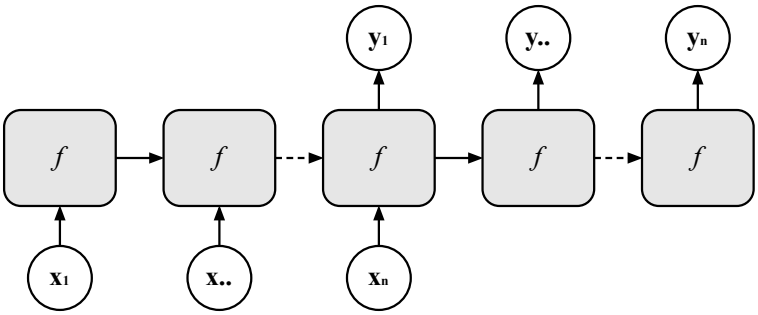


Figure 5.8: The number of inputs does not need to equal the number of outputs, nor do the outputs need to align temporally.

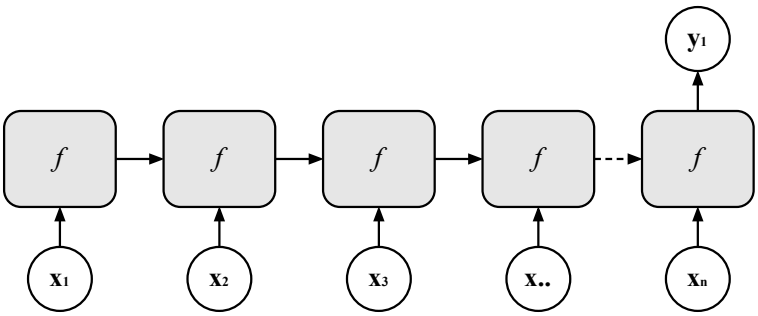


Figure 5.9: Using a many-to-one network, the RNN can create a compressed latent representation of temporal data.

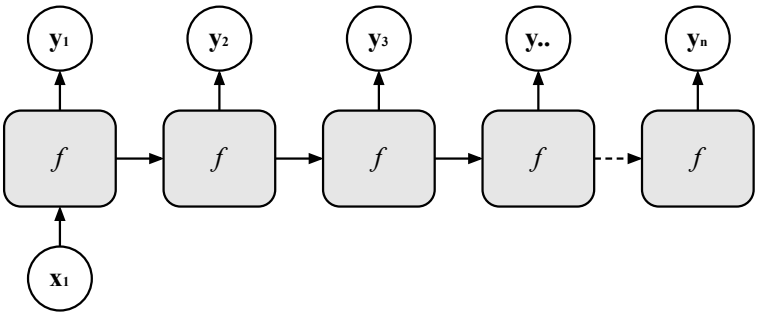


Figure 5.10: A one-to-many RNN can inflate a latent representation of temporal data.

tasks. Therefore we describe the LSTM as we implement, in the Equations (5.3) - (5.8), where σ is the sigmoid function and $\boldsymbol{f}$, $\boldsymbol{i}$, $\boldsymbol{C}$, $\boldsymbol{o}$ are the forget gate, input gate, memory cell, and output gate respectively. We also show a graphical representation in Figure 5.11 to give an alternative understanding of how data flows through the graph. Our implementation

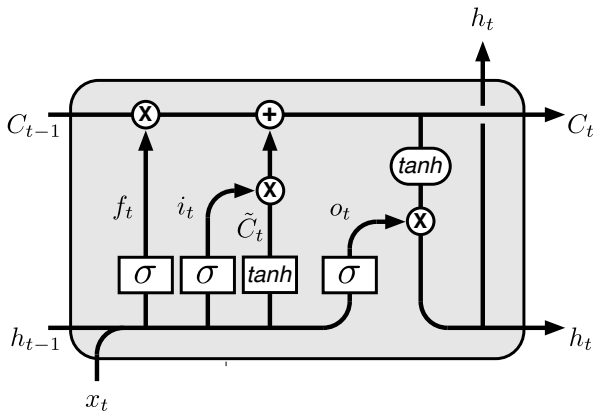


Figure 5.11: LSTM data flow diagram. To give a different view of how the data flows through the LSTM, we show this figure that corresponds to the Equations (5.3) - (5.8).

initially used the Computer Algebra System (CAS), Theano [Al-Rfou et al., 2016], later we embraced the Keras framework [Chollet et al., 2015], which allowed easy transition to Tensorflow [Abadi et al., 2015], as unfortunately Theano has reached the end of its development. We exclusively train our networks on the GPU.

5.4 Bi-Directional Long Short Term Memory

We are modelling predictions of head pose position at a moment in time. Head pose has properties constrained by physical (anatomical) limits, and kinematics. In addition, certain events, for example a pause for breath, have a defined beginning and end that may be a considerable distance apart. The Bi-Directional Long Short Term Memory (BLSTM) introduced by Graves [2012], models events in the future as well as the past over a long time term. This is directly equivalent to concatenating the output of one LSTM with

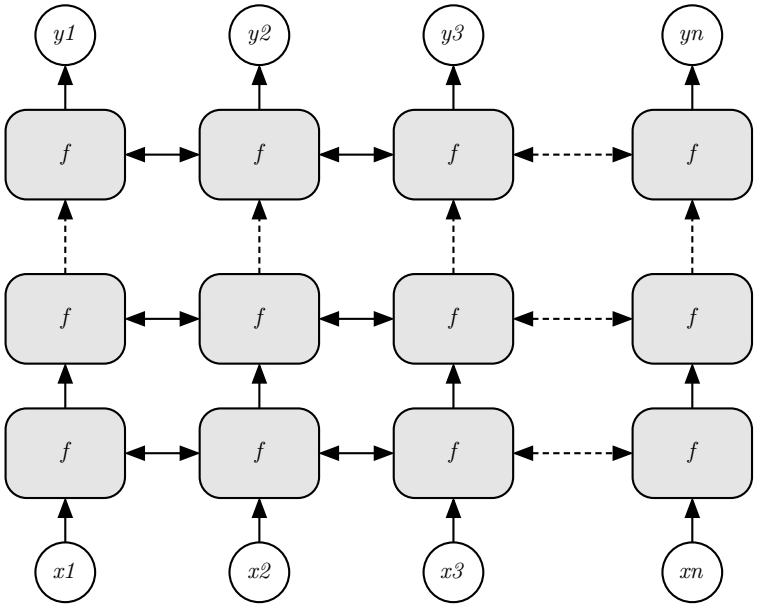


Figure 5.12: The deep BLSTM features in all our modelling solutions. Here we illustrate the concept of passing state forward and backward in time, and through the graph from input to output.

another, identically configured, but with the time series reversed. By recognising the work of Graves, modelling discrete future events such as text prediction, or real values of hand writing trajectories, as an analogy of our own problem, we identify the BLSTM as suitable for our task. We graphically illustrate the topology of the deep BLSTM that is a feature of all our work (Chapters 6, 7, 8) in Figure 5.12. Occasionally we need to compress our graphical illustration even further (e.g., Figure 8.13), and in that case we will simply label a hidden layer with the BLSTM text.

5.5 Generative Models

In Section 6.3.2 we consider the rigid head pose of a speaker repeating the same transcript. Not surprisingly, the outcome of the head pose is different each time, to some degree. This

is not however, an example of the same *utterance* being repeated. Utterances can not be repeated exactly, they are a human action, and like any human action repetition has variation. What if we *want* to synthesise variation for the same utterance? Generative modelling allows us to draw from a normal distribution of a latent space, giving large variation in output for single input.

5.5.1 Autoencoders

Autoencoders are a lossy compression algorithm, where the compression and decompression functions are learnt from data. Typically, the *encoder* and the *decoder* are neural networks that have input $\boldsymbol{x}$ and output $\boldsymbol{x\prime}$. The encoder compresses input to a latent representation $\boldsymbol{z}$, the decoder inflates $\boldsymbol{z}$ to $\boldsymbol{x\prime}$ (Figure 5.13). The encoder and decoder are optimised simultaneously to minimise the loss $\boldsymbol{x} - \boldsymbol{x\prime}$. Autoencoders generally do not outperform standard compression algorithms, but some interesting applications are noise reduction or data visualisation of high dimensional data. For example, we use t-SNE [van der Maaten and Hinton, 2008] to visualise the word embedding of our corpus in Figure 4.11.

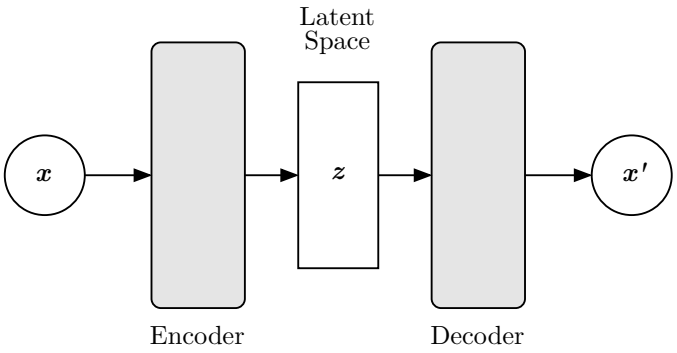


Figure 5.13: Autoencoders learn a lossy compression of data via a latent space $\boldsymbol{z}$.

For our purposes, autoencoders are not particularly useful as the latent space lacks useful structure or may be discontinuous. We can influence the distribution of the latent space with a Variational Autoencoder (VAE).

5.5.2 Variational Autoencoders

VAEs, introduced at the same time by two independent groups, Kingma and Welling [2014] and Rezende et al. [2014], mix the concept of autoencoders and Bayesian inference. The VAE, instead of learning a fixed arbitrary latent vector, learns the parameters, mean μ and variance σ, of a normal distribution (Equation 5.9).

$$\boldsymbol{z} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{5.9}$$

The decoder then inflates a sample from that latent space $\boldsymbol{z}$, and in the same way as the autoencoder, the VAE optimises the encoder and decoder simultaneously to minimise the loss $\boldsymbol{x} - \boldsymbol{x\prime}$. More formally, the encoder, $Q_\theta(\boldsymbol{z}|\boldsymbol{x})$, represents input data $\boldsymbol{x}$ in a latent space $\boldsymbol{z}$

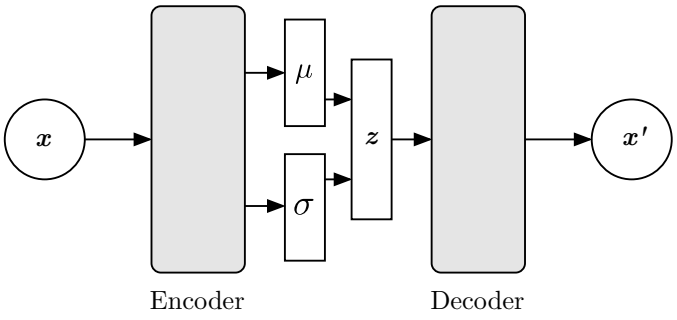


Figure 5.14: Variational Autoencoders learn the parameters, mean μ and variance σ, of a normal distribution.

with weights and biases θ, where the encoder outputs the parameters of a Gaussian proba-

bility density. The decoder, $P_\phi(\boldsymbol{x}|\boldsymbol{z})$, with weights and biases ϕ, transforms the parameters to the distribution of the original data.

The parameters of a VAE are optimised by *two* loss functions. A reconstruction loss, and a regularisation term that minimises the Kullback–Leibler (K-L) divergence. Minimising K-L divergence optimises μ and σ to closely resemble the target distribution. In the paper, Kingma and Welling [Kingma and Welling, 2014, Equation 10] use negative log loss as the reconstruction loss. We could consider scaling our data to the $[0, 1]$ interval and following their implementation, but instead choose MSE as our reconstruction loss. We implement the K-L regularisation term in 5.10, ignoring subscripts. At training time we minimise the sum of the reconstruction loss and the K-L loss.

$$\frac{1}{2}\sum(1 + log(\sigma^2) + \mu^2 - \sigma^2) \tag{5.10}$$

Generating examples is done by sampling from the normal distribution, and feeding forward using just the decoder. This would give us a variation on an example of head pose trajectory. For our purposes, there is one more step to take. We want to generate an example of head motion that is appropriate for the utterance the speaker is making. We need to *condition* our model on speech features.

5.5.3 Conditional Variational Autoencoder

The Conditional Variational Autoencoder (CVAE) [Sohn et al., 2015] adds a conditioning element to the VAE, such that the encoder becomes $Q_\theta(z|x, c)$, and the decoder is $P_\phi(x, c|z)$. Figure 5.15 illustrates how we do this by concatenating our speech features $\boldsymbol{c}$ to both the input vector of motion features, and the latent vector $\boldsymbol{z}$. Specifically, our CVAE implementation uses BLSTM to encode and decode. For training we concatenate the speech features and the motion samples at every time step as input to the encoder. We also concatenate every time step of the speech with the latent variable $\boldsymbol{z}$. For inference, we only use the de-

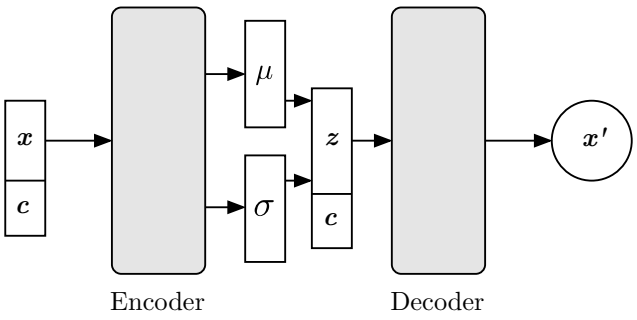


Figure 5.15: Conditional Variational Autoencoders extend the VAE by introducing a conditioning element, $\boldsymbol{c}$, concatenated with $\boldsymbol{x}$ and $\boldsymbol{z}$ respectively.

coder to *generate* examples of head pose. At this time, we concatenate the speech features of our test example with a random sample of noise, which we repeat $t = n$ time steps, of our speech features. Figure 6.14 more explicitly shows the topology of our CVAE model that we train to predict speaker rigid head pose.

5.6 Dropout

Regularising the training of neural networks is an important part of avoiding overfitting (when the model fails to generalise to the validation examples). There are a number of strategies to consider, and the general principal is to apply some sort of penalty to the weights of each layer. These penalties are part of the loss that the network optimises during training. For the practical reasons of finite computational resources, we standardise the regularisation by using *dropout* [Hinton et al., 2012; Srivastava et al., 2014].

The inspiration for the technique is explained in an anecdote by Hinton himself.

> "I went to my bank. The tellers kept changing, and I asked one of them why. He said he didn't know but they got moved around a lot. I figured it must be because it would require cooperation between employees to successfully defraud the bank. This made me realize that randomly removing a different subset of neurons on each example would prevent conspiracies and thus reduce overfitting."

Not only does dropout provide regularisation, it can be considered as a form of ensemble method [Hinton et al., 2012], or even as data augmentation [Konda et al., 2015]. Dropout is recognised as particularly effective for speech and language modelling [Józefowicz et al., 2016]. For all of these reasons dropout is our chosen method of regularisation. All of our models use dropout to regularise training, 0.5 is the default value.

5.7 Data Augmentation

Supervised learning is to present an example input and optimise parameters to minimise a loss to the example output. The main *caveat* of deep learning is the amount of data required. Our corpus has 3600 utterances, a relatively small quantity for deep learning. For comparison, the 'Hello World!' for learning methods, the MNIST handwritten digit recognition task [Lecun et al., 1998], has 60,000 examples. We ensure we have sufficient examples for learning by choosing $t = n$ time steps that temporally span the beginning and end of events we wish to model, where n is in the interval $[29, 129]$. We choose this interval as a compromise of data consumption and event capture, as of course a pause for any reason has no logical bound. We then segment our utterance into sequences of length n. We generally get more pleasing results that capture longer time events, e.g pauses for breath, with $n = 129$. Larger n reduces the total number of examples for training, as we do not pad short ends, or short utterances less than that duration, and thus compromises the goal of maximising the data. An entire utterance is consumed by advancing the temporal span by one sample at a time. We now have unique examples that number in the region of 3×10^5, sufficient for many supervised learning methods.

This penultimate section on data augmentation hides some of the work we carried out to arrive at our technique. Bridging early experiments exploring feature engineering, we discovered we could discard the concatenation of delta features. For example, in ASR the first and second derivatives of MFCCs are commonly stacked with the original as a combined feature. We discovered this was not necessary if we took a small frame span at every time step (instead of a single sample), thus allowing the model to learn higher order features on a frame by frame basis [Greenwood et al., 2017a, Section 3.1]. Ultimately, as we developed better hyper-parameter selection, we could remove this technique, which required post processing to reduce the span back to one sample, and, more training weights for the larger input.

5.8 Alternate Models

There are a number of other strategies that should be mentioned here. We took a decision not to build on some earlier work, based on HMMs, and were unable to implement some other ideas, although we will certainly continue to pursue them in the future.

5.8.1 Hidden Markov Models (HMMs)

We have already mentioned in Section 2.1 that HMMs are an established standard for ASR, and that they are arguably a logical choice for many speech related tasks. There are a number of reasons why a decision was taken not to follow some of the earlier work using this model. HMMs model *state*, and we believe human motion does not fit this paradigm. Ignoring that consideration, labelling discrete output states is a considerable effort in itself. Busso et al. [2005] describe a vector quantization method to provide labels for position states, but only considers rotation of head pose. The larger state space generated by considering translation was noted as prohibitive. Finally, if pose is modelled as a sequence of states, there is a post processing problem to solve.

5.8.2 Generative Adversarial Networks (GANs)

The Generative Adversarial Network (GAN) [Goodfellow et al., 2014] is an exciting modelling approach that trains a generative model by simultaneously training a discriminator. As the discriminator improves in its ability to distinguish a real from fake example, so the generator must improve to fool the discriminator. GANs are notoriously difficult to train, and at this time their greatest success has been in the computer vision community. Unfortunately, we were unable to change this situation, and attempts to train models of this type proved fruitless for our task.

5.8.3 Discriminative Evaluator

Although we were unable to train a GAN, the concept raises an interesting prospect for evaluating predictions. Our hypothesis is a well trained GAN has a discriminator that measures the distance between a real or fake example - minimising this distance is the objective for the generator. This distance could be a useful metric for many tasks that have perceptual aspects that are difficult to evaluate empirically.

5.9 Discussion

In this chapter we have introduced the BLSTM to the task of predicting head pose from speech, after identifying its potential from arguably related tasks in the literature. Not only does the LSTM network have state of the art performance for many speech and language tasks, it has the essential property of modelling the long term dependencies that we see in head pose. Finally, describing complete networks mathematically or graphically can become complex, so a compact diagram form explains the topology of the networks in the forthcoming chapters. We have also given detail on our implementation of a CVAE that uses the BLSTM as both encoder and decoder.

6 Speaker Head Pose

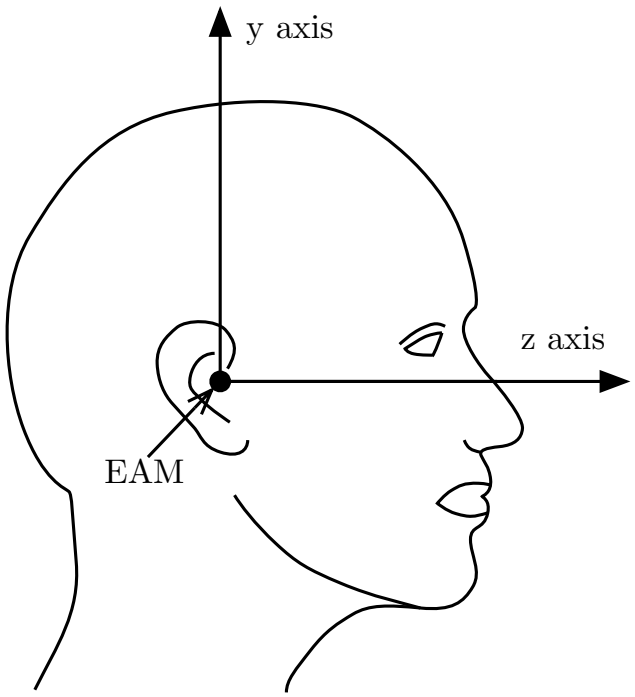


Figure 6.1: The axes of head rotation and pivot location. The $\boldsymbol{x}$ axis passes through each external auditory meatus (EAM), pointing into this figure, The $\boldsymbol{z}$ axis is on Reid's baseline. The $\boldsymbol{y}$ axis is perpendicular to those axes, from a point equidistant from each EAM. We term the rotation about the $\boldsymbol{x}$ axis *nod*, about the $\boldsymbol{y}$ axis *yaw* and about the $\boldsymbol{z}$ axis *roll*.

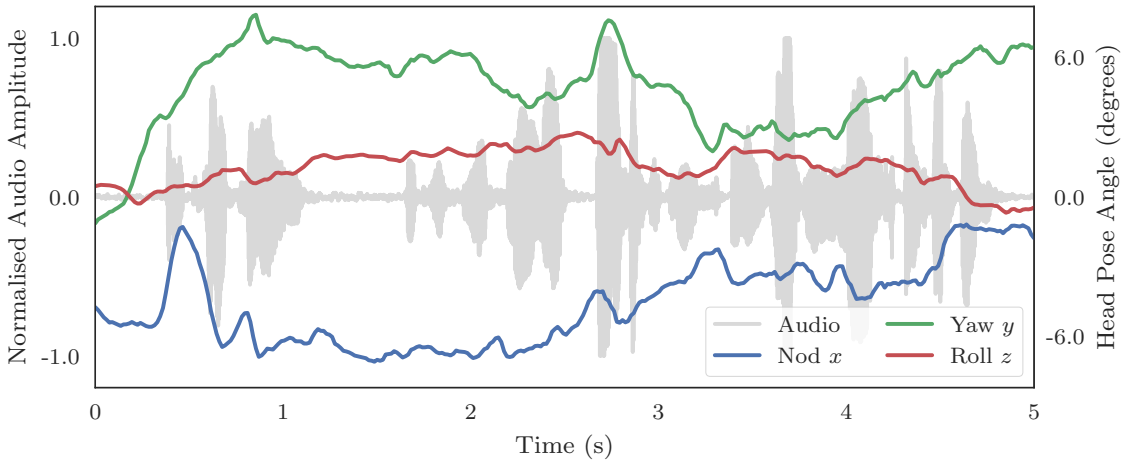


Figure 6.2: The head pose trajectory of Subject A making the utterance: "I can't breathe because you smell like garbage juice or rotten meat or something!" Major events in the speech signal align with major events in the head pose trajectories.

Speaker head motion is a rather intriguing aspect of visual speech. Head motion has been shown to contribute to speech comprehension, [Munhall et al., 2004] yet, unlike the articulators, it is under independent control. As the speech mode contains the most complete information stream in an utterance, it is a reasonable strategy to seek a mapping from within this stream that might enable plausible predictions of head pose. Indeed, there is significant measurable correlation between speech and head motion that has motivated much of the prior art [Hofer and Shimodaira, 2007; Busso et al., 2007].

In this chapter we consider the rigid motion of a speaker's head during speech. Concretely, we seek to answer the question: can we predict the rigid head pose from the speech signal?

We show an example of an utterance from our corpus in Figure 6.2, in which the head pose trajectories are plotted over time on top of the time domain audio, to give a solid impression of how head pose changes during speech. The axes are labelled x, y, z of a right handed system and represent the nod, yaw; the left-right rotation around the vertical axis or head shaking and the left-right roll around the forward facing axis. The speaker, Subject A of our two actors, is saying

> "I can't breathe because you smell like garbage juice or rotten meat or something!"

The style of the speech is expressive, prosodic and emphatic. It is interesting to note, major events in head pose are aligned with major events in the audio.

6.1 Related Work

There have been a number of researchers interested in predicting head motion from speech in recent years (Chapter 2). We can exclude rule based systems that rely on micro-analysis of human motion patterns and subsequent annotation to drive models. We are concerned with data-driven approaches and we should compare our work there.

Many early studies took the approach of clustering head motion patterns and giving class labels [Deng et al., 2004; Busso et al., 2005, 2007]. HMMs were trained for each cluster, learning the relation between the speech features and head motion. These early studies did not produce continuous real output and required post-processing steps to complete the mapping to head pose.

Recently, the GPU has enabled efficient training of DNNs, and within many aspects of speech and language processing, DNNs are now state of the art [Huang et al., 2015; Deng et al., 2013b,a]. Ding et al. [2015] introduced BLSTM networks to the head motion task, noting improvements over their own earlier work with MLPs. More recently Haag and Shimodaira [2016] uses BLSTMs and Bottleneck features [Gehring et al., 2013] and noted a subtle improvement for their own corpus. We should draw our comparisons with these most recent works.

6.2 Corpus

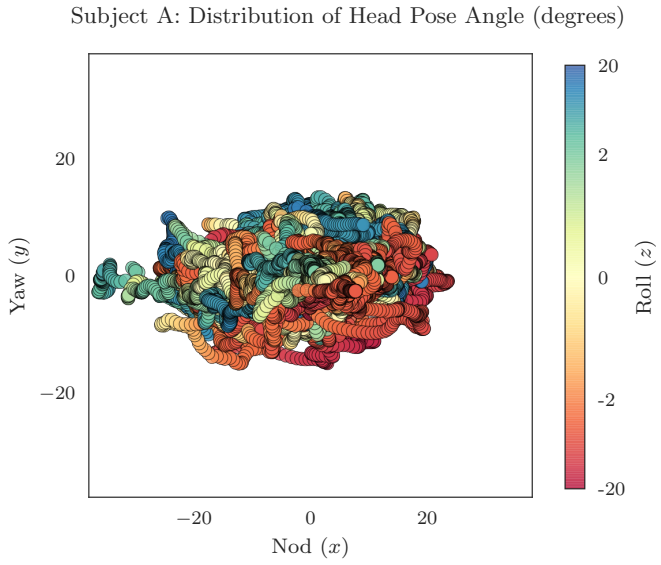
We describe in detail the development of our corpus in Chapter 3. Here we can make some observations regarding the data specifically for the head pose task. We define the axes and pivot point of the head rotation in Figure 6.1. Of course the rotation of the head and neck is a complex biomechanical system with many degrees of freedom [Kunin et al., 2007], which we do not attempt to model. Our representation is akin to motion capture and animation skeletons found commonly in industry. We simplify all the combined rotations below the uppermost animation joint as a translation, giving 6 Degrees of Freedom (DoF) at each motion sample point.

Emphasis should be placed on the *expressive* nature of the corpus. Prior work gathering multi-modal corpora can use speaking styles in data collection that retain motivations of ASR such as phonetic balance or emotion taxonomy, or, collected from a rigidly produced source [Taylor et al., 2012; Busso et al., 2005; Ding et al., 2014].

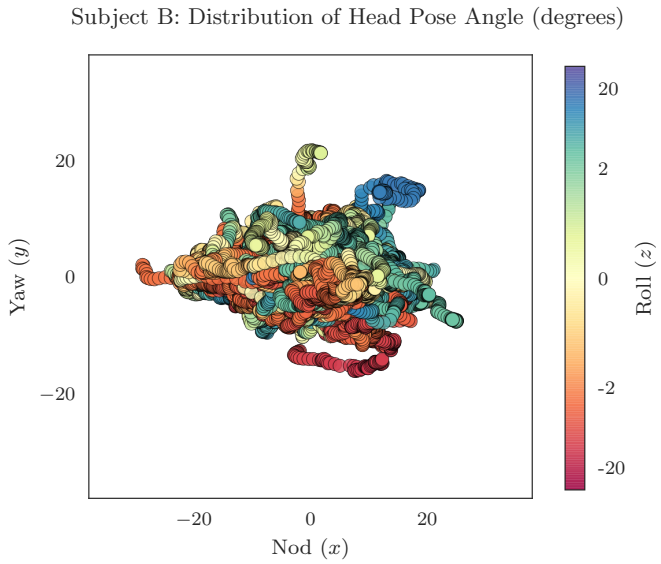
In Figures 6.3a and 6.3b, we plot the distribution of speaker head pose for Subject A and B. Subject A has a wider distribution that is reflected in person as a more expressive and animated demeanour. There is some evidence of correlation in yaw and roll for both speakers, and this is slightly more pronounced in A over B. This observation is also made in a number of the example trajectory plots we show in this chapter. We show numerical statistics in Tables 6.1 and 6.2 that confirm the visual information. The nod is limited for both subjects by normal anatomical limits as the head tilts downward.

6.2.1 Data Augmentation

Supervised learning depends on data, and for many tasks, it is often difficult to collect sufficient quantities of unique samples. Data augmentation is an attempt to increase the number of samples by manipulating existing samples within a corpus. Our primary means



(a)



(b)

Figure 6.3: (a) Distribution of Subject A head pose angle. (b) Distribution of Subject B head pose angle. The Nod angle is constrained in the positive angle (head down), by normal anatomical limits. Attention is drawn to the greater range of pose for Subject A, which can be attributed to more expressive behaviour when Subject A is speaking. The Roll (z) axis is described by the colour bar, that has linear scale in $[-2, 2]$ and log scale outside of that interval to avoid too many light colours.

Table 6.1: Distribution of Subject A head pose angle during speech. We show maximum, minimum, mean and standard deviation for each of Nod, Yaw and Roll. The angle unit is degrees.

Subject A	Nod (x)	Yaw (y)	Roll (z)
Min	−36.38	−15.00	−28.55
Max	23.47	13.30	23.57
Mean	−0.00	−0.00	0.01
STD	5.92	2.69	3.66

Table 6.2: Distribution of Subject B head pose angle during speech. We show maximum, minimum, mean and standard deviation for each of Nod, Yaw and Roll. The angle unit is degrees.

Subject B	Nod (x)	Yaw (y)	Roll (z)
Min	−28.76	−16.15	−17.85
Max	24.96	21.60	19.63
Mean	0.01	0.00	0.00
STD	5.31	2.36	2.77

of augmenting data is to define a sample as a short period that is a sub section of an entire utterance. In this way, we can increase significantly the number of unique samples presented at training time, and we found this an essential part of the training regime. In addition, we trained our models with *dropout* (Section 5.6), where sample points in a single exemplar are randomly set to zero. This was also an important part of training, and as well as augmentation, provides regularisation.

We also considered augmentation of the audio component of each sample. Our data always consists of speech synchronised with the motion of head pose. We made an observation during audio feature development (Chapter 4), that perturbing the interval of the filter banks was equivalent to the same utterance made with different frequency energies. Although the method was able to provide a considerable number of additional samples, we did not gain any training advantage, and could not detect any difference compared with simply training for longer.

6.3 Evaluation and Existing Baselines

Previous authors have published quantitative results on head pose prediction by reporting Root Mean Square Error (RMSE) for angle and Canonical Correlation Analysis (CCA) for correlation. From the literature we show baselines in Table 6.3. Only Ding et al. [2015] report RMSE, however they discuss standardised scaling of the head pose trajectories for model training, but do not state if the RMSE is for scaled or real world angle values. Haag and Shimodaira [2016] report local CCA for a 300 frame window at 120 FPS. To remove all doubt, our results are reported for the reconstructed predictions, rescaled to original scale. We report CCA for an entire utterance for all rotation axes, without any truncation. When we show collated results, we show the mean of these values for all test examples.

Table 6.3: Baseline results for head pose prediction. Showing the best results from the cited works. Only Ding et al. [2015] report RMSE, they discuss scaling and it is unclear if this figure is for scaled trajectories.

Author	RMSE	CCA
Ding et al. [2014]	N/A	0.561
Ding et al. [2015]	0.775	0.711
Haag and Shimodaira [2016]	N/A	0.390

Qualitative assessment is less consistent in the literature, but a number of authors show plots of ground truth trajectories with predictions, but perhaps with only one axis of rotation e.g., [Deng et al., 2004, Figure 7], [Ding et al., 2014, Figure 6]. To offer the most complete qualitative assessment we plot results of predictions with the ground truth for all of nod, yaw and roll axes (x, y, z). We always show the full duration of any test utterance and our head pose trajectories will be presented in real world angle values in degrees. We also show the time domain audio in the same plot, so the reader may assess when events occur in the audio, the ground truth, and the prediction. We feel this makes the clearest qualitative evaluation of a prediction and when used in conjunction with the quantitative measurements gives a good impression of our results.

6.3.1 Canonical Correlation Analysis (CCA)

CCA [Hotelling, 1936] measures the linear relationship between two multi-dimensional variables. For each of the two variables it finds the basis vector, such that the projections of the variables on to these bases are optimal with respect to correlation. For two random variables, $\boldsymbol{x}$ and $\boldsymbol{y}$ with zero mean, the covariance matrix is shown in Equation 6.1.

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{xx} & \boldsymbol{C}_{xy} \\ \boldsymbol{C}_{yx} & \boldsymbol{C}_{yy} \end{bmatrix} \tag{6.1}$$

The canonical correlations between $\boldsymbol{x}$ and $\boldsymbol{y}$ can be found by solving for the eigenvectors, shown in Equation 6.2.

$$\begin{cases} \boldsymbol{C}_{xx}^{-1}\boldsymbol{C}_{xy}\boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{w}_x &= \rho^2\boldsymbol{w}_x \\ \boldsymbol{C}_{yy}^{-1}\boldsymbol{C}_{yx}\boldsymbol{C}_{xx}^{-1}\boldsymbol{C}_{xy}\boldsymbol{w}_y &= \rho^2\boldsymbol{w}_y \end{cases} \tag{6.2}$$

where the eigenvalues ρ^2 are the squared *canonical correlations* and the eigenvectors $\boldsymbol{w}_x$ and $\boldsymbol{w}_y$ are the normalised canonical correlation *basis* vectors.

In this study, we always project to a single base, and measure Pearson's product-moment correlation coefficient, or Pearson's r, using Equation 6.3 on those projections.

$$r = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y-\bar{y})^2}} \tag{6.3}$$

To allow firmer conclusions of the value of CCA, as a measure of the effectiveness of a prediction, we perform the analysis against two types of synthetic data. We choose a sinusoid with a frequency of 1Hz, and a linear value in the interval $[-1, 1]$, shown in Figure 6.4. Table 6.4 shows the results of this comparison for each of our test scenes. As well as CCA correlation, shown as r in the table, we show the p value. The p value is the probability one would have found the current result if the correlation coefficient were actually zero - the *null hypothesis*. If $P < 0.05$ then the correlation coefficient is statistically significant. One can clearly see high correlation for linear data, whereas the sinusoidal data, that is closer in

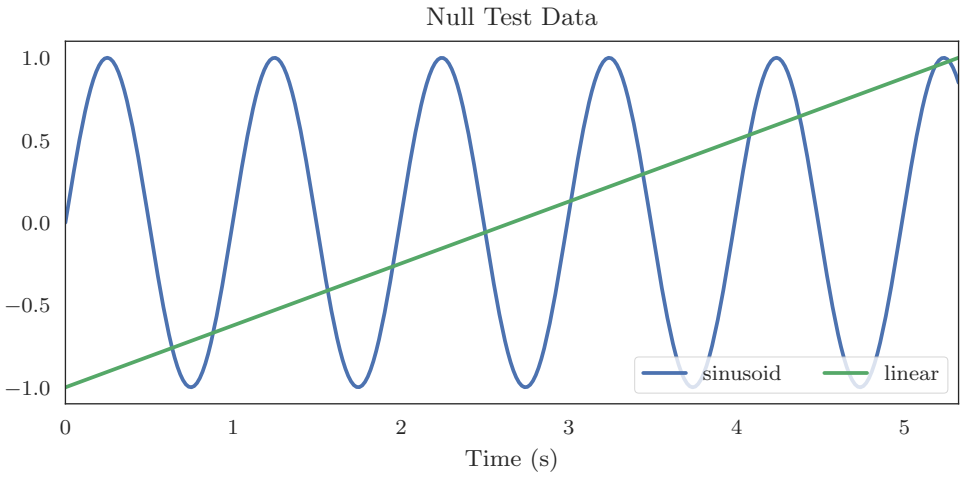


Figure 6.4: To assess the value of CCA as a measure of the effectiveness of a prediction, we perform the analysis against two types of synthetic data.

character to quasi-periodic head pose trajectories, has significantly lower correlation. This demonstrates that we need to take into consideration the qualitative results shown in the plots, and the results of subjective tests, to get a full picture of a prediction.

Table 6.4: We show results for CCA of each test scene with a sinusoid and a linear variable. In addition to the CCA correlation, r, we show the p value.

Scene ID	sinusoid		linear	
	r	p	r	p
A-01-0184	-0.32	3.46×10^{-09}	0.76	4.97×10^{-63}
A-02-0120	-0.33	2.96×10^{-06}	0.91	8.15×10^{-77}
A-03-0203	0.18	5.65×10^{-03}	0.62	3.00×10^{-27}
A-04-0056	0.38	6.27×10^{-09}	-0.75	5.73×10^{-42}
A-05-0263	-0.28	1.42×10^{-06}	0.65	3.24×10^{-37}
A-06-0089	0.52	8.70×10^{-14}	0.90	3.81×10^{-67}
J-01-0153	0.30	2.55×10^{-05}	0.64	3.07×10^{-23}
J-02-0089	-0.39	1.52×10^{-10}	0.71	8.75×10^{-40}
J-03-0263	-0.26	1.61×10^{-05}	0.85	1.23×10^{-73}
J-04-0052	-0.22	7.43×10^{-05}	-0.88	2.04×10^{-16}
J-05-0256	0.22	1.56×10^{-03}	0.90	1.87×10^{-76}
J-06-0276	-0.55	1.08×10^{-14}	0.96	2.09×10^{-92}

6.3.2 Head Pose Expectation

What might one expect to predict? Using our evaluation criteria, can we determine if we have synthesised a plausible head pose sequence? One comparison we can draw is to look at what happens when a speaker repeats the same utterance. This is clearly not the same as an alternative head pose for the same audio, that ground truth can never exist. Figure 6.5 shows the same utterance, or rather the same *transcript*, repeated three times by Subject B.

> "But I want to be a real person with hair that grows and skin that sweats and a heart that beats!"

The three utterances were not of equal length so were interpolated to the mean length of the three, ensuring all series were processed somewhat. We were then able to make an empirical comparison for each sequence which we show in Table 6.5. Figure 6.5 is quite revealing of the *style* of head pose. Of the three utterances the first, scene *J-05-0074*, is least similar in delivery, yet all utterances have in common a nod trajectory that has three prominent peaks. Yaw is more interesting. Between 1 and 4 seconds, in scene *J-05-0074*, the Yaw is roughly parallel to nod, or highly correlated. In scene *J-05-0076*, for the same sub-sequence, yaw has a symmetry to nod. In the last example, scene *J-05-0078*, the yaw changes from parallel to symmetry. What is clear from the plots, is the way head pose is modulated by the speech signal, but does not have a linear relationship. When we

Table 6.5: Comparing the same transcript repeated by Subject B. These values indicate a reasonable target to achieve when we synthesise predictions.

Scene ID	*J-05-0074*		*J-05-0076*		*J-05-0078*	
	RMSE	CCA	RMSE	CCA	RMSE	CCA
J-05-0074	0.00	1.00	2.06	0.75	2.38	0.76
J-05-0076	2.06	0.75	0.00	1.00	2.27	0.86
J-05-0078	2.38	0.76	2.27	0.86	0.00	1.00

look closely at the quantitative values in Table 6.5, we first acknowledge that RMSE and

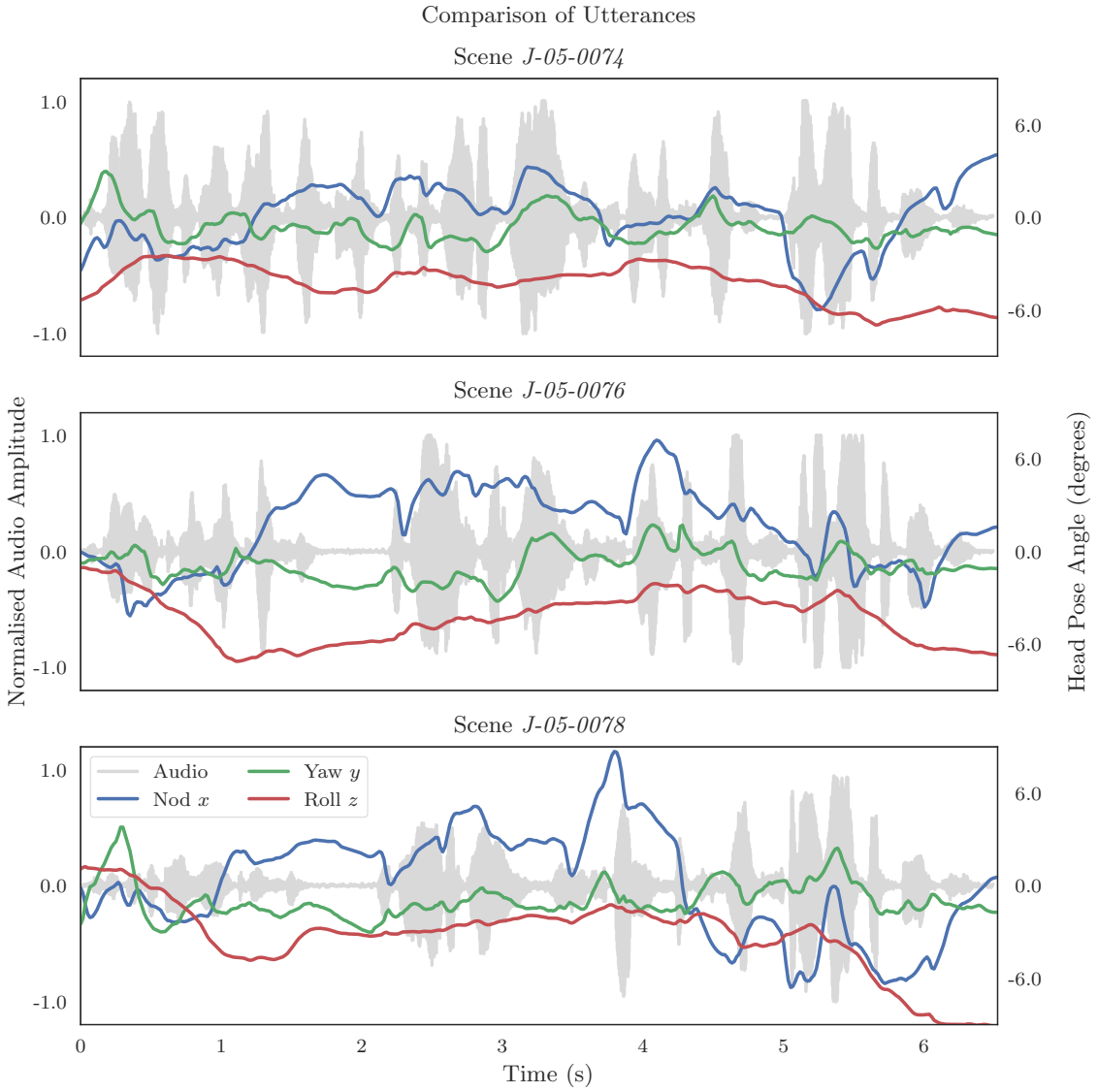


Figure 6.5: Comparing the same transcript repeated by Subject B from the corpus. "But I want to be a real person with hair that grows and skin that sweats and a heart that beats!" The three utterances were not of equal length so were all interpolated to the mean length.

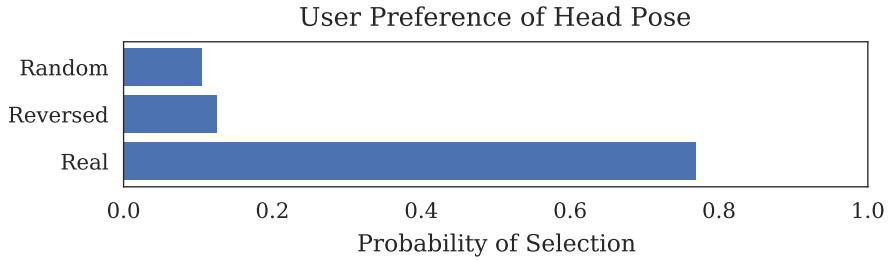


Figure 6.6: User preference of head pose. Users strongly prefer correct head pose with speech, motivating the synthesis of plausible head motion for speech animation.

CCA are 0 and 1 when we compare an utterance to itself. Secondly, if a human speaker repeats a transcript, we might expect that they would be somewhat similar. The plots of the trajectories confirm *similarity*, the values indicate a reasonable target to achieve when we synthesise predictions.

6.3.3 User Preference

We conducted a user study to understand the importance of rigid head pose when presenting speech animation. We collected data by asking a speaker to wear an Inertial Measurement Unit (IMU) based on a Sony PS3 motion controller, and recorded 3 DoF of nod, yaw, and roll of head pose while speaking. The dialogue was emphatic, prosodic and descriptive of a recent car journey. We segmented the data to 30s clips, and applied the recorded sound to a simple 3D model without any facial features to minimise perceptual noise. In addition to the recorded motion, we showed the same audio with the motion reversed, and, random head positions that had no association with the audio. The results of the test are shown in Figure 6.6, and we see a clear preference for the proper motion, providing further motivation for plausible head motion synthesis.

6.4 Model Topology

All of our modelling strategies feature BLSTM networks, which we cover in more detail in Chapter 5. An illustration of the core model topology is shown in Figure 6.7.

Our model accepts speech features as input, and predicts real values of head pose Euler angles. We use a many to many topology, emitting the state of the model at every time step at each layer of the model.

When considering model size, our aim is to minimise the number of trainable parameters in the model without compromising the ability of the model to make useful prediction. We adopt a strategy of starting small and increasing width and depth while managing over fitting. Curiously, the first model we show here is much smaller than others report in the literature. For example, Ding et al. [2015] refers to networks with hidden layers between 128 hidden units and 1024 hidden units. We show a model for our first speaker, Subject A, of 3 bi-layers of 32 hidden units. The impression formed over several model types and training periods is that model topology will be determined by the quantity and quality of the data. An heuristic that emerged from experience, dictates that the total number of trainable parameters is limited by the number of training examples to no more than one order of magnitude greater than the number of examples, lest the model will fail to converge.

6.5 Bi-Directional Long Short Term Memory

Bi-Directional Long Short Term Memory (BLSTM) are a suitable choice for modelling speaker head pose. Certainly, changes in speech characteristics are correlated with head pose. We can also say at any time the kinetics; the position, velocity and acceleration of the head, are influenced by prior and future motion. The BLSTM develops memory of both the speech and head motion in the past *and* the future.

Our first model is BLSTM, that we train using LogfBank features to predict real head pose Euler angles. Early experiments were conducted to determine the best audio features to select for this task. Many feature types could be rejected, but a clear advantage is not shown by any of MFCC, fBank or LogfBank with or without delta 1 and 2. We select LogfBank as our standard audio feature for two reasons: firstly, it has been shown the decorrelation of the DCT stage of MFCC processing is not necessary in speech related tasks with DNN [Deng et al., 2013b], secondly, the log of fBank reduces the feature dominance of lower frequency values (although arguably, our standardised scaling would also remove this aspect of the feature vector). We examine feature extraction in detail in Chapter 4.

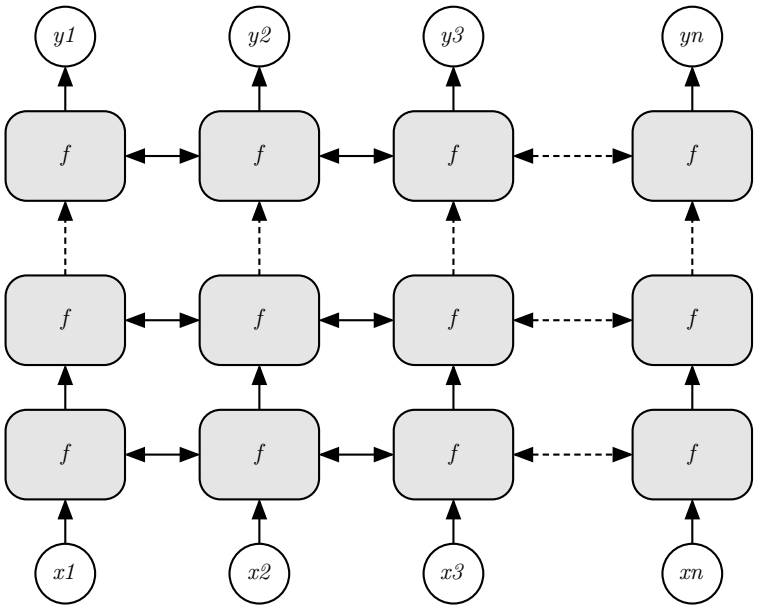


Figure 6.7: Modelling head pose with deep BLSTM. We show a model trained with X speech features to predict Y head pose Euler angles. At each time step we emit the state of the network, in a many to many model.

6.5.1 Training

We trained the networks on our data, split 90% for training, 10% for validation and we extracted a small random selection of utterances distributed uniformly across the corpus

that were never used for training nor validation and model selection. We scale the features such that each feature has zero mean and unit variance. We also scale the target rotation trajectories in the same way, and remove scaling after prediction with the inverse of our scaling function. We regularise during training using dropout (Section 5.6) with a value of 0.5. Our objective function is MSE. Our optimising function is *RMSprop* [Tieleman and Hinton, 2012], we set an initial learning rate of 10^{-3}. Training continues until no further improvement on the validation set is achieved, with a patience of 10 epochs. Model weights are saved at each epoch. We reload the best weights, decrement the learning rate by a factor of 10 until 10^{-5}, finally stopping at the best validation error. We then select the model with the lowest overall validation error. We augment our data as described in Section 5.7 and set $n = 129$ time steps to capture long term events.

6.5.2 Audio Features

We first show results for prediction of head pose from a model trained on audio features for Subject A. To be clear, this is a single speaker model. Figure 6.8 shows predictions for a deep BLSTM trained on the LogfBank features of Subject A as input, with the head pose trajectories of Subject A as the objective. Immediately we observe, that head pose is modulated by the audio in a similar way to the ground truth. Of particular note is the first example scene (*A-01-0184*), that shows the nod (x) angle switching from mirroring to following the ground truth, in much the same way as we observed in the true utterances shown in Figure 6.5. We show a second example from our single speaker BLSTM model in Figure 6.9. We choose this example as the CCA correlation value is the worst in this test set. Regardless, many of the key events in the ground truth are also represented in the prediction, and the range of motion is very much in keeping with the original. Viewing the plot also allows us to appraise the quality of the trajectory. We also remark, there is no post filtering of the result, and show the direct output of the model, simply removing the scaling we applied for training.

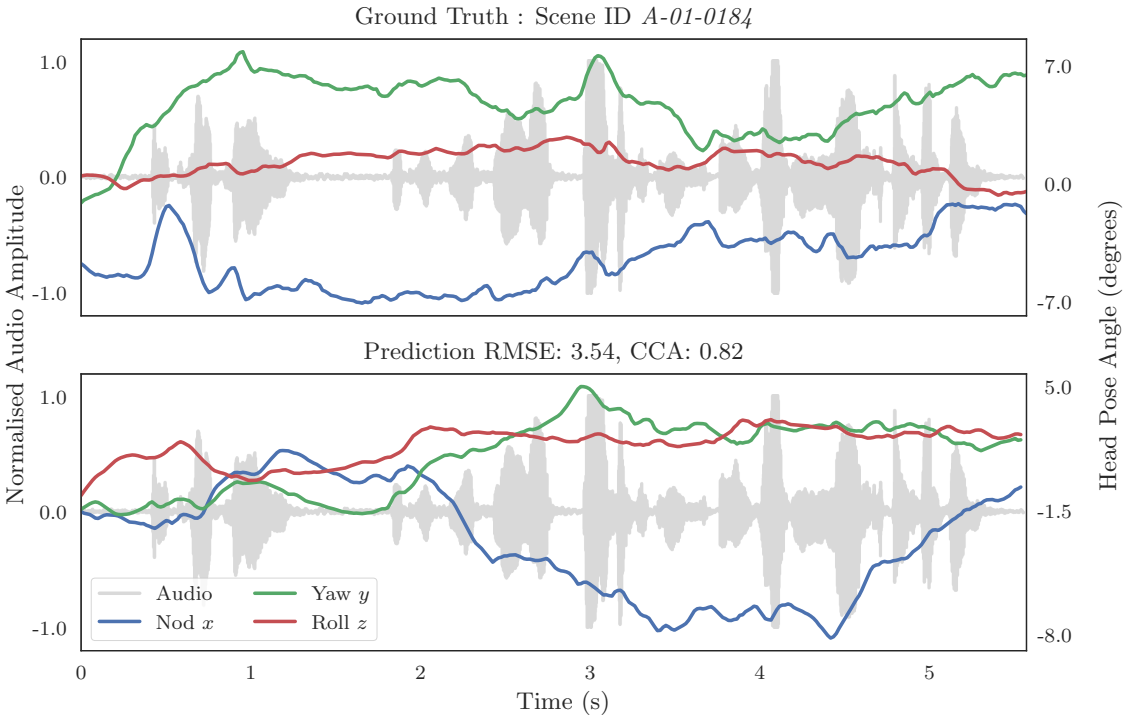


Figure 6.8: Head pose results for Subject A from audio features. We show the ground truth of the utterance with the model prediction. We show the trajectories with the time domain audio to present a qualitative assessment to support the empirical scores of RMSE and CCA. Speech to head pose is a many to many mapping so we do not expect predictions to follow ground truth sample by sample. We do expect predictions to have similar characteristics of the ground truth, and to be modulated by the events in the audio signal.

Table 6.6: Head pose results for Subject A from audio features. This is a BLSTM model trained on a single speaker, using LogfBank features to predict head pose angles.

Scene ID	*A-01-0184*	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
RMSE	3.54	4.45	2.19	3.26	3.34	1.83
CCA	0.82	0.96	0.69	0.92	0.92	0.92

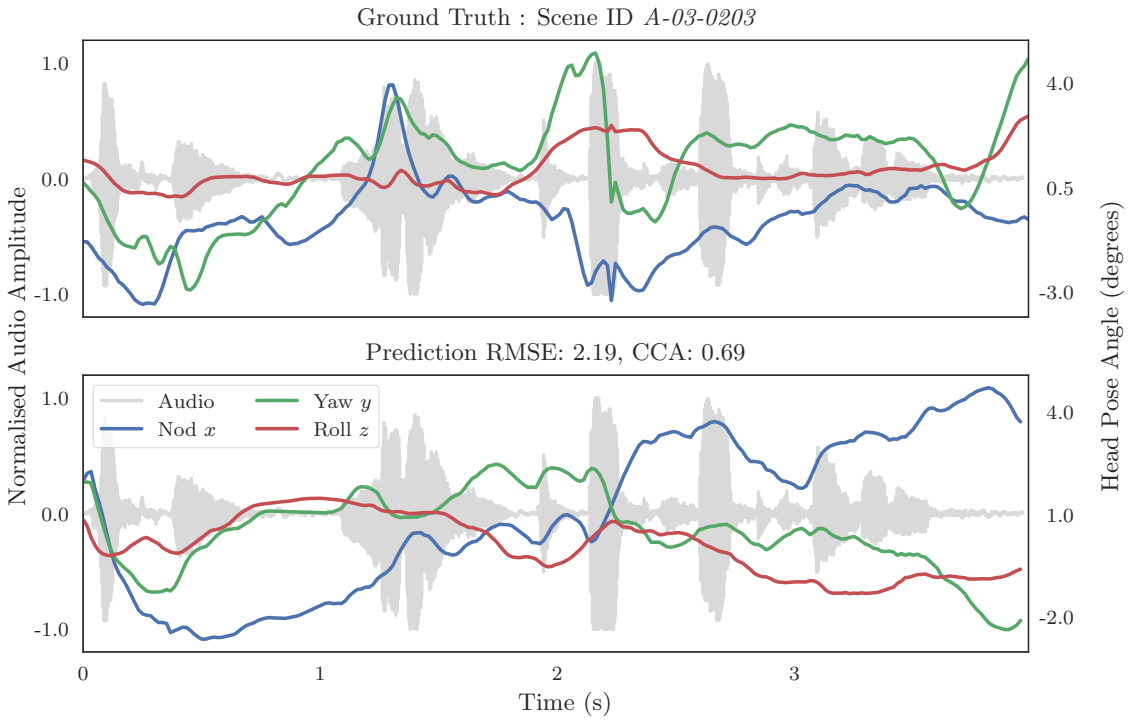


Figure 6.9: Head pose results for Subject A from audio features. We show the ground truth of the utterance with the model prediction. The plots provide a qualitative assessment to support the empirical scores of RMSE and CCA. In this second example, we show a result that has lower correlation, yet even here the trajectory is appropriately modulated by the audio.

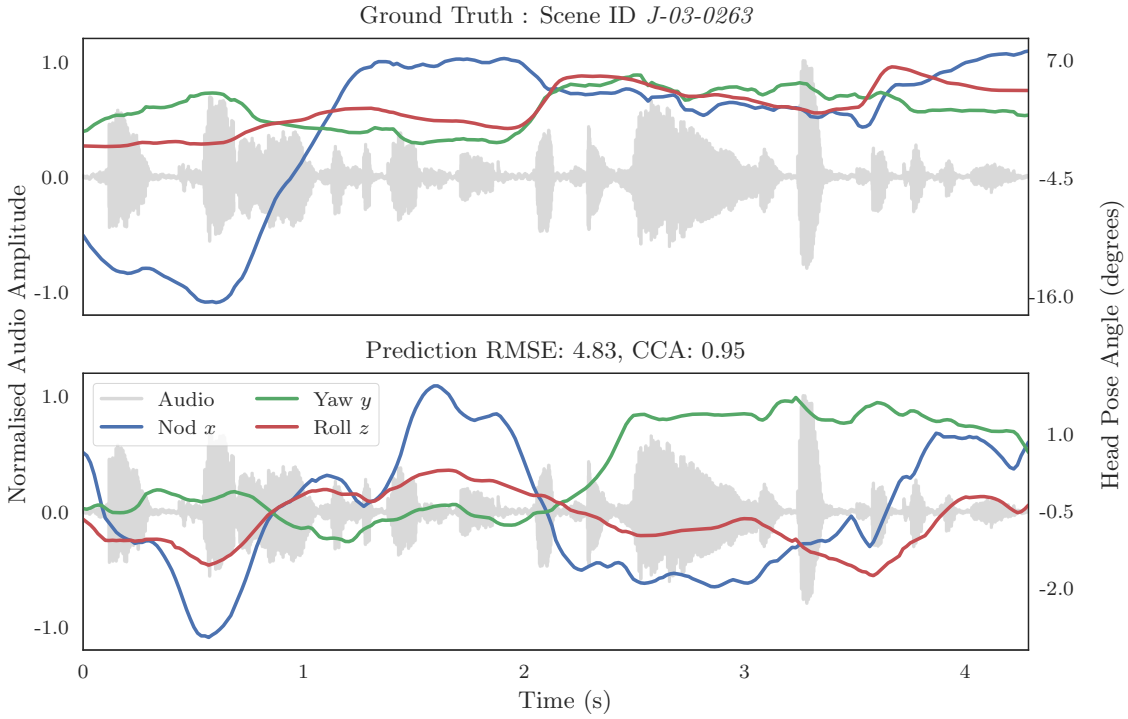


Figure 6.10: Head pose results for Subject B from audio features. We show the ground truth of the utterance with the model prediction. In this figure, we can see again how the audio modulates the head pose. Overall the results are not quite as convincing for Subject B.

Table 6.6 collates the empirical measurements of the test examples. All of the examples show high correlation. Considering the RMSE, this is the most discriminating measurement if we wish to predict a sequence exactly as the ground truth, but this is rarely the case. Parallel offsets of a sequence of values may give an error of two or three degrees, whereas such an offset in animation is not perceptually significant. CCA is probably the most valuable quantitative measure, projecting the multi-variate time series to a single base.

We now show results for prediction of head pose from a second model trained on audio features for Subject B. This model has the same topology as the Subject A model, 3 layers of 32 hidden units in each forward and reverse direction. Our training regime remains the same as described earlier. Again, this is a single speaker model.

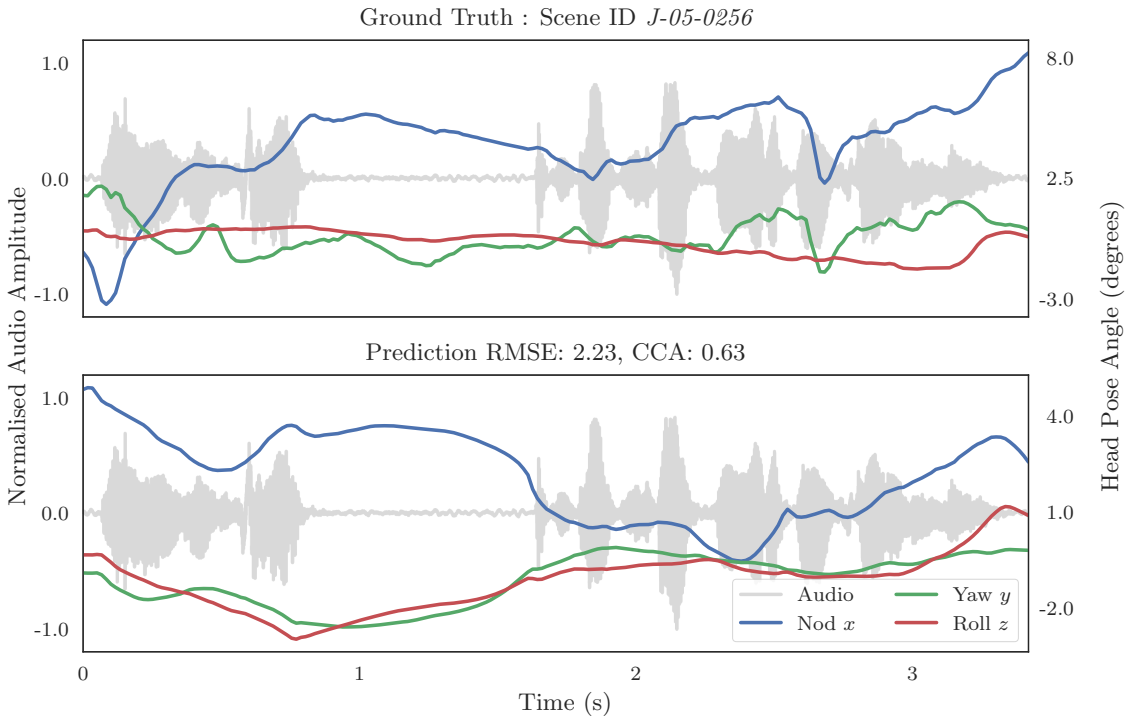


Figure 6.11: Head pose results for Subject B from audio features. We show the ground truth of the utterance with the model prediction. In this figure, we can see again how the audio modulates the head pose. Overall the results are not quite as convincing for Subject B.

Table 6.7: Head pose results for Subject B from audio features. This is a BLSTM model trained on a single speaker, using LogfBank features to predict real head pose angles.

Scene ID	*J-01-0153*	*J-02-0089*	*J-03-0263*	*J-04-0052*	*J-05-0256*	*J-06-0276*
RMSE	3.75	1.78	4.83	4.06	2.23	2.15
CCA	0.93	0.91	0.95	0.92	0.63	0.87

Figures 6.10 and 6.11 show the plots of two examples for our second subject. When we scrutinise the plots, we see that the head pose angles have a similar behaviour to the ground truth, but do not closely follow the original trajectories. We should not expect that they should, as we have previously shown the variance in head pose for the same transcript (Section 6.3.2).

For the Subject B audio model, we collate results for our test samples in Table 6.7. This model shows similar performance as the Subject A model, confirming our choice of modelling strategy. We get poorer performance if we concatenate the data for both speakers. Each subject has a definitive characteristic motion which we can identify by looking at the global statistics of the head pose angles in Figures 6.3a and 6.3b, and in Tables 6.1 and 6.2. When we train our model on both subjects the predictions become less dynamic than either, with the model heading toward the mean. We speculate that it is possible that a much larger data set may allow training of a larger, more expressive model, that could mitigate this situation and directly learn how to separate identity. A larger corpus is not available, so we must consider other approaches.

6.5.3 Phone Features

Training a model on Audio features is perhaps the most desirable work flow. Processing audio on modern hardware is fast and convenient. One drawback is undoubtedly how much identity is embedded in audio, it is very easy for a human listener to distinguish between two speakers, for example. Our neural network models require training on one speaker to effectively predict that same speaker's actions, at least from our experiments with our corpus. We described the extraction of phoneme features, or more accurately, phone features in Section 4.4. The key potential of phone based features, or any non-audio feature, is a degree of automatic speaker normalisation. Concretely, any speaker can read from a text, but the text is the canonical version of the utterance.

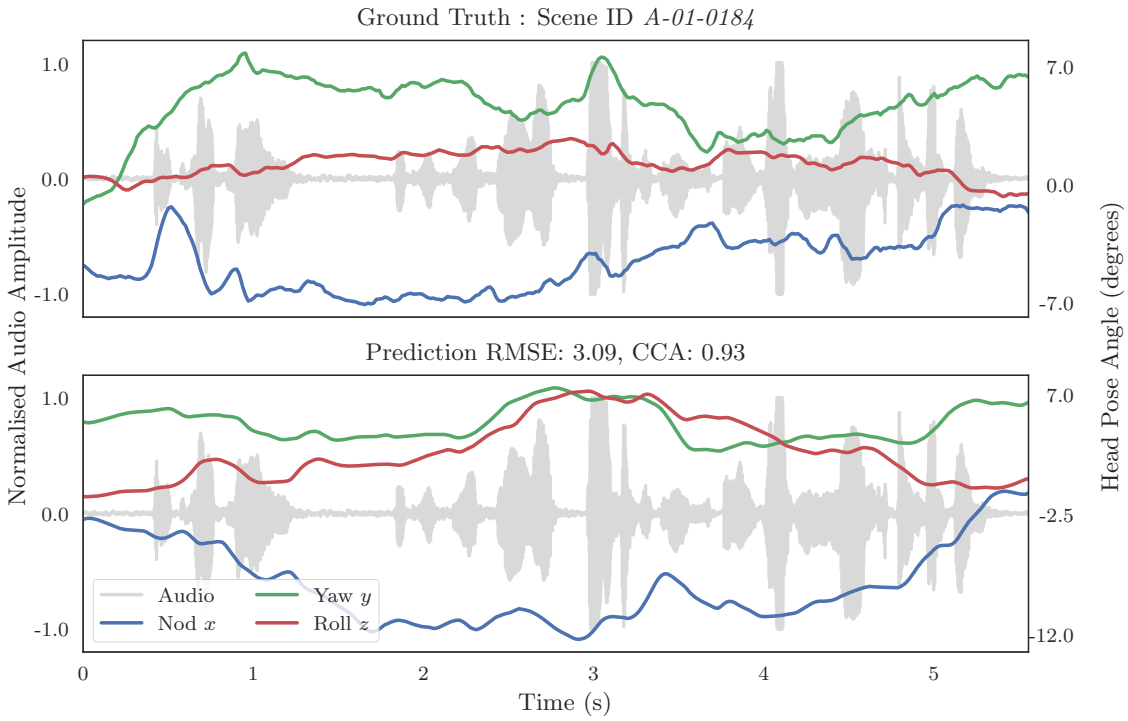


Figure 6.12: Results for Subject A from our BLSTM model trained on phone features. On each plot we show the ground truth and prediction angles for each of Nod (x), Yaw (y) and Roll (z). We show the time domain audio waveform to show how the head pose angle is modulated by the activity in the speech.

We implement our model with the same architecture as our previous audio features model. A stack of BLSTM layers, with each hidden state feeding forward to the next layer. To make direct comparison, we also retain the same size of model, which is 3 layers of 32 hidden units, in each direction. Again, we use dropout to regularise the training, with a value of 0.5. We refer to the same schematic in Figure 6.7, but this time the input speech features are the temporally aligned phones (Section 4.4). We show results for our model, trained on phone features to predict rigid head pose for Subject A, in Figures 6.12 and 6.13, and in Table 6.8.

We can see, for example in Figure 6.12, that we get generally good predictions of head pose for phone features. Closer scrutiny of the plot suggests that the model produces

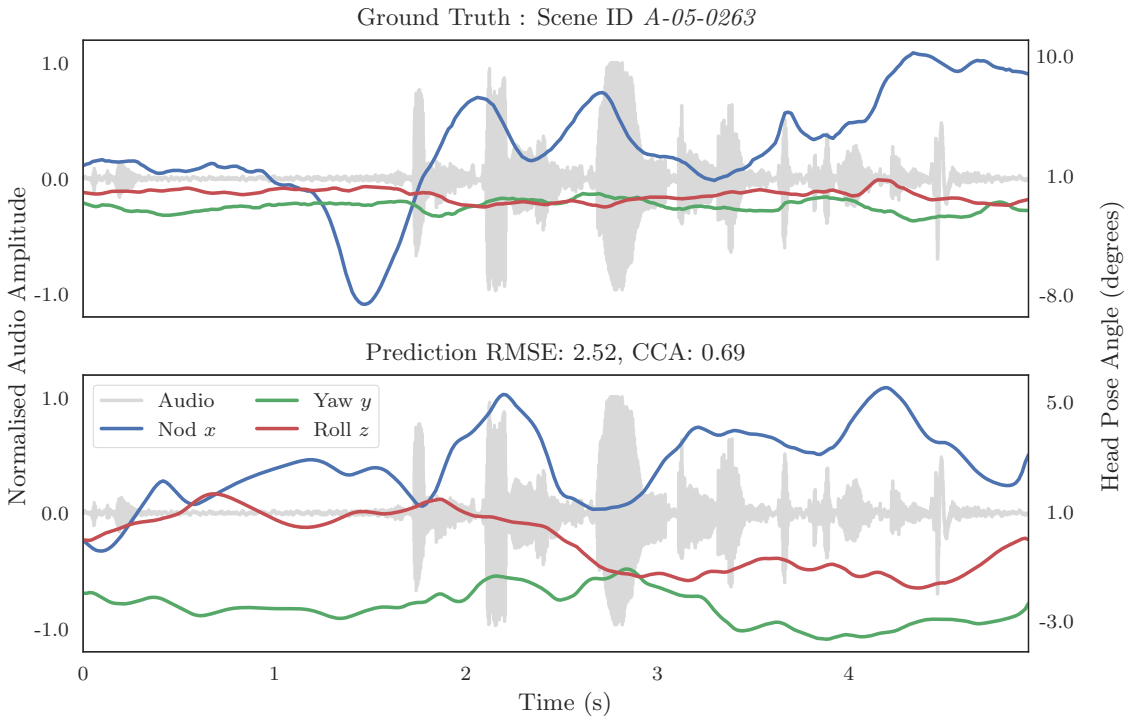


Figure 6.13: Results for Subject A from our BLSTM model trained on phone features. On each plot we show the ground truth and prediction angles for each of Nod (x), Yaw (y) and Roll (z). We show the time domain audio waveform to show how the head pose angle is modulated by the activity in the speech.

Table 6.8: Head pose results for Subject A from phone features. We show RMSE angles in degrees and correlation using CCA for our test scenes. All test scenes have been permanently excluded from training, validation and model selection.

Scene ID	*A-01-0184*	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
RMSE	3.90	2.99	4.43	4.08	2.48	2.01
CCA	0.82	0.92	0.77	0.89	0.77	0.90

results that are slightly 'smoother' than those produced by the audio features model in the previous section. One possible reason for this, is that despite the model topology being the same, the size of the input feature is not. Our LogfBank feature vector has 40 features (derived from 40 filter banks), and our phone feature vector has 71 features; the length of the One Hot Encoding (OHE) vector for each of the phone categories with the lexical stress markers. Of course, any change in the structure of a model, results in entirely different weight convergence. Regardless, the small difference in the qualitative aspects of the prediction are not necessarily a model failure, we see 'smoother' trajectories in ground truth examples too. Figure 6.13 is an example that shows similar qualities in the ground truth. When we look at the prediction in this example, we see that the model captures the dynamic qualities of the original, including much of the relatively large range of movement.

The motivation to use phone features is to remove, or at least reduce, speaker dependence while accepting the increased processing to produce this type of feature. Concretely, we want to be able to use the data from both the speakers in our corpus to extend the training of our model. It transpires the phone model has similar limitations in this regard to the first audio features model. There are some interesting possible explanations for this. The BLSTM retains information from past (and future) input, and also *output*. We suggest that the identity of the speaker is also embedded in the characteristics of the head motion. We can see that to separate the speakers using global statistics (Section 6.2) is a trivial task, with Subject B having markedly less dynamic head motion. Another possible factor is that phone features do not remove identity in the way we expected. Phonemes are linguistic units, that exchanged with another change the meaning of a word in a language. Phones are distinct speech sounds without regard to the meaning of the spoken word. It is possible our feature extraction method retains differences in the sound each speaker makes. Finally, we emit phone features over time, embedding some of the speaking *style* of the subject. We need to consider another approach to take advantage of both our subjects for model training.

6.6 Conditional Variational Auto Encoder

In the past few years, generative models [Kingma et al., 2014; Rezende et al., 2014], trainable with back propagation [Bengio et al., 2014] have taken an important step in learning, with models that can perform probabilistic inference and make diverse predictions. For example, Bowman et al. [2016] employed a VAE for natural language generation and Walker et al. [2016] used a CVAE to predict video motion vectors conditioned by a single image. To our knowledge, generative models have not yet been used for head motion prediction so we introduce a CVAE to the head motion synthesis task here [Greenwood et al., 2017a].

We discuss details of the CVAE in 5.5.3, specific to our implementation. Here, we draw attention to the topology of the model we develop specifically for the purpose of predicting the rigid pose of the speaker's head from audio features. Our model has a tapering stack of BLSTM layers in the *encoder* of 128, 64, 32 hidden units for each direction. We use a latent vector size of 6, and a *decoder* with BLSTM layers of 128, 64, 32 hidden units for each direction. We use dropout of 0.5 to regularise the model during training. In Figure 6.14, we show how the model is conditioned at every time step with speech features, specifically LogfBank features.

6.6.1 Training

The CVAE model is trained by splitting our data 90% for training, 10% for validation. We have the same previously removed random selection of utterances distributed uniformly across the corpus that were never used for training, nor validation and model selection. Our standard technique is to scale our input features and output target such that they have zero mean and unit variance for training. We de-scale the target rotation values after prediction, using the inverse of the scaling function, so they are restored to real world values. Our optimising function is *RMSprop* [Tieleman and Hinton, 2012] and we set a learning rate of 10^{-3}. The network is trained until no further improvement on the validation examples

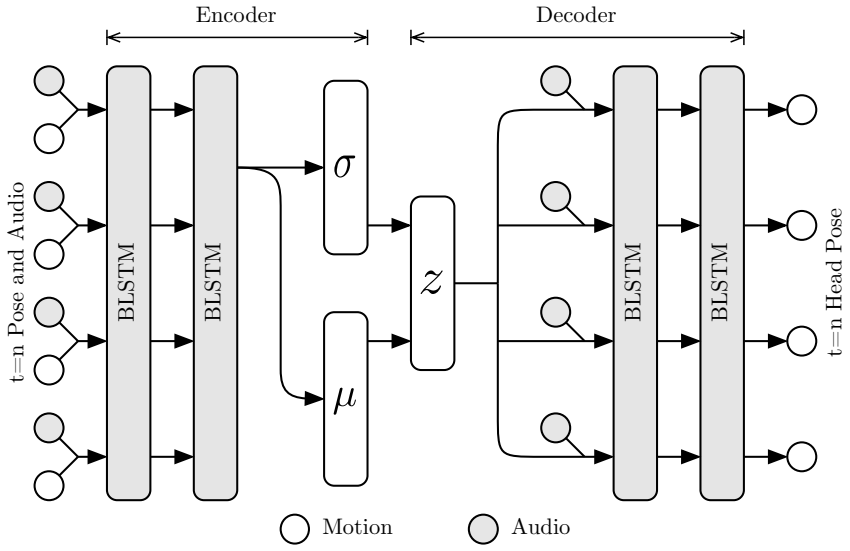


Figure 6.14: A sketch of the topology of the CVAE model we use to predict speaker head pose. The model is conditioned on LogfBank features at every time step.

with a patience of 10 epochs. We augment our data as described in Section 5.7 and set $n = 129$ time steps to capture long term events.

For this model we train a *single* model on the speech and head pose of *both* subjects. This is one of the goals of this approach, to take advantage of the additional training of both our subjects combined. This also partly explains the larger number of trainable parameters in this model compared with the simpler model of Section 6.5.

6.6.2 CVAE Results

So we can make a direct comparison, in Figure 6.15 we show the same example utterance reviewed for our two previous models. Again, we see that the head pose trajectories correlate with events in the audio. The plot also shows that the character of the prediction is similar. This is an important aspect of viewing plots of the predictions. Point wise comparisons or global statistics do not reveal transitions from smooth motion to more staccato motion sections. In this example we have exactly this situation in the ground truth, and the

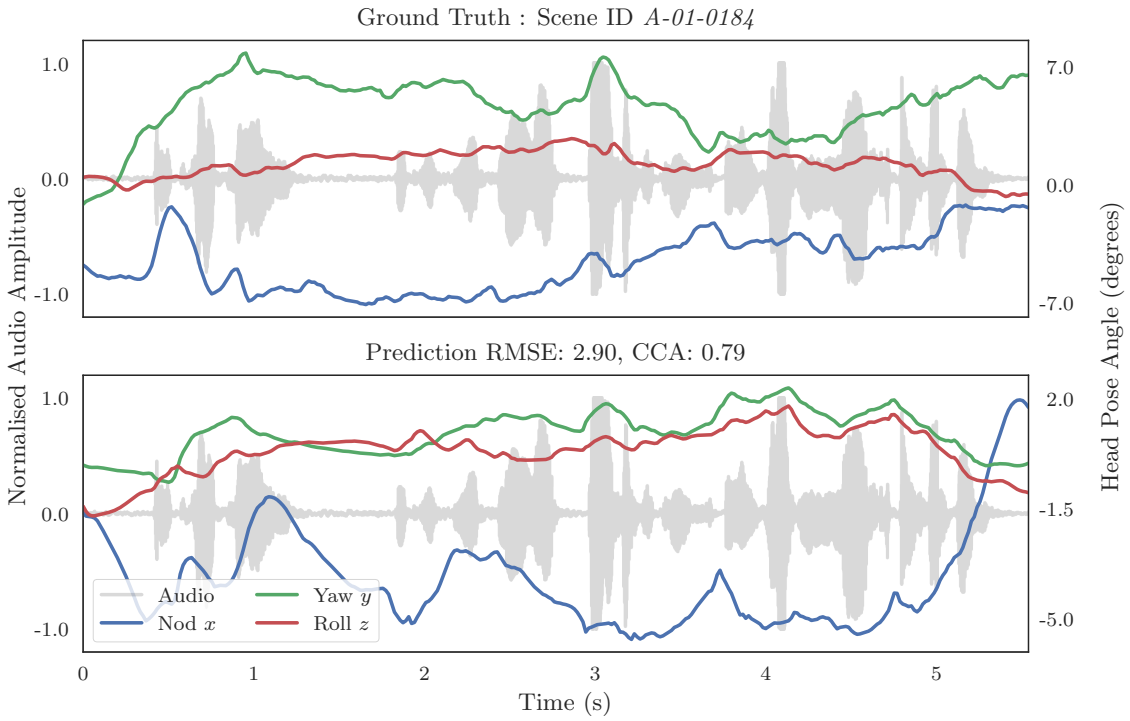


Figure 6.15: Results for Subject A CVAE model, example scene 1. Here we can see the head pose is modulated by the audio n the prediction, with similar characteristics to the ground truth.

prediction accurately reflects this. Another example for Subject A is shown in Figure 6.16. Although the CCA values are not the highest, this is not the only criteria we should consider. The plot shows again how well the prediction is modulated by the audio, with similar timing to the ground truth and similar correspondence to events on the audio. One of the goals of this model is to be able to exploit additional training data available across a range of speakers. We have two subjects in our corpus, but their identities are quite separate, one is male, one is female. They differ physically and also in character. A property of the CVAE model is the model learns the distribution of the subject's identity that is embedded in their motion characteristics. This identity is captured in the latent variable of the model. Figures 6.17 and 6.18 show plots for Subject B, predicted from the same model as Subject A. A particularly interesting part of the utterance in Figure 6.18, between 1 and 2 seconds, shows a pause in the delivery. This is anther example of a change in the style of the head

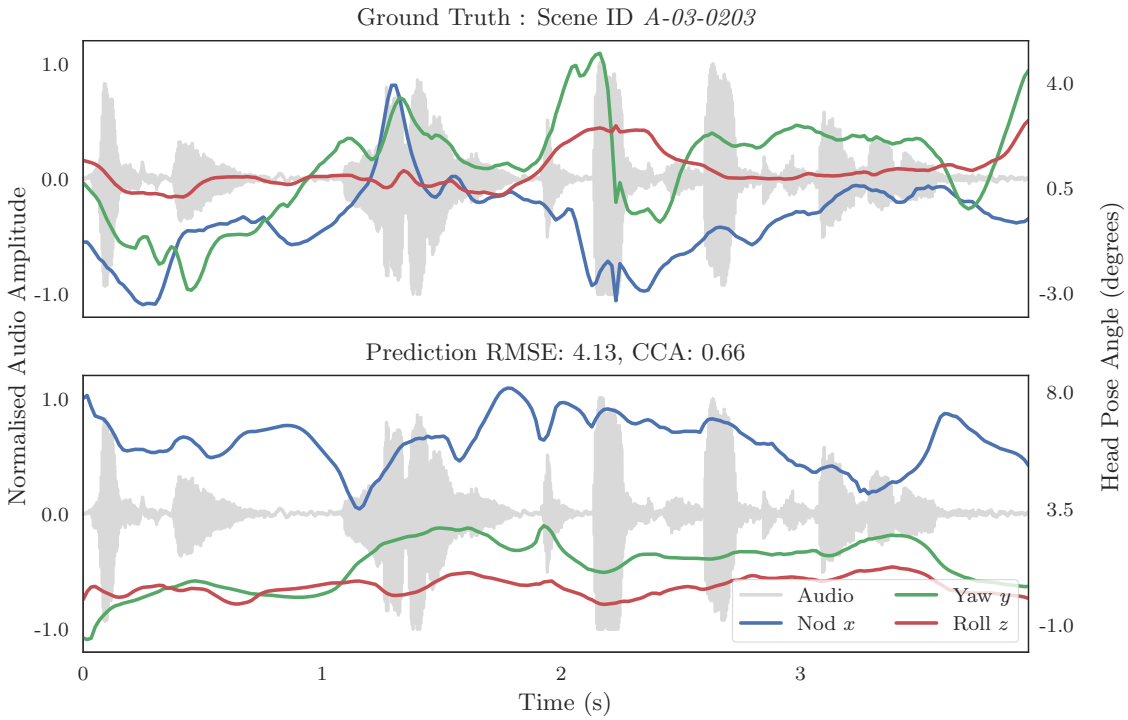


Figure 6.16: Results for Subject A CVAE model, example scene 2. Even though the CCA value is lower, the timing of events in the ground truth is reflected in the prediction.

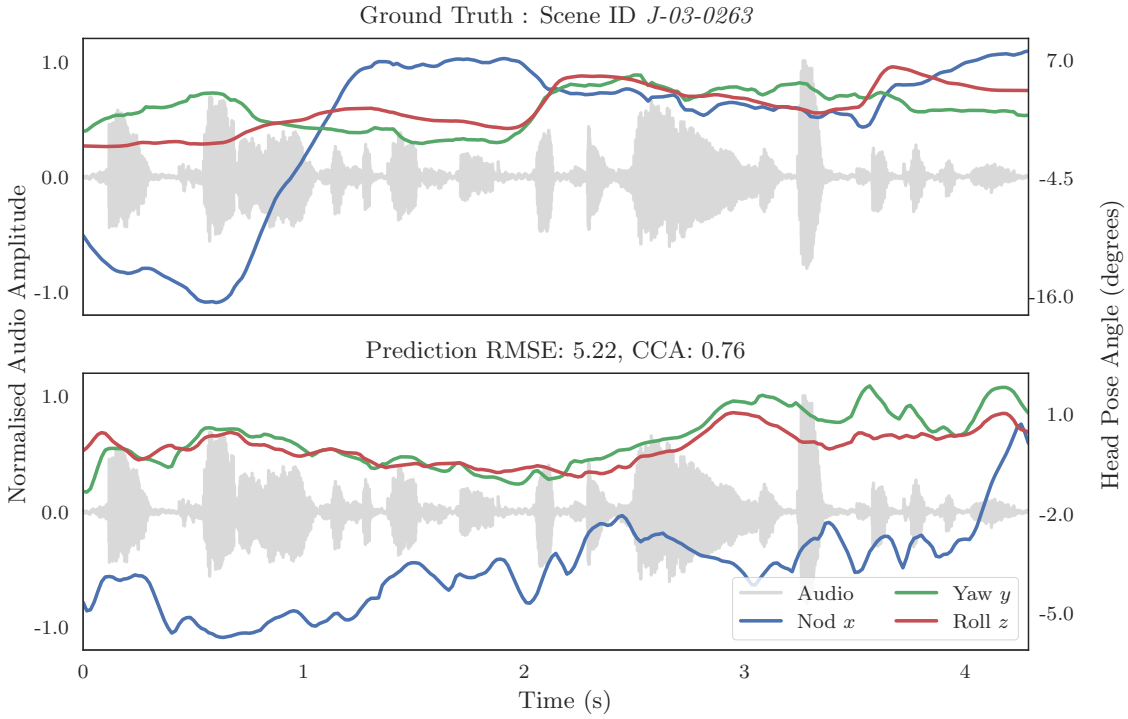


Figure 6.17: Results for Subject B CVAE model, example scene 1.

pose in the ground truth that is reflected in the prediction. We highlight the importance of showing the plotted trajectories for qualitative assessment, as statistical measures do not reveal these characteristics.

Table 6.9: Head pose results for Subject A for CVAE model. We show RMSE in degrees for rotation and millimetres for translation, and correlation using CCA for our test scenes. All test scenes have been permanently excluded from training, validation and model selection.

Scene ID	*A-01-0184*	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
RMSE Rotation	2.90	5.76	4.13	3.42	4.56	1.63
RMSE Translation	2.51	2.70	2.71	2.70	2.54	3.87
CCA Rotation	0.79	0.82	0.66	0.82	0.86	0.94
CCA Translation	0.79	0.83	0.68	0.66	0.72	0.79

We collate all the quantitative results for the CVAE model in Tables 6.9 and 6.10. We show results for both subjects predicted from one model. The collated results show good

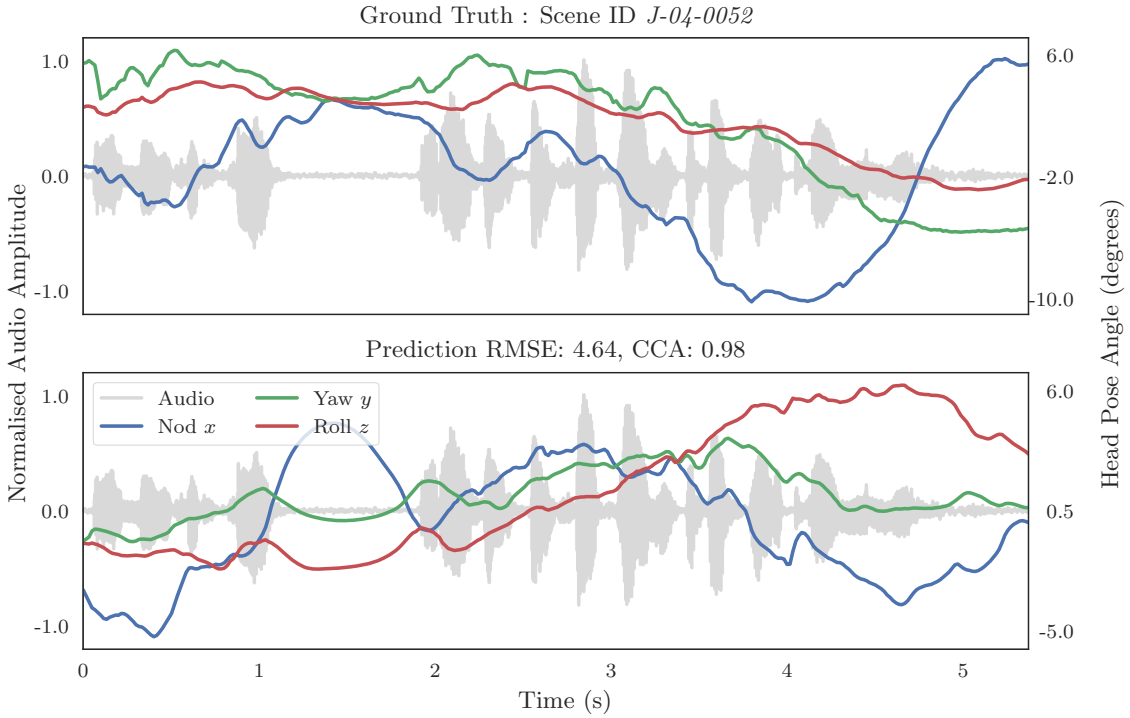


Figure 6.18: Results for Subject B CVAE model, example scene 2.

Table 6.10: Head pose results for Subject B for CVAE model. We show RMSE in degrees for rotation and millimetres for translation, and correlation using CCA for our test scenes. All test scenes have been permanently excluded from training, validation and model selection.

Scene ID	*J-01-0153*	*J-02-0089*	*J-03-0263*	*J-04-0052*	*J-05-0256*	*J-06-0276*
RMSE Rotation	3.48	2.66	5.22	4.64	4.20	4.46
RMSE Translation	1.04	2.13	2.60	3.41	4.97	3.41
CCA Rotation	0.92	0.73	0.76	0.98	0.92	0.96
CCA Translation	0.71	0.75	0.83	0.89	0.84	0.89

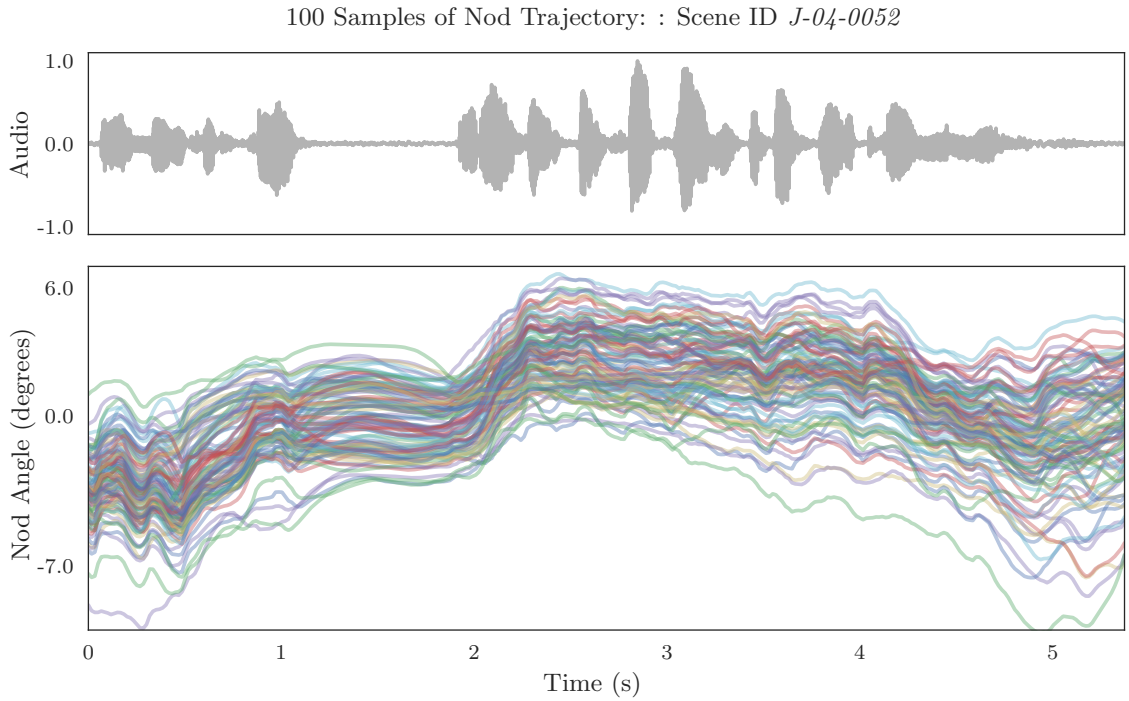


Figure 6.19: 100 random samples of nod trajectory. The CVAE model can predict a large variety of trajectories by sampling a normal distribution.

correlation across all examples, in line with the expectations that we established in Section 6.3.2. Note we also show results for the effective *translation* for head pose. To be clear, we show results for a single model that makes predictions of 6 DoF head pose from audio speech feature input.

A significant advantage to the CVAE model is to draw samples from a normal distribution and make a large number of plausible predictions from a single utterance. An example use case might be a single voice actor driving an animation of a crowd speaking in unison. All the agents are animated plausibly, but are all moving differently. Figure 6.19 shows the same utterance we illustrated earlier (Figure 6.18). To focus attention on the diversity, we plot only the nod trajectory, and show 100 randomly drawn samples to give 100 variations of that trajectory. This qualitative result demonstrates the character of the pose is maintained, despite the variety. Note the action in the region of the pause in the audio.

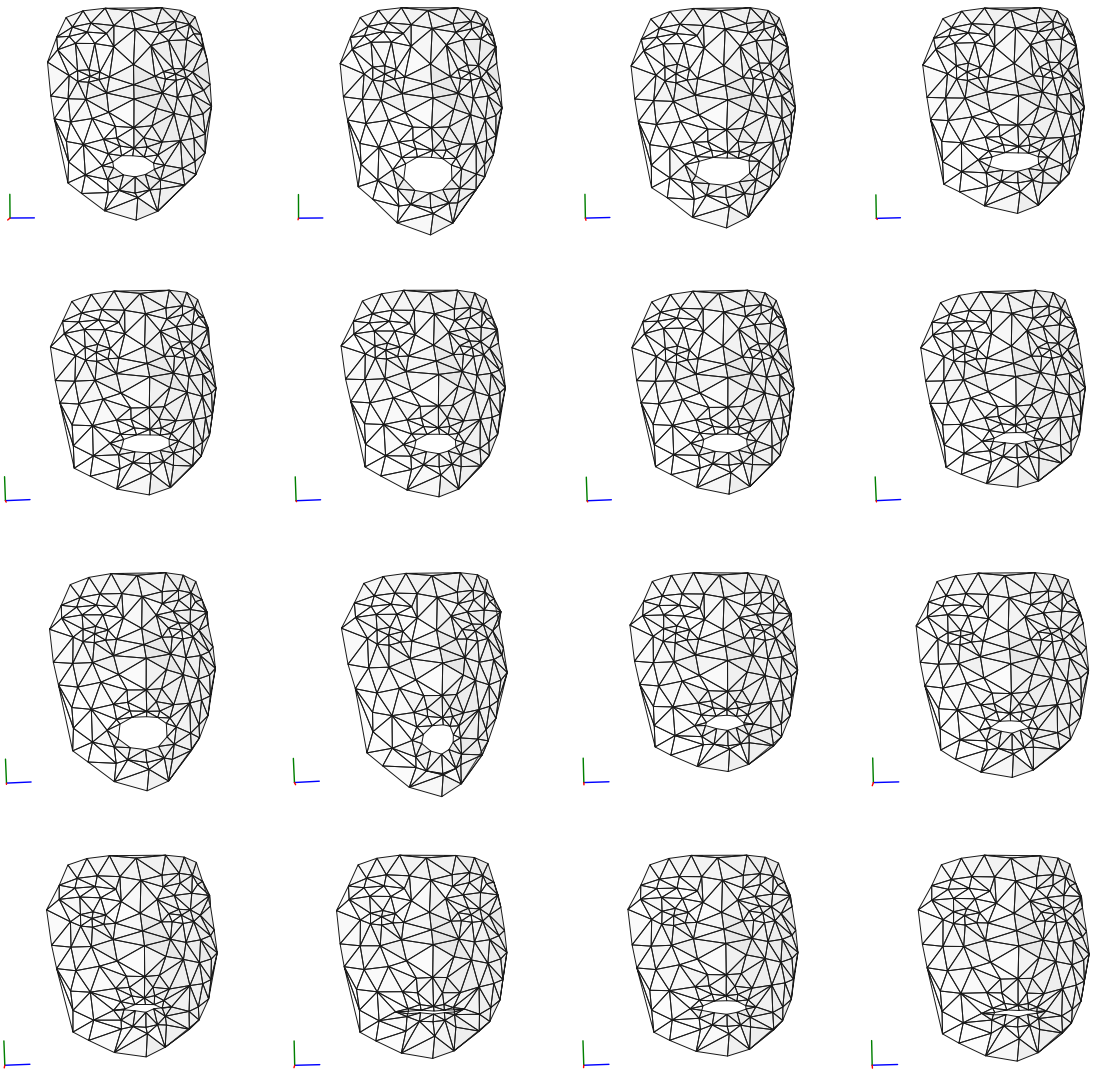


Figure 6.20: Here we show a number of head pose variations, extracted from an animation of Subject A, making the utterance "I can't breathe because you smell like garbage juice or rotten meat or something!". We display a frame at 20 sample intervals. Animations of the sparse point distribution model are the data for subjective testing in Section 6.7, and in the following chapters.

6.7 Subjective Testing

To further support our results, we conducted a subjective user study by predicting samples from our CVAE model, for the held out test data.

From our model, we predict 6 DoF, which we show by exchanging the values of the ground truth example with those of the predicted example. We then render animations of the complete example, for both the ground truth and the predicted values. To be clear, the real examples are simply renders of our corpus data, and the predictions are the same examples, with the 6 DoF head pose replaced with the model output. No other processing is applied.

Participants were shown the predictions, and were asked to distinguish between the prediction and the ground truth counterpart in a forced choice one stimulus discrimination test. We argue all forms of animation have some amount of perceptual noise. An example of (near) zero noise is a face to face meeting with another person, although even here, an actor could establish some degree of perceptual dissonance by adopting a particular behaviour.

We used Signal Detection Theory (SDT) [Macmillan and Creelman, 2004] to calculate the sensitivity index $d\prime$, the distance between the mean of the stimulus and the mean of the noise in dimensionless units of standard deviation. Here, the noise is the ground truth 'Real' example and the stimulus, or signal, is the predicted example which we dub 'Fake'. We regard this as a suitable test as it is not effected by bias, eg. a user selecting the same answer repeatedly. As an aside, we use the terms real and fake in deference to GAN terminology [Goodfellow et al., 2014], but of course here, our discriminator is a human viewer.

We show the raw results of the data collection in Figure 6.21, with the x axis showing the observed probability of all the answers. Table 6.11 shows the Hit and Miss scores for the collected data. A 'Hit' represents an answer of 'Fake' when the example is a prediction.

Table 6.11: Results of our user study for speaker head pose predictions from this chapter. A 'Hit' represents an answer of 'Fake' when the example is a prediction. A 'Correct Reject' is to answer 'Real' for a ground truth example. When a user responds 'Fake' to a real example, we report 'False Alarm', and finally, a 'Miss' is a 'Real' response to a prediction.

	Probability
Correct Reject	0.65
False Alarm	0.35
Hit	0.44
Miss	0.56

Table 6.12: We calculate the sensitivity index $d\prime$ to gain insight to viewer acceptance. In dimensionless units of standard deviation of the distribution of our examples, $d\prime$ gives a measure of how strongly viewers discriminated between real and predicted examples. Small values, less than one, indicate plausible animation predictions. We also show Z score for Hit and False Alarm, where $Z(p), p \in [0, 1]$, is the inverse of the cumulative distribution function of the normal distribution.

	$d\prime$	zH	zFA
Score	**0.241**	-0.14	-0.38

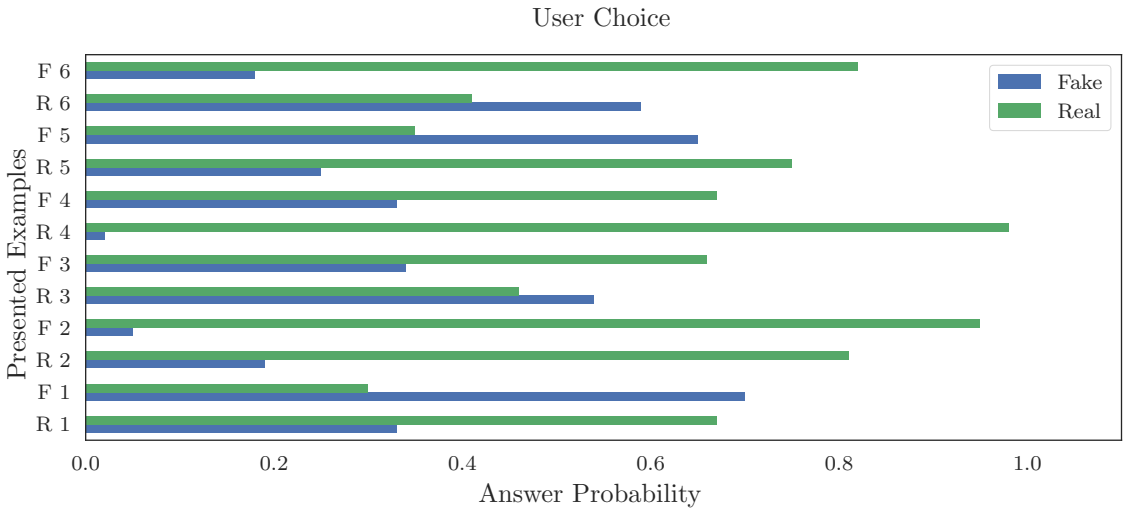


Figure 6.21: The data collected from a user study. We show the raw data as the probability of response to the question 'Real or Fake', when the user was shown an example from the corpus, or, a counterpart prediction respectively.

A 'Correct Reject' is to answer 'Real' for a ground truth example. When a user responds 'Fake' to a real example, we report 'False Alarm', and finally, a 'Miss' is a 'Real' response to a prediction. We use the equal variance model [Wickens, 2002] and the distance between the two distributions is calculated as the difference of the Z scores, $Z(Hit) - Z(FalseAlarm)$. Finally, we report the Z scores in Table 6.12 to calculate the *Sensitivity Index*, $d\prime$, which gives us a measure in standard deviations, of the distance between the noise and signal distributions.

6.8 Comparison with Prior Work

Table 6.13: Here we show recent results from related work compared with our own. The common quantitative measure is CCA for this domain. We project to a single base and report Pearson's r for the best values of our own and other work. Values are expected in the interval $[-1, 1]$. No correlation has a value of 0, -1.0 is negative correlation, with 1.0 as maximum correlation.

Author	CCA
Ding et al. [2014]	0.56
Ding et al. [2015]	0.71
Haag and Shimodaira [2016]	0.39
Our method	**0.96**

In this section we compare our own results with the best results of other recent work. Previous work may have different goals to our own work. We are not, for example, interested in categories of emotion [Busso et al., 2007]. We are interested in *plausible* animation of head pose for speech animation. We also want to predict real values of head pose without requiring post processing. We therefore compare our work with previous authors with similar goals, that publish evaluations with a common metric, Canonical Correlation Analysis (CCA). Table 6.13 shows the best results of similar prior art compared with our own.

6.9 Discussion

A number of previous authors evaluate results with a correlation measure or some other point-wise comparison, and of course we must also show these measurements. Taken in isolation, these methods of comparison can be unreliable for the following reasons: We do not expect a prediction to closely match the ground truth, rather, each should be one example of many possible but appropriate trajectories. Further, the *character* of the motion is not measured easily; one could measure quite high correlation without necessarily having appropriate motion. To illustrate this point we show a counterpart example in Figure 6.22, showing acceptable RMSE and CCA, but unacceptable motion characteristics. Incidentally, this is an example of a model that is *overfitting*.

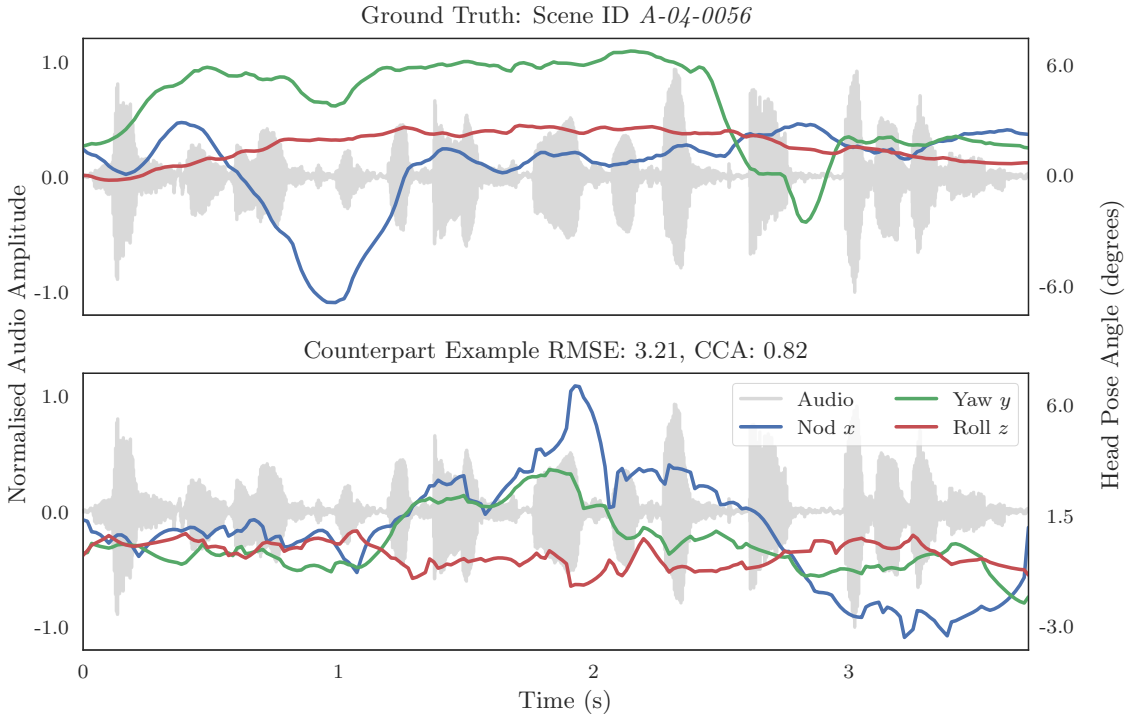


Figure 6.22: Qualitative counterpart example. To emphasise the importance of including a qualitative means of evaluation, we show an example that has acceptable quantitative measurement. Viewing the plot clearly reveals the head pose trajectories do not share the same motion characteristics of the ground truth.

The question of what represents appropriate or suitable head motion during speech is unclear. Subjectively, we have observed certain key events support viewer acceptance (Figure 6.6), so it appears some amount of correspondence with the audio is important. Qualitative assessment, by plotting trajectories, is useful to identify the character of the motion. Yet if we rely only on moment statistics, that can identify noisy motion, we may still have unacceptable motion; consider the reversed motion in Figure 6.6, that has the same moment statistics as real motion, but not the correlation. Finally, as well as all of these assessments, we can show our predictions to human viewers, and here we gain more insight to what is accepted and plausible head motion (Section 6.7).

We may not yet have a single measure of how 'good' a head pose prediction is, we do know however, that it is important to have correct motion [Munhall et al., 2004], and also that we can identify when it's not correct [Mori, 1970]. Developing a standard measurement of correct head motion, or indeed more broadly gesture, is an open and difficult problem, and we are actively pursuing this goal.

Our most interesting results come from the CVAE model, that solves the one to many mapping problem. We can predict a number of plausible motion trajectories by choosing new values for z, but with the same audio features. This model also allows us to train the model on multiple speakers without the predictions heading toward the mean. The model learns to distribute the identity of the speaker normally, maximising our training data. With a corpus of many speakers, we could place a categorical label as a further conditioning feature and explore the manifold of identity.

For subjective testing, we replace the ground truth rigid motion, with the predicted rigid motion. We do this to unify the comparison of all our methods. One downside here, is that we give a visual clue embedded in the facial expression as to whether or not the head pose is 'Real' or 'Fake'.

In this chapter we have developed deep learning techniques with RNNs, specifically BLSTM. By carefully preparing a corpus of expressive speech examples and recognising the appro-

priate topology of model, we have been able to exceed the performance of previous authors. The most prevalent measure of performance in the literature is CCA, and we agree this is an important measure. The best figure previously being 0.711 (Table 6.3), we comfortable exceed this figure, our best figures typically greater than 0.95. In addition to predicting head pose from audio features, we also make predictions from aligned phone features. We revisit this feature in Chapter 8. It should be made clear that we make predictions of the real values of head pose directly; we do not use any form of post processing, filtering or smoothing.

We go further than other recent authors in the scope of our ambitions. We show models that predict the six DoF of head pose; the effective translations as well as rotations. Finally, to underline the view that there are many appropriate head motions during speech, we show 100 examples of the nod trajectory in Figure 6.19.

7 Listener Head Pose

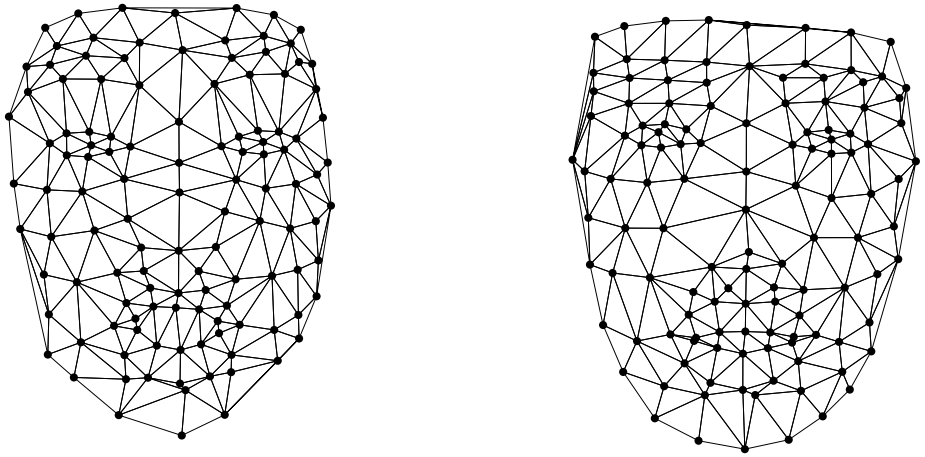


Figure 7.1: In this chapter we model the head pose of the listener in dyadic conversation. Listener head pose can express attention, agreement, empathy and provide feedback to the speaker.

In this chapter, we consider the rigid motion of the head of the *listener* in dyadic conversation. Specifically, we explore the notion that the listener responds to the speakers *voice* in some corresponding way. Listener interaction, or "backchannels" [Yngve, 1970], provide both acoustic and visual signals that inform turn taking, express attention, agreement, empathy and give feedback to the speaker. Some obvious use cases we have already discussed in previous chapters for speaker animation, hold also for the listener, such as automatic animation for the entertainment industry, or on-line avatars. Another, very important domain, are Embodied Conversational Agents (ECAs). ECAs can be a compelling model for

Human-Computer Interactions (HCIs), and, as well as providing a natural interface are useful, for example, in CBT.

7.1 Motivation

Ward and Tsukahara [2000] establish that prosodic aspects of the speakers voice elicit listener back channel response, so we want to examine our corpus for indications to support their claims. A motivating example is shown in Figure 7.2. The plot shows the head pose trajectory of Subject A, listening to the voice of Subject B, making the utterance: "It's laughable to me that you assume I have any interest in touching you." The pose does appear to show a degree of correspondence with the audio, particularly in the first 2s. To confirm this observation, we measure correlation with our standardised audio feature, LogfBank, using CCA. We report Pearson's r of 0.87 for this example projected to a single base, a figure that indicates significant correlation. Recall that many researchers studying speaker head pose regard this degree of correlation to be a motivation for predicting head motion from audio (Chapter 2).

7.2 Corpus

We thoroughly discuss the collection of data in Chapter 3, so here, highlight the relevance for making predictions of the listener's head pose. Our data was recorded as a set of dyadic *conversational* vignettes. Specifically, the two actors were engaged in a verbal exchange where they could both see each other and hear each other.

Two further details of our collection process provide the data for the work in this chapter. First, we ensured the tracker was fitted to the Subjects' faces at *all* times. Secondly, *all* the cameras were synchronised for the recording of both Subjects. These facts allowed us to take the annotation of the audio for the speaker, and use the timing of the utterance end

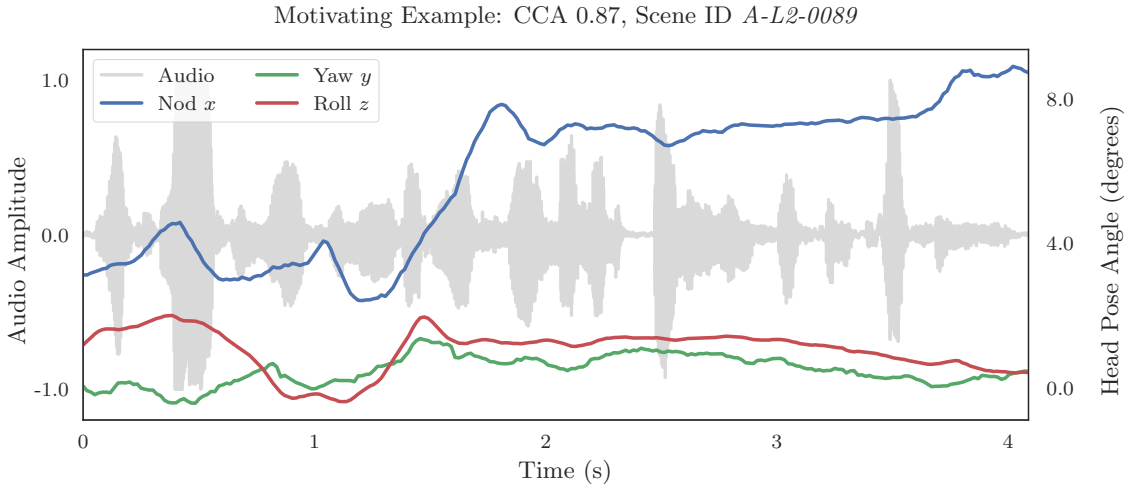


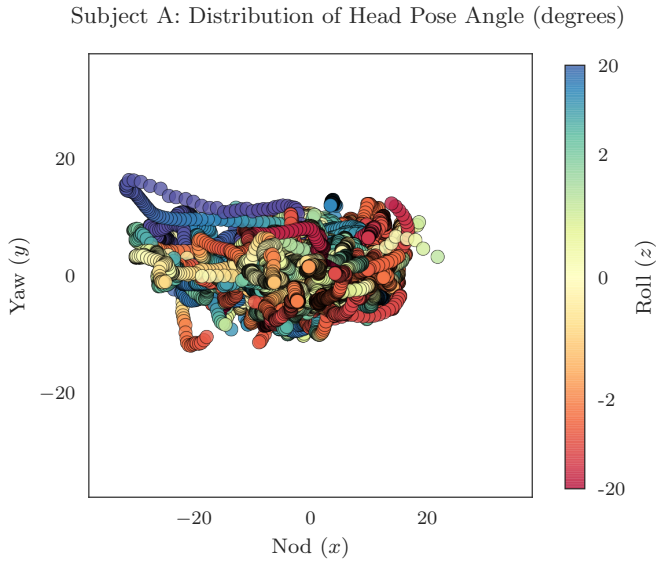
Figure 7.2: A motivating example for learning listener head pose from audio. The head pose of the listener is modulated by the speaker's voice in a similar way to the speaker head pose. Correspondence is lower, but measuring correlation using CCA shows potential for learning.

Table 7.1: Distribution of Subject A head pose angle while listening. We show maximum, minimum, mean and standard deviation for each of Nod, Yaw and Roll. The angle unit is degrees. For Subject A we observe much greater roll than B, that we can also see in the scatter plot shown in Figure 7.3a

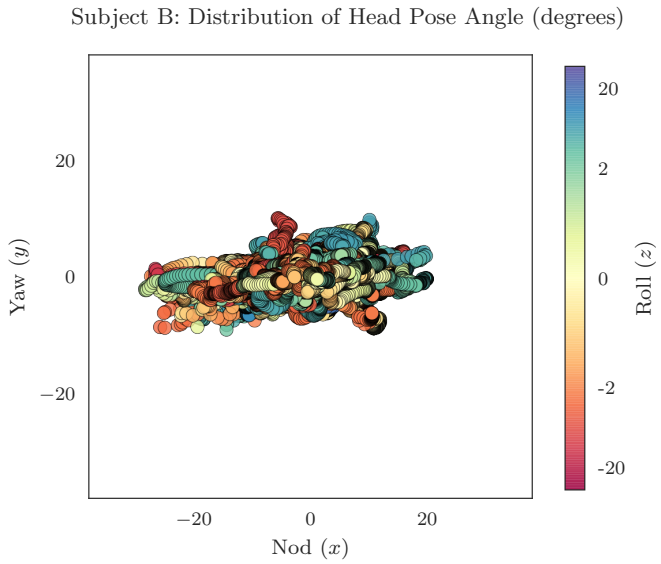
Listener A	Nod (x)	Yaw (y)	Roll (z)
Min	−31.59	−11.86	−32.07
Max	21.65	16.29	43.25
Mean	0.09	0.01	0.01
STD	5.09	2.24	3.66

points, to identify the tracking data for the *listener*, in the same period. We now have the motion data of the listener while the speaker makes an utterance.

It is useful to get an overview of the motion of the subject while listening. We can visualise the distribution of the head pose angle for our subjects while they are listening in Figures 7.3a and 7.3b. These figures make an interesting comparison with our same subjects when speaking (Figures 6.3a and 6.3b). The first observation is that Subject A has a larger range of motion, while listening, than Subject B. This is consistent with our observations for our



(a)



(b)

Figure 7.3: In the same way as the speaker distribution, the nod angle is constrained in the positive angle (head down), by normal anatomical limits. Attention is drawn to the greater range of pose for Subject A, which can be attributed to more expressive behaviour of Subject A generally. The Roll (z) axis is described by the colour bar, that has linear scale in $[-2, 2]$ and log scale outside of that interval to avoid too many light colours.

Table 7.2: Distribution of Subject B head pose angle while listening. We show maximum, minimum, mean and standard deviation for each of Nod, Yaw and Roll. The angle unit is degrees.

Listener B	Nod (x)	Yaw (y)	Roll (z)
Min	−28.29	−8.99	−16.98
Max	19.92	9.96	18.14
Mean	−0.04	−0.01	−0.00
STD	5.13	1.48	2.60

subjects while speaking. It is also consistent with our view that Subject A has a more animated demeanour in person, and certainly represents part of the subject's identity. Not only are the pose angle values more widely distributed for Subject A, much of the difference is in the head roll; an action that could indicate greater empathy or attention.

7.3 BLSTM model

We now introduce the deep BLSTM to the problem of modelling the listener's head pose [Greenwood et al., 2017b]. We establish a model topology similar to our approach when modelling the speaker's head pose (in Chapter 6) and we describe the motivation for this choice and the detail of implementation in chapter 5. Specifically for the task of modelling the listener's head pose, we build a network with 3 bidirectional layers each of 32 hidden units. The count is for each direction, so is doubled. We arrived at this topology by conducting experiments with both larger and smaller networks and choosing the most effective based on quantitative and qualitative evaluation of the predictions. It was not possible to rely on the minimum loss of the model objective function. Figure 7.4 illustrates this topology. Our model accepts LogfBank audio features as input, and the real values of rotation angles of the listener's head. We augment our data as described in Section 5.7 and set $n = 129$ time steps to capture long term events.

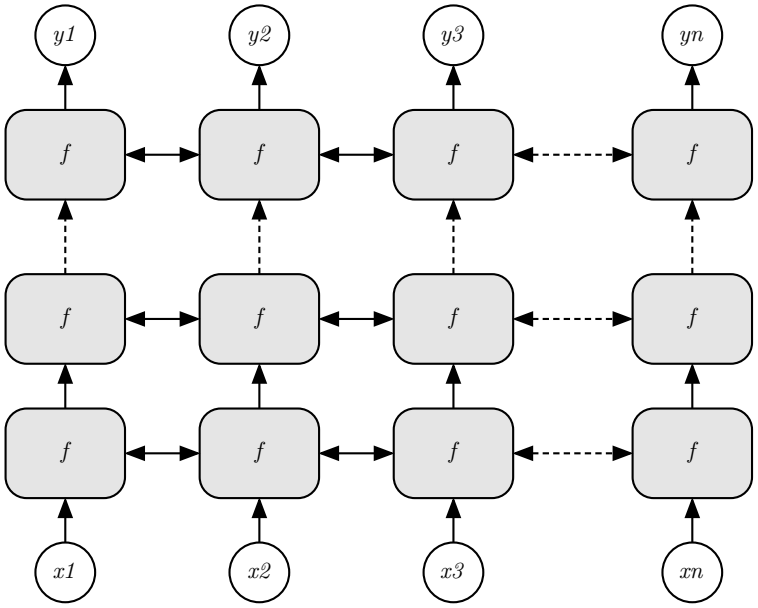


Figure 7.4: Modelling listener head pose with deep BLSTM. We show a model trained with X speech features to predict Y head pose Euler angles of the *listener*. At each time step we emit the state of the network, in a many to many model topology. For our listener experiments the model has 3 layers each of 32 hidden units for each direction, and we train the model on only one speaker for the corresponding listener, i.e., this model is subject dependent.

7.3.1 Training

We trained the networks on our data, split 90% for training, 10% for validation and we extracted a small random selection of utterances distributed uniformly across the corpus that were never used for training, nor validation and model selection. We scale the features such that each feature has zero mean and unit variance. We also scale the target rotation trajectories in the same way, and de-scale them after prediction with the inverse of the scaling function. Our objective function is MSE. Our optimising function is *RMSprop* [Tieleman and Hinton, 2012].

Previously, for speaker head pose prediction, we would set a patience for stopping, then reload the best weights, decrement the learning rate and restart training. We found this

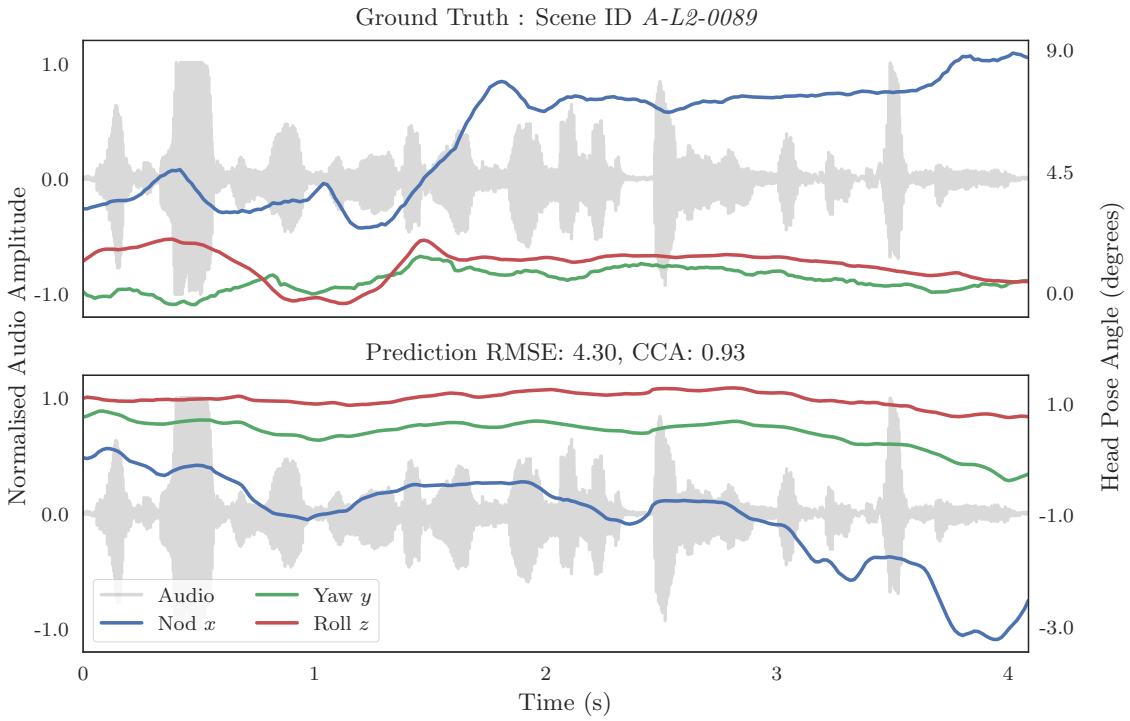


Figure 7.5: Listener head pose trajectories for the example we used previously to motivate the experiments of Subject A listening to the speech of Subject B. We can see that there is a correspondence between the audio and the head pose in both the ground truth and the prediction, the trajectories of the prediction look almost mirror symmetrical to the ground truth.

unnecessary for listener head pose, in fact qualitatively, this was detrimental to the result. As such, we set a learning rate of 10^{-3}. Training continues until no further improvement on the validation set, with a patience of 10 epochs. It was useful to view plots of samples predicted from earlier epochs of training rather than just judging by the smallest MSE. We suggest this is a useful example of *early stopping* [Prechelt, 1998].

7.3.2 Results

We first view the utterance example from Section 7.1, and we show our model prediction in Figure 7.5. By now we are used to not seeing something that looks just like the ground truth, but looking more closely we see the plot of the trajectories has many of the broad

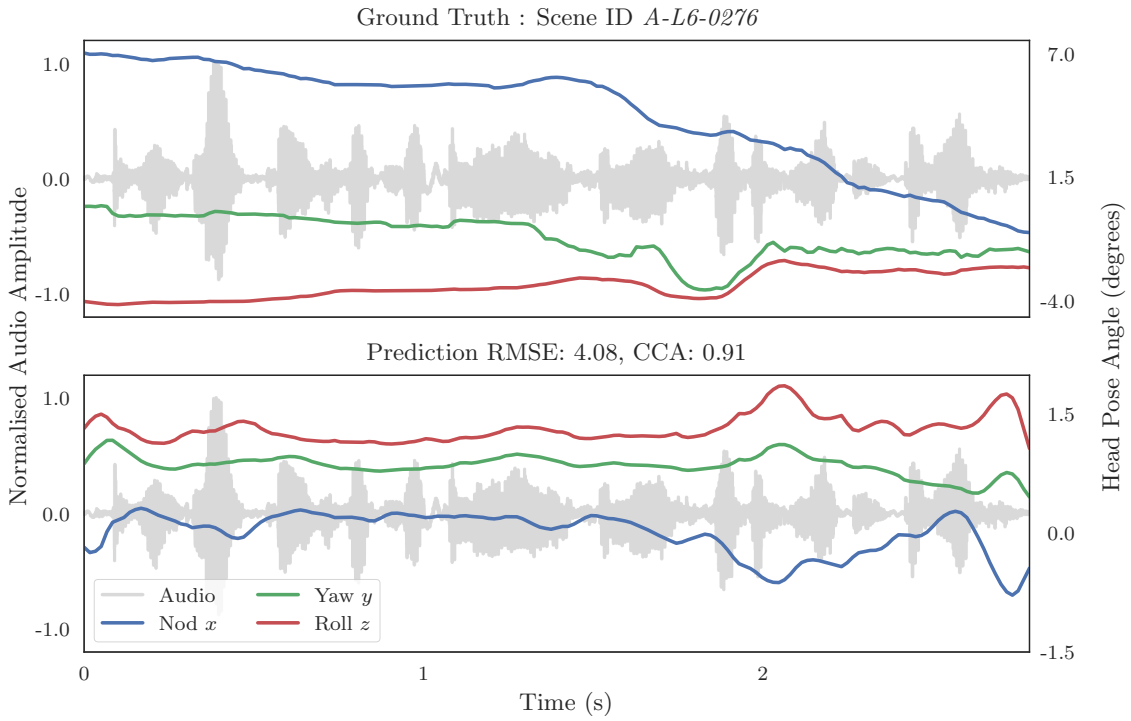


Figure 7.6: Here we show a second example for Subject A listening to Subject B. The latter half of the utterance shows increasingly frequent modulation in both the ground truth and the prediction.

aspects of the real utterance. In fact the trajectories look like the mirror symmetry about the time axis of the plot. The CCA figure of 0.93 shows considerable correlation. We show another example of Subject A is listening while Subject B speaking in Figure 7.6. Here, we see in the ground truth, that the second half of the utterance provokes the most movement in the listener. The same can be said in the prediction, where there is more activity in the second half of the utterance. Another interesting observation is the activity is a subtle rhythmic nodding, with the head rolled to the side; something very characteristic of Subject A's motion (Figure 7.3a). We accumulate results for our testing examples for Subject A listening in Table 7.3. Here we can see quantitatively, that we achieve good correlation with the ground truth, comparable with the figures we achieve for speaker head pose prediction. The model we have demonstrated so far is trained on the speech of

Table 7.3: Listener head pose results for Subject A from audio features. Subject A is listening to Subject B speaking.

Scene ID	*A-L1-0151*	*A-L2-0089*	*A-L3-0264*	*A-L4-0053*	*A-L5-0255*	*A-L6-0276*
RMSE	2.66	4.30	3.68	6.21	2.71	4.08
CCA	0.80	0.93	0.74	0.85	0.93	0.91

Subject B to predict the head pose of Subject A. We now train a separate model to make the counterpart prediction, Subject B listening to subject A. To be clear, this model for listener B has the same topology and is trained with the same parameters as the model for listener A. For subject B, we show an example prediction in Figure 7.7. We achieve good correlation and RMSE with the ground truth, and this example again captures the qualities of the listener's pose. Here we can see that Subject B is simply less dynamic than Subject A and this is supported by the data for all B's listener head pose angles in Figure 7.3b and Table 7.2. We collate the results for Subject B listening in Table 7.4. The values

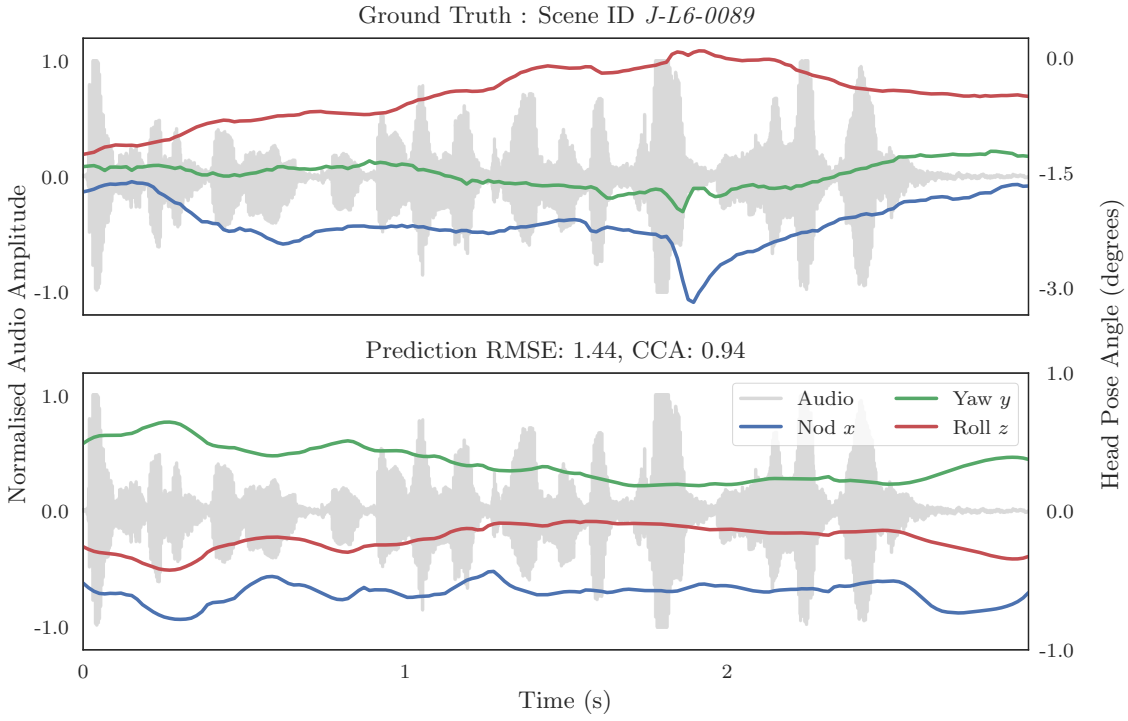


Figure 7.7: Listener head pose trajectories example for Subject B. Subject B is generally less dynamic than Subject A. Here we show a prediction for our test scene that retains the 'character' of the Subject's motion.

Table 7.4: Listener head pose results for Subject B from audio features. Subject B is listening to Subject A speaking.

Scene ID	*J-L1-0185*	*J-L2-0121*	*J-L3-0203*	*J-L4-0056*	*J-L5-0263*	*J-L6-0089*
RMSE	2.18	4.39	1.93	6.92	6.51	1.44
CCA	0.78	0.87	0.68	0.97	0.81	0.94

are less consistent for Subject B, we hypothesise that he is less responsive to the speech of his counterpart interlocutor, or perhaps not as involved in listening as Subject A.

The main limitation of our BLSTM model is the requirement to train a model for each subject. Particularly for our task of predicting the listener's pose, the pose gap between our subjects makes a single model trained on both subjects head toward the mean pose result, diluting the characteristics of each. We address this issue with our CVAE model [Greenwood et al., 2017b].

7.4 CVAE Model

The Conditional Variational Autoencoder (CVAE) model allows us to generate head pose trajectories drawn from a normal distribution, *conditioned* on the audio of the speaker. We formally describe the CVAE model in Section 5.5.3. Specifically for the task of learning listener head pose, our model has a tapering stack of BLSTM layers in the *encoder* of 128, 64, 32 hidden units for each direction. We use a latent vector size of 6, and a *decoder* with BLSTM layers of 128, 64, 32 hidden units for each direction. We use dropout of 0.5 to regularise the model during training. Figure 7.8 shows how we condition the model on speech features at every time step, which are LogfBank audio features. For this model, output is the rotation *and* translation values of the listener's head pose. In the same way as the previous model, we augment our data as described in Section 5.7 and set $n = 129$ time steps to capture long term events.

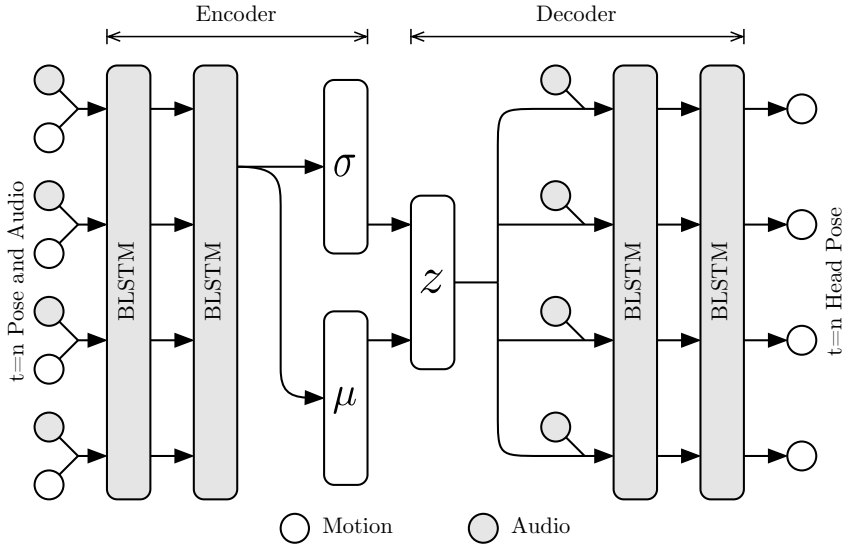


Figure 7.8: The topology of the CVAE model for our listener experiments is similar to the model for speaker experiments. Here the encoder has a tapering topology of 128, 64, 32 hidden units, and we use the same arrangement for the decoder. The model is conditioned on speech features at every time step. At test time we forward propagate through the decoder only.

7.4.1 Training

The CVAE model is trained by splitting our data 90% for training, 10% for validation. We have previously removed a random selection of utterances distributed uniformly across the corpus that were never used for training, nor validation and model selection. Our standard technique is to scale our input features and output target such that they have zero mean and unit variance for training. We de-scale the target rotation values after prediction, using the inverse of the scaling function, so they are restored to real world values. Our optimising function is *RMSprop* [Tieleman and Hinton, 2012] and we set a learning rate of 10^{-3}. The network is trained until no further improvement on the validation examples with a patience of 10 epochs. Just as with the BLSTM model earlier, it proved useful to make qualitative assessments by plotting the head pose trajectories in the region of the minimal loss epoch. To allow direct comparison with our earlier model, we retained $n = 129$ time steps for this model.

7.4.2 Results

To directly compare the two approaches, we first show the example used to motivate our learning approach, and tested with other listener model. Figure 7.9 shows Subject A, listening to Subject B making the utterance: "It's laughable to me that you assume I have any interest in touching you." Qualitatively, we can immediately see how the significant events (in the first half) of the nod trajectory of the ground truth is successfully predicted by the model. Recall, we train the model only on audio features of the speaker, to predict the head pose of the listener, so this represents a good example that this *is* indeed possible, and supports the work of Ward and Tsukahara [2000].

One of the advantages of the CVAE model, is we can train the model on both the subjects of our corpus. We can do this because the model learns the distribution of each subject's

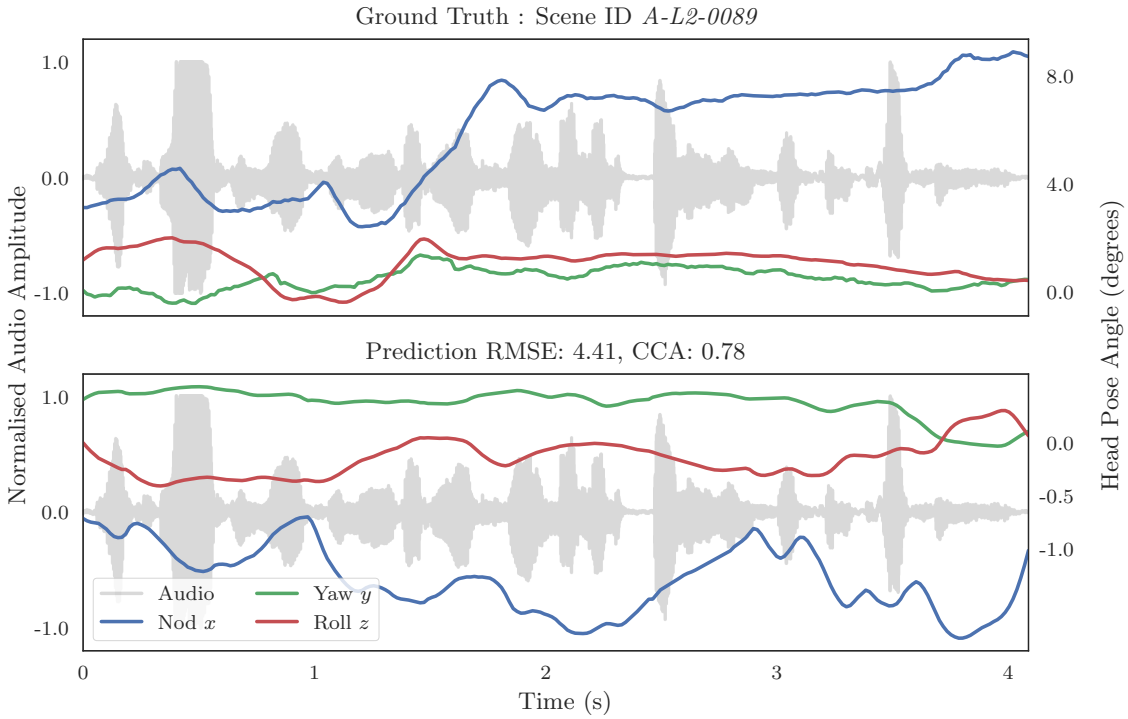


Figure 7.9: CVAE result for Subject A listening. We revisit our motivating example again with our generative model. Here we see especially how the nod trajectory of the listener moves to the speaker's voice. The peaks in the first 2s are very much alike in ground truth and prediction. Not only does this show the model is performing appropriately, but that the audio contains relevant information to predict the listener's head pose.

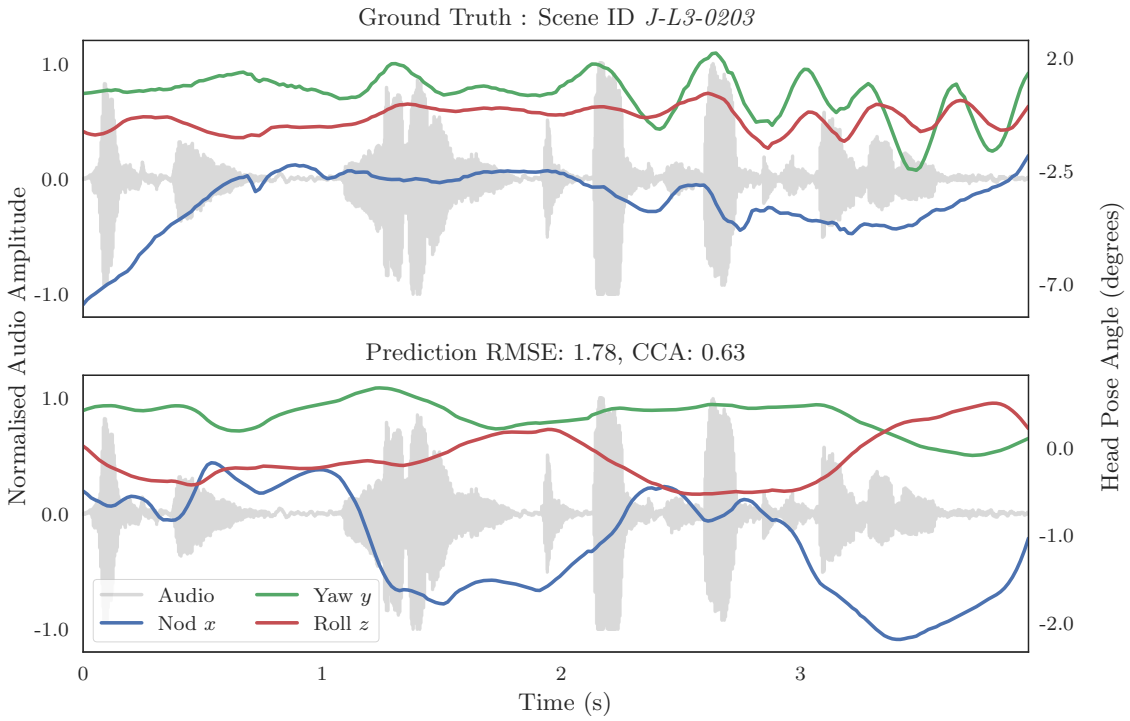


Figure 7.10: CVAE result for Subject B listening. This is a particularly illuminating example. Here we see how an event at 2.5s is captured by the model, yet the yaw (y) activity is not. The hypothesis is the *sentiment* expressed by this action is a result of the listeners global analysis of the utterance, i.e., he rejects the utterance.

pose such that the latent variable contains aspects of the subject's *identity* with respect to their pose.

In Figure 7.10 we show a test example from our other listener, Subject B, listening to Subject A making the utterance: "This is literally the best sausage stand on earth." The example here is quite interesting, in that it represents, at least partially, a counter example. The ground truth of the example shows the yaw oscillating throughout the utterance, but most pronounced in the latter half. This is not properly reflected in the models prediction. We hypothesise this is atypical of the subject's normal response. This rhythmic head shaking, representing some form of disagreement or lack of belief in the speaker's statement, does not appear with sufficient regularity in the collected data. If we consider the nod trajectory, we do see similar events in the ground truth and prediction.

Table 7.5: Listener head pose results for Subject A for CVAE model. We report RMSE angles in degrees for rotation, and millimetres for translation. Correlation using CCA is also reported. All test scenes have been permanently excluded from training, validation and model selection.

Scene ID	*A-L1-0151*	*A-L2-0089*	*A-L3-0264*	*A-L4-0053*	*A-L5-0255*	*A-L6-0276*
RMSE Rotation	4.08	4.41	4.34	5.89	2.79	4.15
RMSE Translation	3.28	19.55	12.37	9.29	22.27	33.77
CCA Rotation	0.80	0.78	0.90	0.86	0.85	0.92
CCA Translation	0.86	0.83	0.87	0.78	0.77	0.71

Table 7.6: Listener head pose results for Subject B for CVAE model. We report RMSE angles in degrees for rotation, and millimetres for translation. Correlation is measured using CCA for our test scenes. All test scenes have been permanently excluded from training, validation and model selection.

Scene ID	*J-L1-0185*	*J-L2-0121*	*J-L3-0203*	*J-L4-0056*	*J-L5-0263*	*J-L6-0089*
RMSE Rotation	2.34	4.06	1.78	6.53	6.94	1.31
RMSE Translation	14.99	33.39	28.58	25.67	36.47	21.70
CCA Rotation	0.88	0.88	0.63	0.94	0.95	0.84
CCA Translation	0.87	0.93	0.92	0.86	0.75	0.74

We collate the results for both our listener's in Tables 7.5 and 7.6. Immediately we notice that the error for *translation* is large. The unit is millimetres, so we are not viewing model collapse, but when compared to predictions of the speaker's head pose (Tables 6.9, 6.10), for some examples the error is almost an order of magnitude. Clearly the translation, which is an effective translation from all the joint positions below the skull, does not have sufficient correspondence with the listener's response to the speaker's voice.

Another advantage for the CVAE model is the ability to draw from a normal distribution and generate alternative predictions for the same audio input. This certainly has interesting possibilities for some of the use cases we draw attention to at the head of this chapter. For example, we could bias output to more or less dynamic regions of the distribution for particular users. We show an example in Figure 7.11 of multiple possible outcomes by drawing 100 random samples from a normal distribution. We return to our original example utterance and show 100 possible nod trajectories for Subject A listening to Subject B.

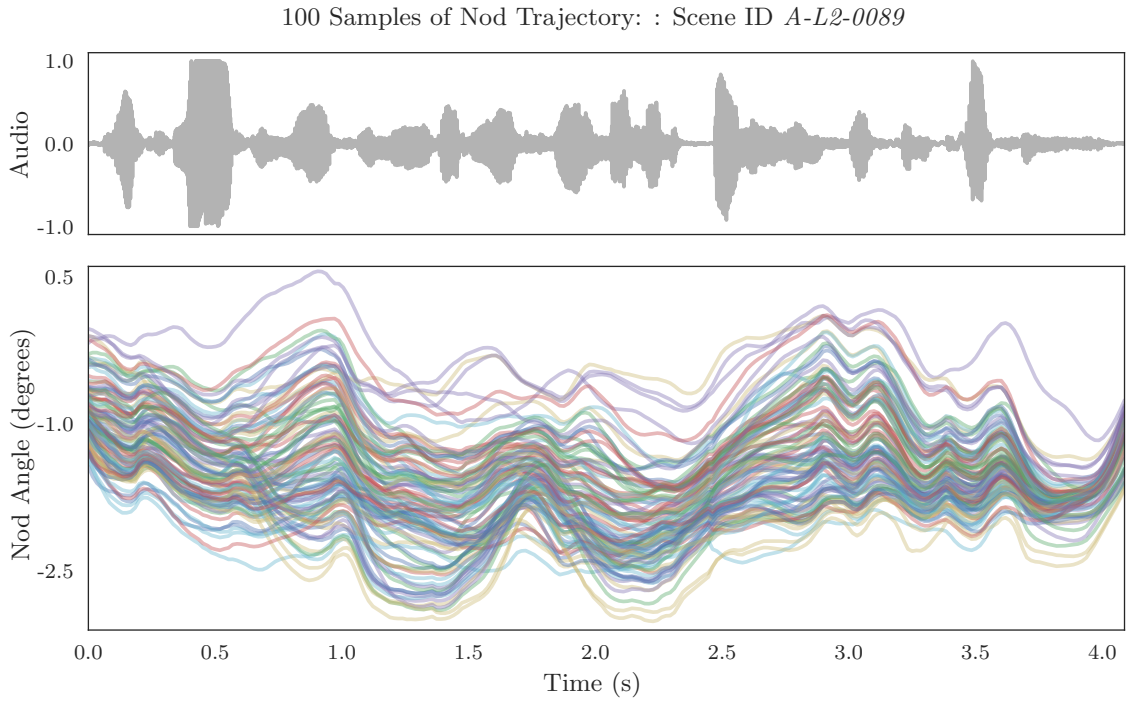


Figure 7.11: 100 variations on listening. A benefit of the generative model is the ability to sample from a normal distribution to predict many variations of plausible motion. Here we show 100 such samples of the nod trajectory of our original example.

7.5 Discussion

Modelling the pose of the listener in dyadic conversation is certainly an interesting problem and a number of questions arise from our approach in this chapter. In line with other authors, we can confirm that the speakers voice has an influence on the listener's pose. The first question is how *much* correspondence is relevant for learning? At the start of this chapter we found an example from our corpus that shows high correlation between the speakers voice and the listener's pose. As we trained models motivated by this example, we found that we could make reasonable predictions, however we also found counter examples that showed motion that did not model well.

We might conclude that only some of the information a listener model needs can be learnt from the audio. Another view, and one presented as more likely, is our corpus does not adequately represent attentive listening, especially when Subject B is the listener.

Modelling the valence and arousal of the speaker, is quite possible from their speech, perhaps by sentiment analysis, or based on pitch and cadence. To model the listener's emotional state appears to be too far removed from the speakers voice and more strongly linked to the listener's character.

To build a better model of the listener, aspects of the listener's response need to be more clearly defined. A corpus where the listener had clear agreement, or disagreement would allow categorical labelling of the response in the positive or negative.

Our modelling approach, using the CVAE, is entirely appropriate for this additional modality. At present we condition our model on the speaker's voice. We can extend the idea to condition on the listener's valence as well. In fact, for a large corpus of many participants, we could condition on *identity* too. We then have a model that, in response to the speaker's voice, could move in positive or negative manner. Adding the identity label, allows us to explore a manifold of learnt identity in just the same way as we can predict multiple possible trajectories at present.

It does appear that our listener model is constrained by our corpus, rather than our method, so there is a clear path to resolve those limits. The view that using only the speaker's voice is a limitation for this task is not one we hold. In the next chapter, we show how to model the speaker's facial expression *and* rigid head pose from just the speaker's voice. Using ideas from the next chapter would allow our listener model to hear and *see* the speaker; an exciting prospect.

8 Facial Expression

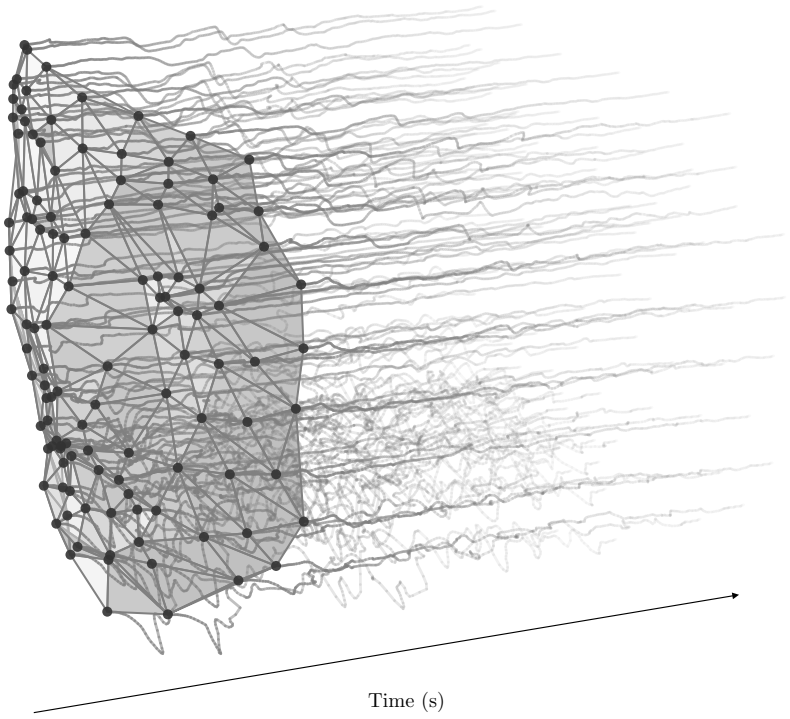


Figure 8.1: The 3D deformations over time. We show the 3D shape model for Subject A, with the perturbation of the landmarks as the shape deforms during an utterance. The shape model has 120 3D points, which for this utterance of 300 samples, is over 100,000 degrees of freedom.

In this chapter we will explore the activity of the facial features during speech. We will consider three questions:

- Can we predict the facial expressions, including lip shapes from sound?
- Can we predict facial expression, including lip shapes from text, specifically a phone alignment?
- Can we predict head pose from facial expressions?

By answering these questions here, we raise further possibilities for rigid head pose prediction, that we cover later in the Chapter, and in Greenwood et al. [2018].

8.1 Related Work

We enter a very broad area of research in this chapter, but will highlight a brief overview of some pertinent previous work. Much more detail can be found in Chapter 2. Facial animation, predicted from speech is not the primary goal of this work. We discover, however, by making predictions of facial expression, including lip syncing, that a new solution for making predictions of head pose is available. This allows us to make comparisons between our own method and some recent examples of facial prediction [Suwajanakorn et al., 2017; Karras et al., 2017; Taylor et al., 2017].

8.2 Dimensionality Reduction

The activity of the face is complex during speech. With our parametrised shape model, described in detail in Section 3.7.2, we have 360 degrees of freedom at every motion sample, with every point of the model under independent control. Figure 8.1 gives one view on this complexity by showing our shape landmarks moving through time for one example utterance. We show a triangulation of the shape model for additional clarity in the illustration.

One can observe each point moves to some degree at each time step, but also, we can see that there is considerable correlation in the movement.

Using PCA, we can reduce the number of dimensions somewhat. Forming a covariance matrix by flattening the shape model at every time step over the entire data set for a single subject, 98% of the variance is captured by 8 principal components. We will refer to this reduction as our PCA shape model, not to be confused with any neural network models we develop using principal components as input. By plotting the reconstruction from those principal components against the original shape, we can visualise the loss. Figures 8.2 and 8.3 illustrate the reconstruction loss for Subject A and Subject B respectively. The loss for each subject is acceptably small: 0.52 mm and 0.71 mm respectively. To place these values in some perspective, the marked landmarks on our actors' faces were approximately 2 mm in diameter. In practice it would be difficult to place a marker within a tolerance of $\pm$ 0.5 mm when annotating the training data (described in Chapter 3). One aspect of dimension reduction using PCA, is that the resulting Eigenvectors do not represent what might be termed 'animation controls'. The predominant work flow in industrial face animation environments is the blend shape model, were a technical artist creates a number of facial poses that are combined in a linear fashion by an animator, using 'animation controls'. It is possible to learn a mapping from PCA component values to blend shape controls as demonstrated by Taylor et al. [2017] using rig re-targeting, although this does still require the services of a technical artist.

Regardless of this reduction in dimensionality, it is still somewhat difficult to visualise the data. For Subject A making the utterance:

> "I can't breathe... because you smell like garbage juice or rotten meat or something!"

Figure 8.4 shows the PCA components as they vary over time, with the time domain audio of the utterance shown on the same plot. Here we can see how significant activity in the time domain audio corresponds strongly to activity in the values of our PCA model. We note

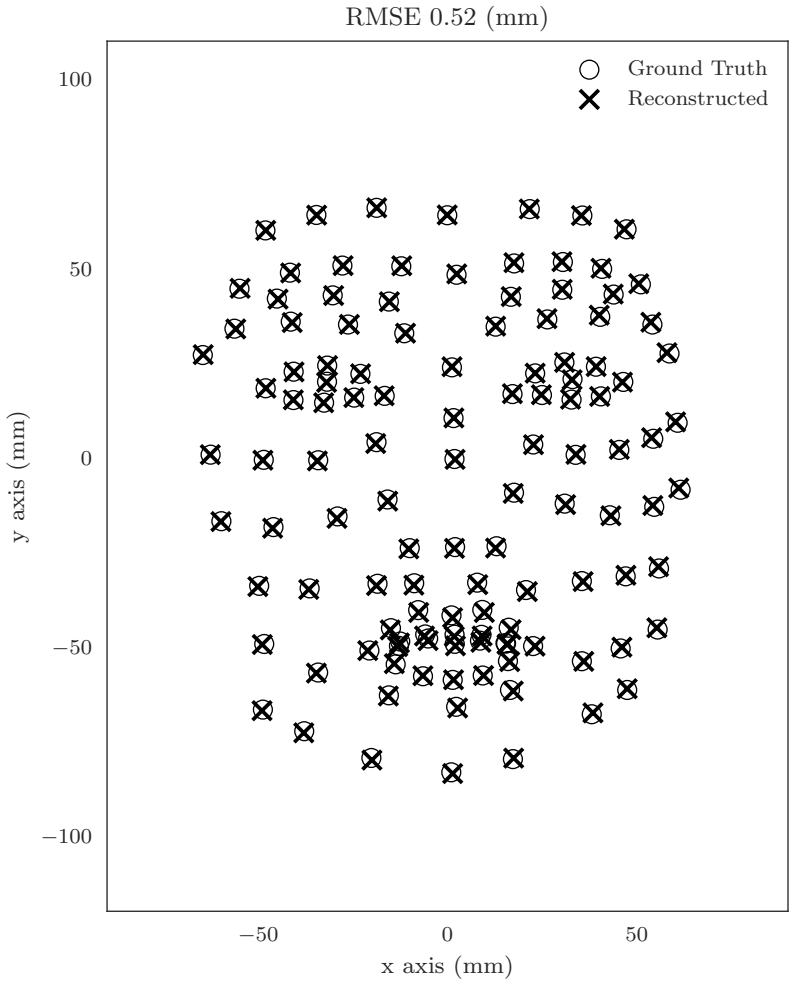


Figure 8.2: Reconstructing the deformation shape from 8 principal components for Subject A. The RMS reconstruction loss, ≈ 0.52 mm is acceptably low. The figure is an orthographic projection to 2D, however the loss is for the 3D data.

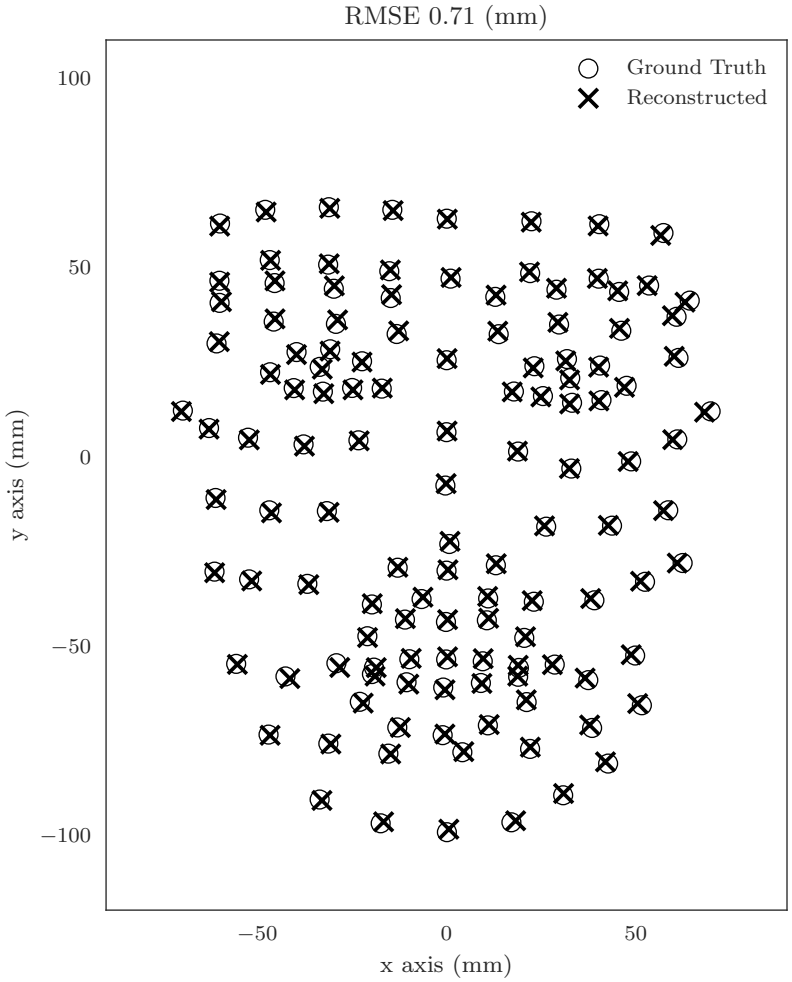


Figure 8.3: Reconstructing the deformation shape from 8 principal components for Subject B. The RMS reconstruction loss, ≈ 0.71 mm , although not as good as Subject A, is still acceptably low. We report loss for the 3D data, although the figure is a 2D projection.

that some components, pc 2 for example, rise in value when amplitude increases, whereas others, *e.g.* pc 5, behave in the opposite manner. It is important to make the remark that these changing values do not directly represent a particular facial activity such as brow raising, or mouth motion. The PCA model merges all such activities as determined by the variation in the data.

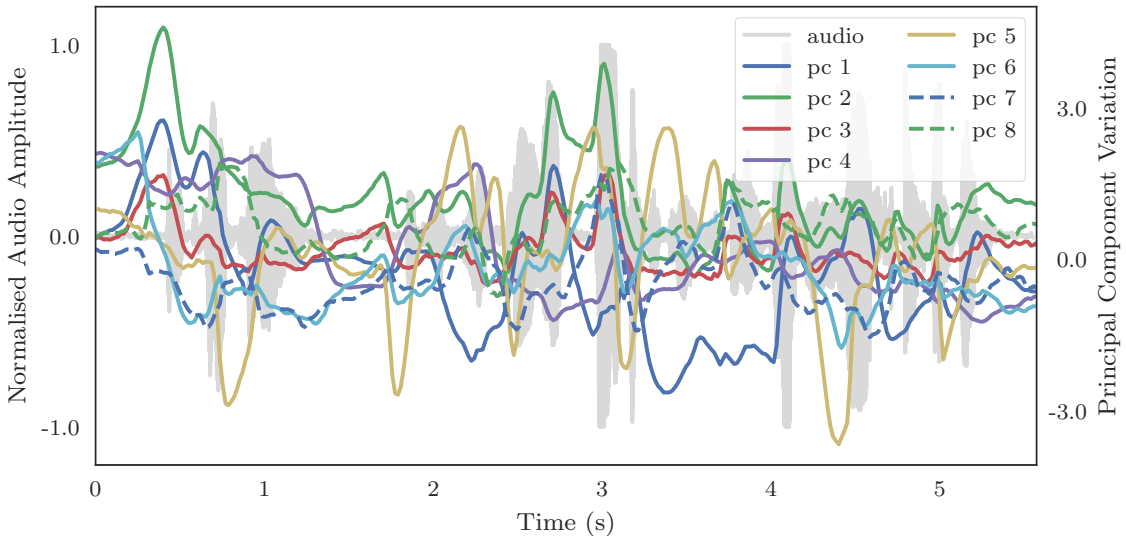


Figure 8.4: Eight principal components vary over time during speech. The components are shown with the audio to give an impression of the variation that occurs within the data during speech.

8.3 Audio Features to PCA Shape Values

We first consider a BLSTM with an input of audio features, and output of the first 8 principal components of the PCA of the facial deformations. Our model topology is described in detail in Chapter 5, and we cover the specific variation for this task here.

The input to the model are LogfBank audio features (Section 4.1). The audio features are standardised by scaling, such that each feature has zero mean and unit variation. The model output are the principal components described earlier in this chapter. When the PCA model was trained we used whitening to ensure unit component-wise variance, so no

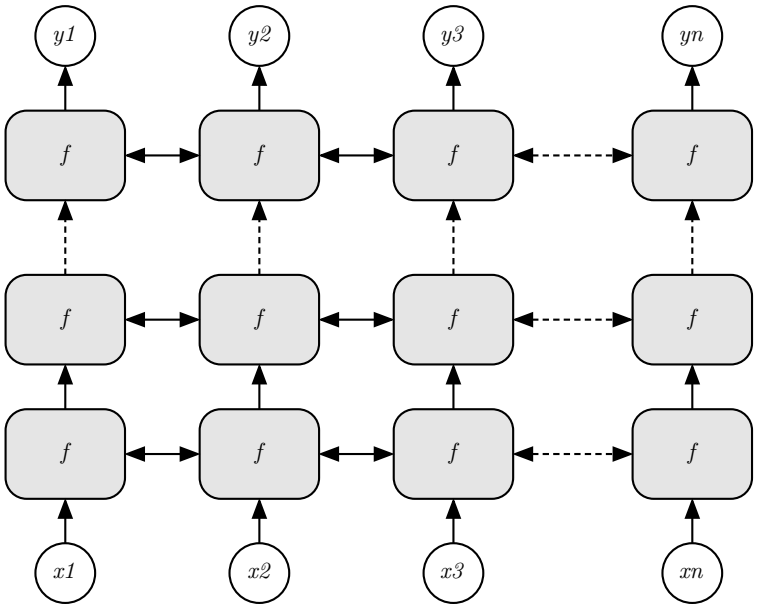


Figure 8.5: All the modelling of motion in this chapter uses a deep BLSTM. We train models for speech to PCA value prediction, and PCA to head pose angle prediction.

further scaling is used. We have a separate PCA model for each subject, (Section 3.8), so train a BLSTM model for each subject accordingly.

The model to predict PCA components is larger than the model we used to predict head pose from audio (Chapter 6). Here we use a tapering model of 3 bidirectional layers of 256, 128, 64 hidden units. This count is for each direction, so we double these figures.

We train our model on 90% of our data set, and use 10% for training validation. Before splitting our data we have removed a small uniformly random sample of data from the full corpus that has neither been used to train a model nor used for validation or model selection. This is the data we report results upon. Models are trained for between 25 and 100 epochs with a patience of 10 epochs. We save the network weights at every epoch, then the lowest

validation loss is used to select the best model. Training time is somewhere between 20 minutes and one hour for each epoch, depending on available GPU resources.

We augment our examples by taking short periods $t = n$, of each utterance in the interval $[59, 129]$ (Section 5.7). To maximise the chance of capturing the end points of long duration motion events we settle on the largest n for the comparisons we make in this chapter. Even though we train our model for n time steps, our model can make predictions of any time step duration, but has no 'memory' outside this receptive field.

8.3.1 Subject A Results

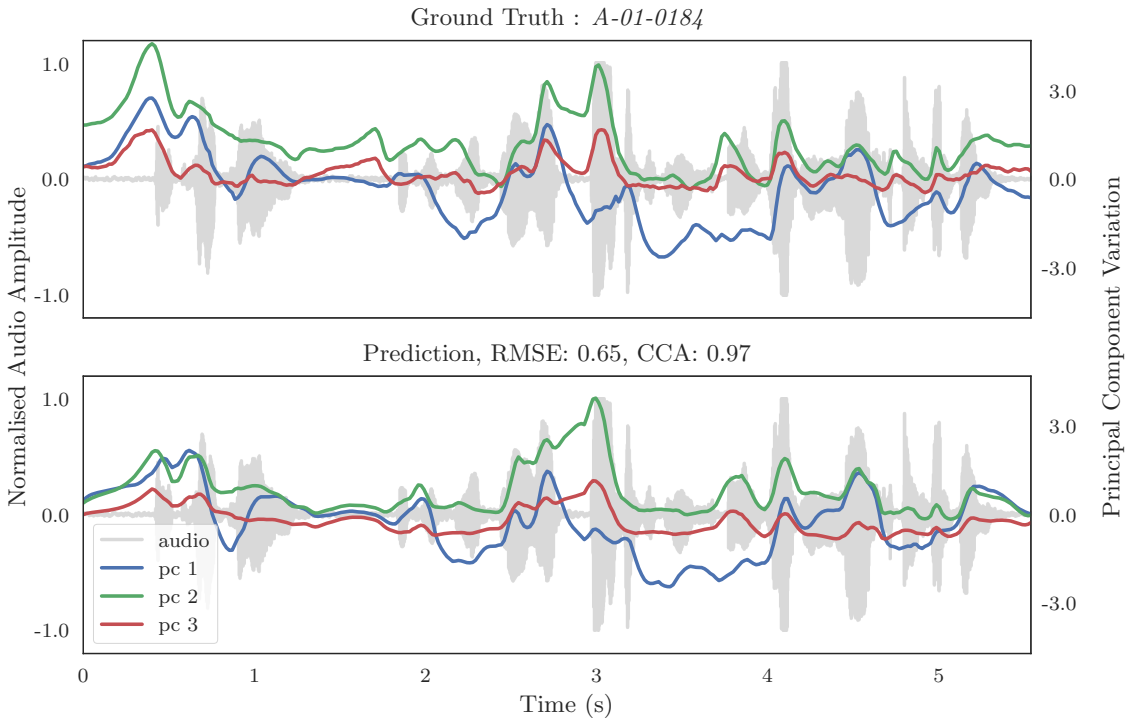


Figure 8.6: In this Figure, we show the results for prediction of the principal components of the shape model for Speaker A from audio features. The plot shows the ground truth and prediction for the first three components over the time domain audio. We report the RMSE and CCA for the combined 8 components. This is the same utterance shown at the start of this chapter, Figure 8.4.

We use our held-out data to make predictions from our trained model. For each example we prepare the data in the same way as the training data. We used the scaling parameters for the whole corpus to scale the input audio features, and we used the PCA model trained for Subject A to inverse transform the predicted component values. First, let us examine an example utterance, scene *A-01-0184*, that we have shown as the first example in this chapter in Figure 8.4. We show a prediction of the first 3 principal components for this utterance in Figure 8.6. We show only the first 3 components solely for reasons of clarity. The initial observation is how remarkably similar the plots of the ground truth and the prediction are. In particular, the most significant events in the time domain audio, where the face also is most active, are very well modelled. The CCA value confirms the qualitative assessment, with 0.97 indicating very significant correlation (recall: positive correlation is reported in the interval $[0, 1]$).

Figure 8.7 shows the same example in greater detail. Here, each component is plotted individually, and for each component we report RMSE in standard variations of principal components, and Pearson r correlation coefficient. As we have reduced the comparison to a pair of 1D time series, we do not need to use CCA. This detailed view shows how well the most significant components are modelled and highly correlated, and also, for the most dynamic of the lower ranked components. Only component 4 has somewhat lower figures, caused by a departure from the true values for the first half of the utterance.

We collate the results for correlation and RMSE for six Subject A test utterances in Table 8.1, where we show the values for each predicted component for each test utterance. We can see that generally the more significant components are predicted with good correlation and low error. There are some scenes with low or negative correlation for some of the lesser components, these components represent motion less well correlated with speech, for example the upper facial region. The accuracy of the more significant components is more important for lip sync accuracy.

Table 8.1: Predicting facial variation during speech from audio features. This table shows the results of predictions made from the held out examples from our data set for Subject A. Each scene identifier is shown across the top of the table, with the prediction values for each component in descending order. The RMSE unit is standard deviations of the PCA parameters.

PCA component		Scene ID					
		A-01-0184	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
1	RMSE	0.43	0.46	0.41	0.49	0.46	0.55
	COR	0.93	0.75	0.82	0.89	0.84	0.80
2	RMSE	0.78	0.57	1.20	0.57	0.50	0.65
	COR	0.75	0.69	0.61	0.59	0.82	0.58
3	RMSE	0.51	0.38	1.72	0.64	0.17	0.72
	COR	0.75	0.32	0.01	0.13	0.90	0.62
4	RMSE	1.04	0.97	1.60	1.03	0.57	2.26
	COR	0.45	0.58	−0.12	0.77	0.65	−0.24
5	RMSE	0.63	0.68	0.90	0.56	0.69	0.85
	COR	0.87	0.73	0.40	0.73	0.55	0.83
6	RMSE	0.62	0.47	0.75	0.59	0.59	1.27
	COR	0.70	0.03	0.13	0.56	0.66	0.12
7	RMSE	0.54	0.58	0.54	0.53	0.54	0.62
	COR	0.54	0.50	0.41	0.72	0.67	0.63
8	RMSE	0.41	0.85	0.34	0.95	0.48	1.89
	COR	0.69	0.56	0.58	0.30	0.48	−0.14

Table 8.2: Subject A reconstruction loss during speech from audio features. By reconstructing the full shape model from the PCA parameters, we can measure the reconstruction loss in meaningful units. We reconstruct the entire sequence in each utterance for both the ground truth and prediction and report RMSE in millimetres for the sequence. We also show CCA for the reconstructed prediction and ground truth for the entire sequence. The RMSE unit is millimetres.

Scene ID	*A-01-0184*	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
RMSE	1.44	1.29	2.49	1.41	1.05	2.32
CCA	0.97	0.96	0.97	0.99	0.98	0.95

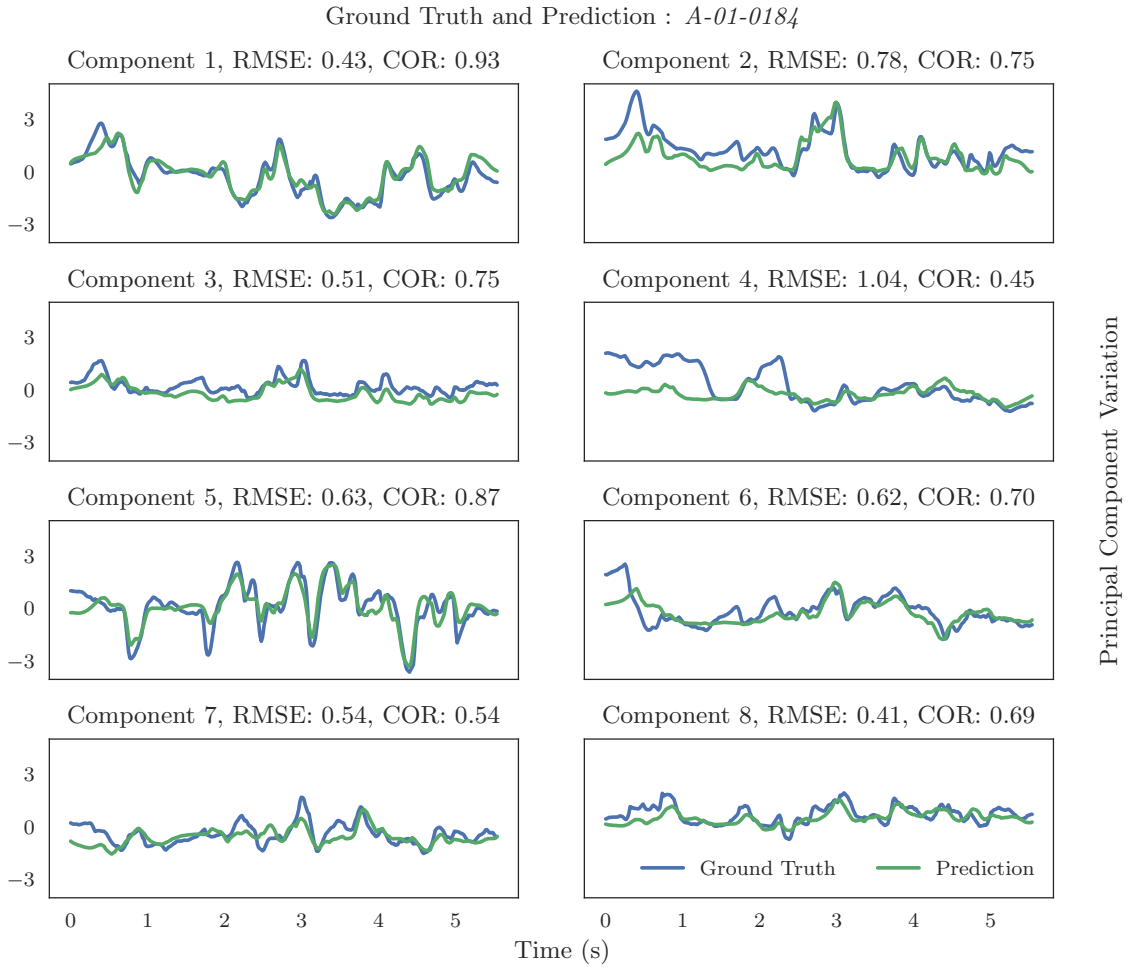


Figure 8.7: Again, using the same utterance in Figure 8.4, here we plot each of all 8 components individually for Speaker A, predicted from audio features. We report the RMSE and Pearson r correlation for each individual ground truth and prediction pair.

Table 8.2 shows a further evaluation of the results for predicting PCA components for Speaker A. Here we show the full reconstruction of the shape model from the components for each of the ground truth and predicted utterances. Here we do not need to concern ourselves with interpreting the value of each component, rather we can compare the real values of the RMSE for the entire utterance measured in millimetres. We also show CCA for the complete utterance.

8.3.2 Subject B Results

For Subject B we have a BLSTM model of the same architecture as Subject A (Figure 8.5), and in this section we show results from that model. Our training regime is also the same, and the models converge in a similar number of epochs of training. We also have a number of held-out samples for Speaker B that we use to evaluate the model. Just as for Subject A, we scale the audio features using the parameters for the whole corpus. Likewise, the PCA component features are not required to be scaled as the PCA model is whitened. In Figure 8.8 we show an example utterance, scene *J-03-0263*, in which we plot the first 3 components upon the time domain audio for the ground truth and the prediction of the utterance. Again we note the very good modelling of these components particularly in the region of high activity in the audio.

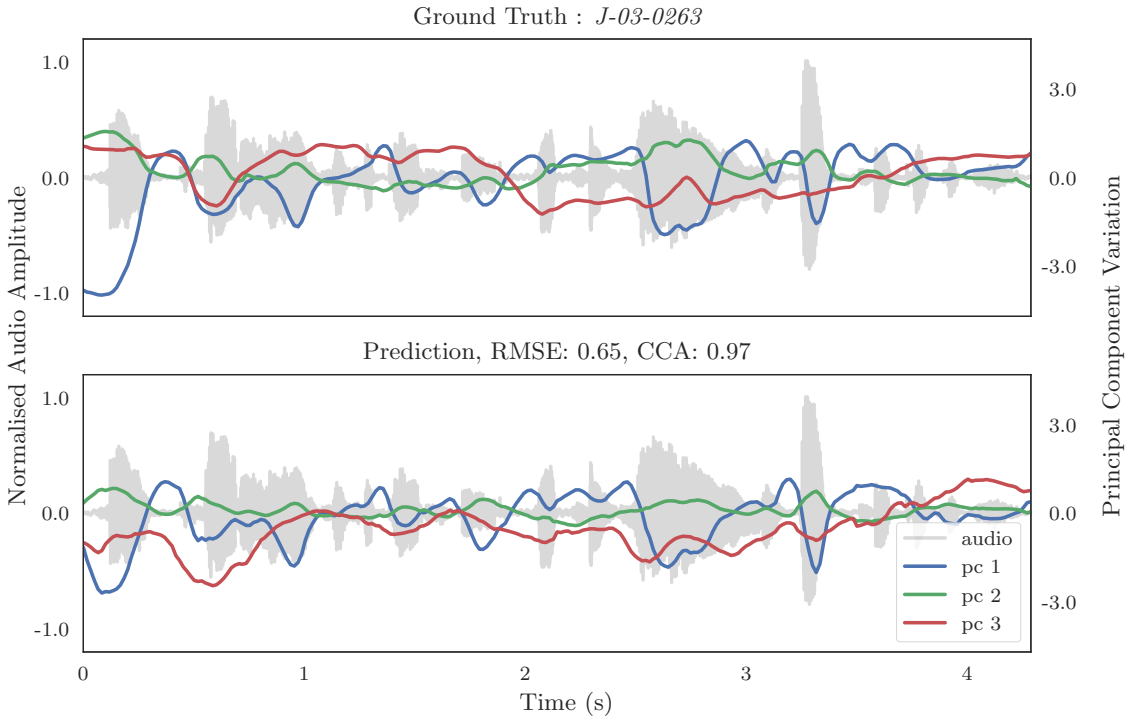


Figure 8.8: In this Figure, we show the results for prediction of the principal components of the shape model for Subject B, from audio features. The plot shows the ground truth and prediction for the first three components plotted over the time domain audio. We report the RMSE and CCA for the combined 8 components.

We examine this example in more detail in Figure 8.9. Here we show each individual component, and the RMSE and Pearson r correlation for each individual ground truth and prediction pair. We can see again that the most dynamic motion is modelled well, although this example is not as good as the best example for Speaker A. This is typical of Subject B, who is measurably less dynamic the Subject A (Figure 6.3b).

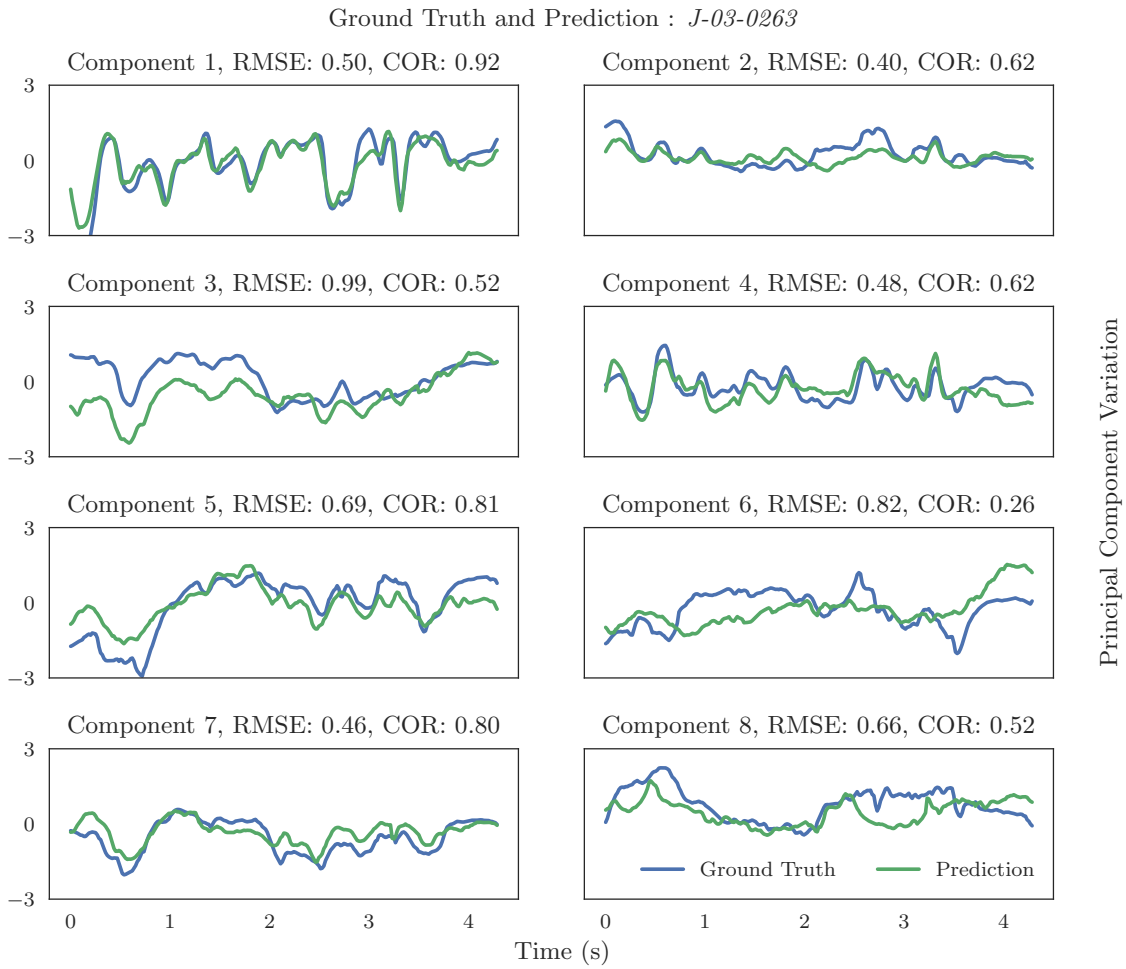


Figure 8.9: Showing the detail result for prediction of components from audio features for the utterance in Figure 8.8. Here we plot each of all 8 components individually. We report the RMSE and Pearson r correlation for each individual ground truth and prediction pair.

If we look at Table 8.3, here we show all the test utterances and all of their individual evaluations for Subject B. We can look through the results and see the example is typical of

our Subject B tests. Table 8.4 shows the reconstruction losses for all the utterances. Here we can see that the CCA is very high for all the utterances and the reconstruction loss is low.

Table 8.3: Speaker B prediction of facial variation during speech from audio features. The RMSE unit is standard deviations of the PCA parameters, we also show Pearson r correlation for each individual ground truth and prediction pair.

PCA component		Scene ID					
		J-01-0153	*J-02-0089*	*J-03-0263*	*J-04-0052*	*J-05-0256*	*J-06-0276*
1	RMSE	0.45	0.52	0.50	0.95	0.50	0.59
	COR	0.85	0.92	0.92	0.45	0.68	0.72
2	RMSE	0.24	0.26	0.40	0.65	0.25	0.26
	COR	0.56	0.81	0.62	0.34	0.28	0.44
3	RMSE	0.83	0.80	0.99	1.02	0.42	0.24
	COR	0.75	−0.30	0.52	−0.01	0.68	0.81
4	RMSE	1.07	0.68	0.48	1.68	0.63	0.56
	COR	0.77	0.41	0.62	0.10	0.58	0.87
5	RMSE	0.40	0.65	0.69	0.68	0.61	0.82
	COR	0.70	0.50	0.81	0.35	0.51	0.80
6	RMSE	1.50	0.96	0.82	1.22	0.82	0.53
	COR	−0.08	0.73	0.26	0.11	−0.40	0.04
7	RMSE	0.75	0.94	0.46	1.54	0.72	0.60
	COR	0.69	0.06	0.80	−0.22	0.66	0.83
8	RMSE	0.51	0.89	0.66	1.68	0.57	0.63
	COR	0.60	0.20	0.52	−0.22	−0.33	0.32

Table 8.4: Speaker B reconstruction loss during speech from audio features. By reconstructing the full shape model from the PCA parameters, we can measure the reconstruction loss. We also show CCA for the reconstructed prediction and ground truth. The RMSE unit is millimetres.

Scene ID	*J-01-0153*	*J-02-0089*	*J-03-0263*	*J-04-0052*	*J-05-0256*	*J-06-0276*
RMSE	1.73	1.55	1.56	2.71	1.26	1.25
CCA	0.99	0.97	0.97	0.96	0.97	0.96

8.3.3 Audio to PCA discussion

When we examine the results for the reconstructed scenes for both our subjects, we can see that they both show remarkably low error. Both subjects have very high CCA correlation

of 0.95 to 0.99. Similarly, RMSE is low, 1.2 to 2.7 mm. Recall that our estimate of the size of the facial markers is approximately 2 mm diameter. Clearly the activity of the face is highly correlated to the sounds we make when speaking. In this section we have shown that modelling the facial activity from audio features with BLSTM is a viable strategy with sufficient training data available for each subject. One significant short coming to highlight is we have speaker dependence. We need to train a model for each speaker, to make predictions only for that speaker. The reason why we have to train individual models is a limitation of our corpus. Due to dissimilar landmark locations, tracking occlusion and 3D extraction (Chapter 3), we are not able to create a unified shape model for both speakers. Re-meshing is a possibility, interpolating to a uniform number of points, and building a shape model from there. The most compelling option, as time allows, would be to re-track the data with a high resolution canonical mesh.

8.4 Phone to PCA Shape Values

In this section we experiment with aligned phones as input to our model to address the limitations of speaker dependence in a model driven by audio features. Aligned phones (Section 4.4) appear to remove many of the properties of speech that one might associate with non-verbal facial activity. Emphasis, intonation, prosody and emotion are all aspects of speech associated with pitch and energy in the speech signal. Although we might expect that phones could predict mouth shapes, and it has been proven to be the case [Taylor et al., 2017], predicting the motion of other regions of the face seem a more difficult task.

To test this difficulty we train a deep BLSTM model (Figure 8.5) similar to the model in the previous section, but with an input of time aligned phones. The output of our model are the values of the first 8 principal components of the PCA of the facial deformations. We train our model to predict only the facial deformations of Speaker A. With rig re-targeting we could consider any trained PCA model to be speaker independent. Our model is trained on approximately 90% of our corpus, with 10% used for training validation. We have the

same reserved sample of data described in the previous section used neither to train nor validate the model. We use this data to report the results. Training is between 25 and 100 epochs with a patience of 10 epochs. The lowest validation loss is used to select the best model.

8.4.1 Phone Model Results

Figure 8.10 shows an example scene, *A-01-0184* of Subject A making an utterance. For ready comparison, we use the same example shown in Section 8.3.1. We can observe from this example, that aligned phones appear to model PCA components very well. For this example we show the same correlation using CCA as the audio features model, with a small difference (less than 0.15) in RMSE.

Following the pattern for showing the audio features model results, Figure 8.11 shows the detail of each component plotted separately as ground truth and prediction pairs. It is worth drawing attention to were each model differs most significantly from the ground truth. In component 2 and 4, the first half of the utterance shows lower values for the predictions compared to ground truth, whereas component 6 shows the first half prediction running higher.

Table 8.5 shows the full results for the test selection for the phones to PCA model, where we can examine the RMSE and Pearson r correlation for each individual ground truth and prediction pair. Although it is interesting to see how well any particular value is modelled, the reconstruction losses shown in Table 8.6 give an overall view of the model performance.

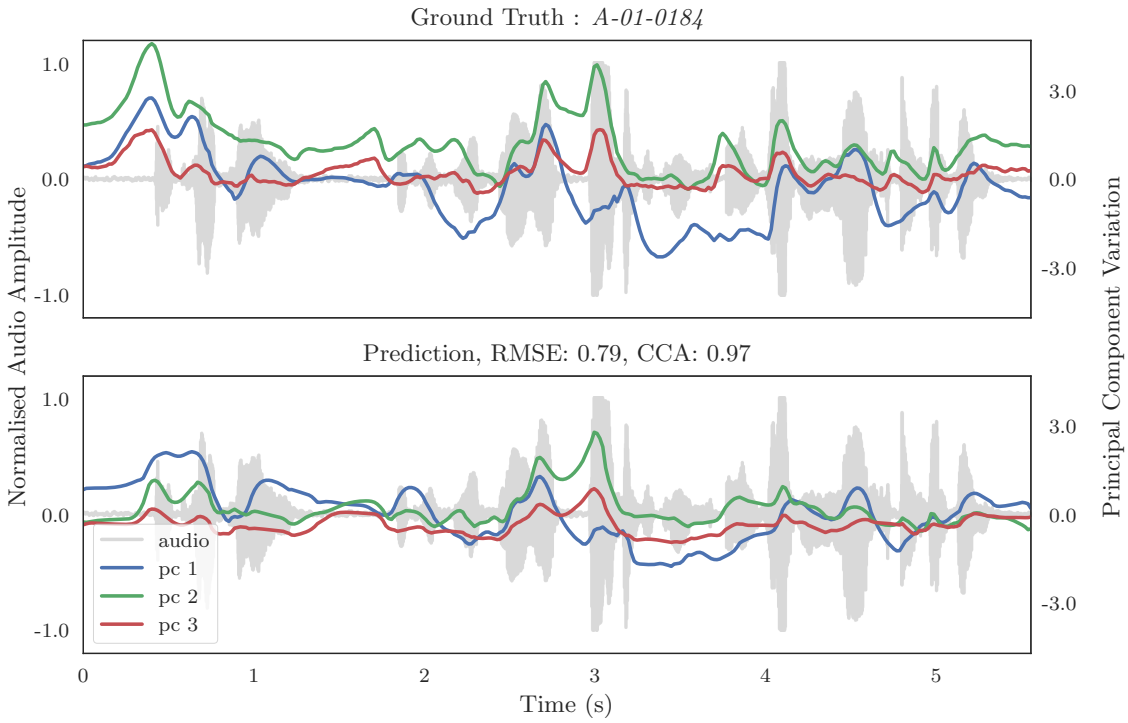


Figure 8.10: In this Figure, we show the results for prediction of the principal components of the shape model, from aligned phones, for Speaker A. The plot shows the ground truth and prediction for the first three components over the time domain audio. We report the RMSE and CCA for the combined 8 components. This is the same utterance shown at the start of this chapter, Figure 8.4 and, in the previous section, Figure 8.6.

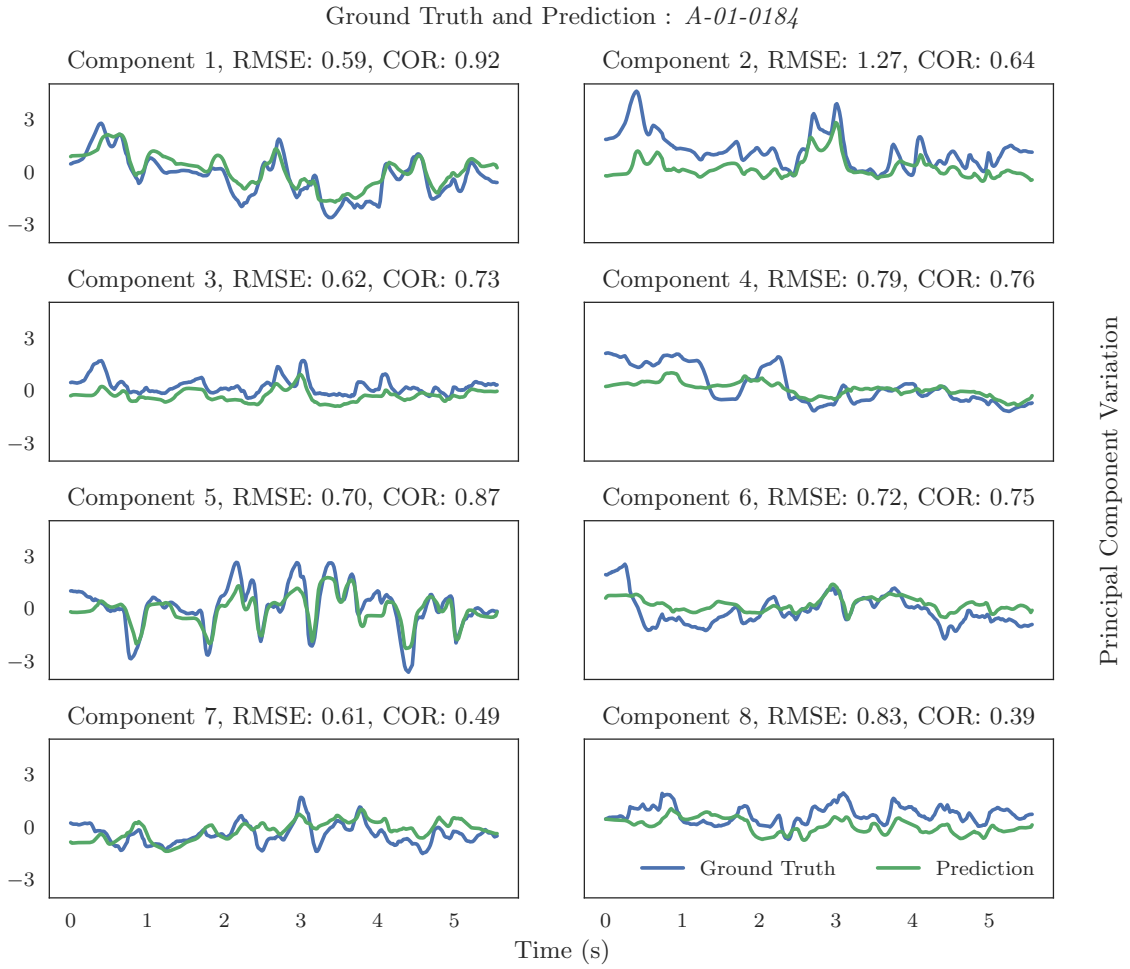


Figure 8.11: Detail result for prediction from phone features for Speaker A. Using the same utterance in Figure 8.4 and in Figure 8.6, here we plot each of all 8 components individually for Speaker A. We report the RMSE and Pearson r correlation for each individual ground truth and prediction pair.

Table 8.5: Speaker A predicting facial variation during speech from aligned phone features. The MSE unit is standard deviations of the PCA parameters.

PCA component		Scene ID					
		A-01-0184	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
1	RMSE	0.59	0.46	0.63	0.49	0.50	0.67
	COR	0.92	0.70	0.75	0.88	0.87	0.78
2	RMSE	1.27	0.70	1.08	0.48	0.74	0.62
	COR	0.64	0.71	0.33	0.67	0.89	0.61
3	RMSE	0.62	0.49	2.17	0.64	0.42	0.42
	COR	0.73	0.32	0.23	0.20	0.60	0.57
4	RMSE	0.79	1.05	1.18	1.47	0.66	1.39
	COR	0.76	0.27	−0.19	0.10	0.17	0.26
5	RMSE	0.70	0.48	0.60	0.51	0.62	0.58
	COR	0.87	0.89	0.89	0.78	0.68	0.85
6	RMSE	0.72	0.65	1.24	0.46	0.67	0.79
	COR	0.75	−0.26	−0.34	0.60	0.52	0.23
7	RMSE	0.61	0.86	0.54	0.52	0.54	0.90
	COR	0.49	0.09	0.64	0.72	0.60	0.80
8	RMSE	0.83	0.94	0.71	1.31	0.44	0.85
	COR	0.39	0.23	0.50	0.08	0.49	0.16

Table 8.6: Speaker A reconstruction loss during speech from phone features. By reconstructing the full shape model from the PCA parameters, we can measure the reconstruction loss. We also show CCA for the reconstructed prediction and ground truth. The MSE unit is millimetres.

Scene ID	*A-01-0184*	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
RMSE	1.85	1.44	2.66	1.61	1.30	1.69
CCA	0.97	0.97	0.97	0.97	0.97	0.96

Table 8.7: Comparison of PCA models. The mean result for each of the previous three models shown together for comparison.

Model	*Audio A*	*Audio B*	*Phone A*
Mean RMSE	1.68	1.67	1.76
Mean CCA	0.97	0.97	0.97

8.4.2 Phone Model Discussion

The main reason to consider a model driven by phones is to remove speaker dependence. Of course, the penalty to achieve speaker independence must not be too great. Table 8.7 shows the mean results of the 3 models discussed so far in this section. Here we see that the penalty is very low. The RMSE unit is millimetres, and the difference of less than 0.1 mm error is negligible. As the correlation measures the same for all models there is no reason not to choose phones as the driving feature of a facial shape model on the basis of performance. One reason might be the extra processing required for a phoneme based model, aligning phonemes to text is largely automatic with modern tools (Section 6.5.3), but not entirely. We did not record error rates of a first pass of our alignment process, but estimate it as in the region of 10%. We would need to conduct further tests to see if this reduces the accuracy of our facial predictions significantly, or, improve phone alignment by developing our own method.

8.5 Principal Components to Head Pose

In this section we consider the third question posed at the beginning of this chapter; can we predict head pose from the facial expression? Specifically we develop a model that accepts as *input* the principal components we have been predicting so far in this chapter, and outputs the head pose we have discussed in previous chapters. The hypothesis being that there is greater correspondence between the motion of the face and the rigid pose of the head. At the very least we do not suffer a domain gap, as we propose a prediction of one motion from another. A possible use case for this model is to introduce plausible head pose to any speech animation model when an animator or automatic system has created the facial animation.

To test this hypothesis we train a deep BLSTM model (Figure 8.5), with PCA component values as input. Output are the real values of head pose, which we describe as x, y, z rotation in degrees, with the x axis representing subject nod, y side to side shake and z being side to side roll. The *topology* of the model is similar to the two previous sections, but here we find we need a smaller model, similar to that for predicting speaker head pose of 3 bidirectional layers of 32 hidden units. Again, the count is for each direction, so should be doubled.

We follow our standardised method of splitting the corpus 90% for training and 10% for validation and model selection. We have reserved a small random selection of examples that never participate in training nor model selection for testing and reporting of results. We train our model for 25 to 100 epochs, with a patience of 10 epochs. The model with the lowest validation error is selected for testing.

8.5.1 Principal Components to Head Pose Results

Let us first examine some plots of head pose trajectories predicted from our principal components to head pose model. Figure 8.12 illustrates a selection from our held out test set of Subject A, with the ground truth and prediction plotted over the time domain audio of each utterance. Again, we show high correlation with the ground truth, with RMSE in the region of 2 to 3 degrees. We notice that some of the individual axes of rotation show remarkably similar trajectories to the ground truth, but show a parallel offset, accounting for a good proportion of the prediction RMS error.

The key observation to make, is the quantitative results for predicting head pose from PCA expression values are better, for the same examples, than from predictions made directly from the same audio features (Table 6.6).

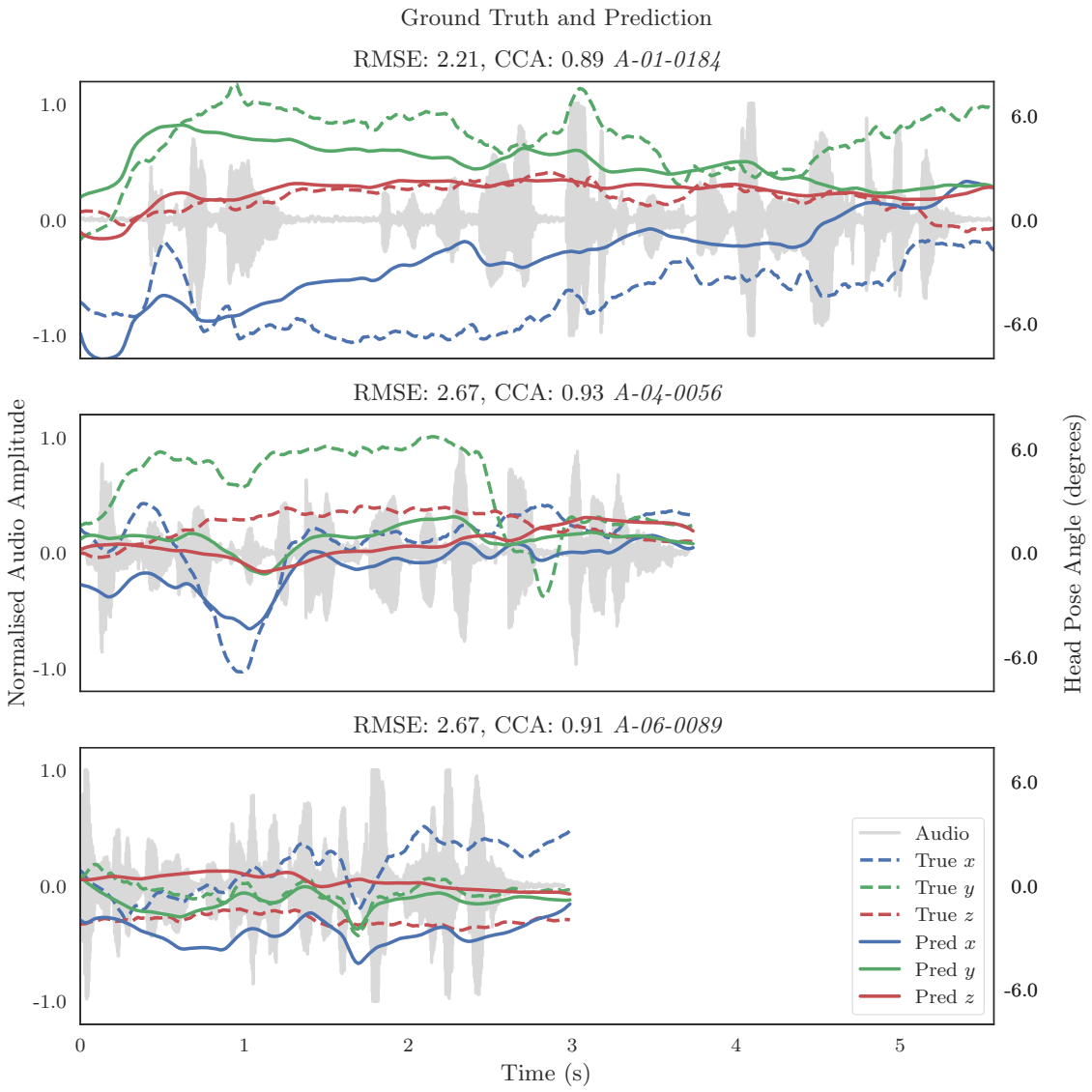


Figure 8.12: Predicting head pose from principal components. Here we show results for Subject A head pose predictions. We show the ground truth head pose and the predicted head pose plotted upon the audio amplitude of the utterance to give a good visual guide to effectiveness of the prediction. Note particularly example *A-06-0089*, where we see very strong correspondence, but global offset of a few degrees.

Table 8.8: Collated Subject A results for PCA to head pose. Here we show RMSE in degrees and CCA for head pose prediction for our selection of test scenes.

Scene ID	*A-01-0184*	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
RMSE	2.21	2.10	2.36	2.67	2.77	2.67
CCA	0.89	0.96	0.88	0.93	0.90	0.91

8.6 Joint Learning of Facial Expression and Head Pose

Of the three questions posed at the start of this chapter, we have been able to answer them all positively, by evaluating our results quantitatively and qualitatively. It is certainly interesting that predictions of the compressed facial activity, which we represent by the PCA model of our shape model over our corpus, are able to be made equally well from audio and phone features. That we can also make predictions of head pose from those components, is equally interesting, and we provided use cases for this application. One can not help ask if we can bridge these two models to form a phone to head pose predictor? Unfortunately, the answer is no, at least not directly. The gap between real PCA components and those predicted by our model is too great to fool our PCA to head pose model.

In this Section we model the complete facial activity during speech, along with the rigid pose of the speaker's head. We extend our earlier work on speaker head pose (Chapter 6) by modelling six DoF: the rotations of nod, yaw and roll and the translations on those axes. We have previously remarked that head pose has properties that make it difficult to model directly from speech. There is high measurable correlation between speech audio and head pose, yet a speaker repeating an utterance several times may move his head in significantly different manner on each repetition. In the previous section we observe a closer correspondence between facial activity and head pose during speech by modelling head pose directly from facial features rather than from audio. We hypothesise the modal gap is smaller between facial activity and head pose as the anatomical, physical and kinetic constraints are closer.

To exploit this observation, we first train a model to predict the facial animation from audio features, then in a second stage, encourage the model to learn head pose from the latent representation of the facial activity by using a separate objective for each mode.

8.6.1 Model Description

Clearly, much of the activity of the orofacial region has significant correspondence with speech production. Other regions of the face, along with head pose, have also been shown to have a relationship with speech Graf et al. [2002]. Our initial experiment was to consider how well we could predict face animation with a deep BLSTM, with audio features as input and our 8 PCA values as output. We observed good modelling, particularly of the more significant components. We further experimented with predicting head pose from facial expression, and observed improved performance over direct prediction from audio features. We hypothesise that the facial activity during speech closes the modal gap to head pose, i.e., the motion of the face is controlled by anatomy and limited by kinetic constraints, and so is the rigid motion of the head. When we try to model head pose directly from audio features we can not force the model to learn via that space.

Simply concatenating the rigid pose and shape values and training a Deep BLSTM did not provide the results we had seen with independently trained models. So we describe a *forked* model, with separate objectives for the six DoF head pose values and the PCA expression values. This allows independent control of each of these modalities, both in the topology, and in the training of the model. We found our best results were achieved by developing a model that only predicted the PCA values, then forking the model late in the latent space to a new stack of layers, with output to head pose values. Figure 8.13 illustrates the topology of the network. Our experience with these networks so far has been that the number of trainable parameters is limited by the quantity of our data. We find a properly converged model has a layer of 256 hidden units at input, with four subsequent layers, tapering in hidden units to 32, to the PCA objective. The head pose branch can be as little as two layers of 32 hidden units, significantly smaller than a model for predicting head pose alone. Recall, we use BLSTM, so the number of hidden units is doubled, as the count is for each direction.

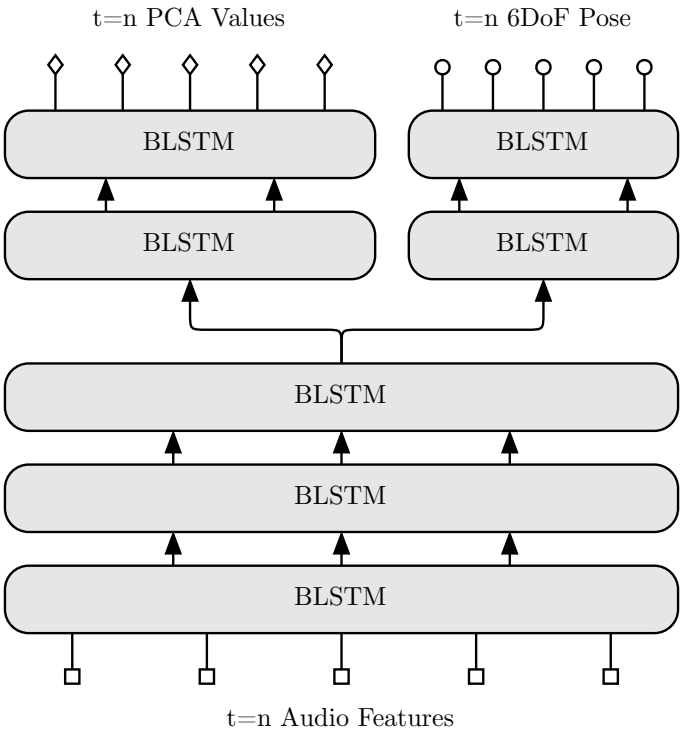


Figure 8.13: The topology of the deep BLSTM model. We pre-train the route to the PCA expression values, then train the whole model with separate objectives for the PCA values and the 6 DoF of the rigid head pose.

8.6.2 Training

We trained the networks on our data, split as before, with our test examples held out from the outset so they are never used for training or model selection. Both our objective functions are MSE, and our recent experiments in this section use *adam* [Kingma and Ba, 2014] for optimisation, with the parameters: $lr = 0.001$, $beta_1 = 0.9$, $beta_2 = 0.999$, $epsilon = 1 \times 10^{-8}$, $decay = 0.0$. Comparing identical models, *adam* converges more quickly than *rmsProp* [Tieleman and Hinton, 2012] for our task.

Training continues until no further improvement on the validation set, with a patience of 10 epochs. We first train a model with the sole objective of the PCA expression values (Section 8.3), then load those weights to the lower layers of our forked network (Figure 8.13). We recommence training of the entire network now with two objectives. Interestingly, the loss for PCA expression values continues to descend from this point. While we monitor both losses, our early subjective tests indicate viewers discriminate on the overall animation quality more by face accuracy than head pose, so we train until no further validation improvement on the PCA fork, with a patience of 5 epochs. We use the Keras framework [Chollet et al., 2015], with Tensorflow [Abadi et al., 2015] back end.

In this Section, the models are trained on one Subject, A, from our corpus.

8.6.3 Joint Learning Results

Table 8.9 shows the results of predictions for the held out utterance examples. We quantitatively evaluate our predictions in the following way: We use CCA to measure correlation for each predicted example by projecting to one base and calculating Pearson's r for the projection to the base. CCA > 0.5 represents significant correlation, and CCA $= 1.0$ is maximum correlation. We show CCA for the 8 predicted PCA component values, CCA for the head rotation values, and CCA for the head translation values. We report RMSE for the reconstructed PCA shape model for the whole utterance measuring the error in millimetres

(mm). We report RMSE for head rotation in degrees, and head translation in mm. We find CCA the more valuable measure for head pose as it indicates comparable modulation by the audio waveform, whereas a uniform offset in the trajectory can increase RMSE without adversely effecting the quality of the prediction. For the facial activity we desire *both* high correlation and low RMSE.

Table 8.9: For a quantitative evaluation of our predictions we show six scenes held out from our corpus. We show the reconstruction RMSE in mm for our shape model for the entire utterance, along with CCA for the true and predicted PCA components. We show the same measure for the six DoF of head pose, though the head pose rotation error unit is degrees.

Scene ID	*A-01-0184*	*A-02-0120*	*A-03-0203*	*A-04-0056*	*A-05-0263*	*A-06-0089*
RMSE PCA Rcn	1.46	1.37	2.84	1.68	1.24	1.88
RMSE Pose Rot	2.68	3.73	3.93	4.27	3.27	1.81
RMSE Pose Trn	2.85	1.81	3.84	2.84	3.64	2.57
CCA PCA Cmp	0.97	0.96	0.96	0.98	0.98	0.96
CCA Pose Rot	0.90	0.74	0.79	0.85	0.89	0.94
CCA Pose Trn	0.79	0.69	0.72	0.54	0.78	0.96

For qualitative assessment, we show plots of the trajectories of the first three Principal Components (Figure 8.14) and the rotation angles of nod (x), yaw (y) and roll (z) (Figure 8.15). On the plot in Figure 8.14 we report CCA for just the three plotted components. These components largely relate to the orofacial area and indicate lip sync performance. On Figure 8.15 we report measure in the same way as Table 8.9.

8.6.4 Subjective User Study

To further support our results, we conducted a subjective user study by predicting samples from our joint learning model, from the held out test data. Participants were shown the predictions, and were asked to distinguish between the prediction and the ground truth counterpart in a forced choice one stimulus discrimination test. We argue all forms of animation have some amount of perceptual noise. An example of (near) zero noise is a face to face meeting with another person, although even here, an actor could establish some degree of perceptual dissonance by adopting a particular behaviour.

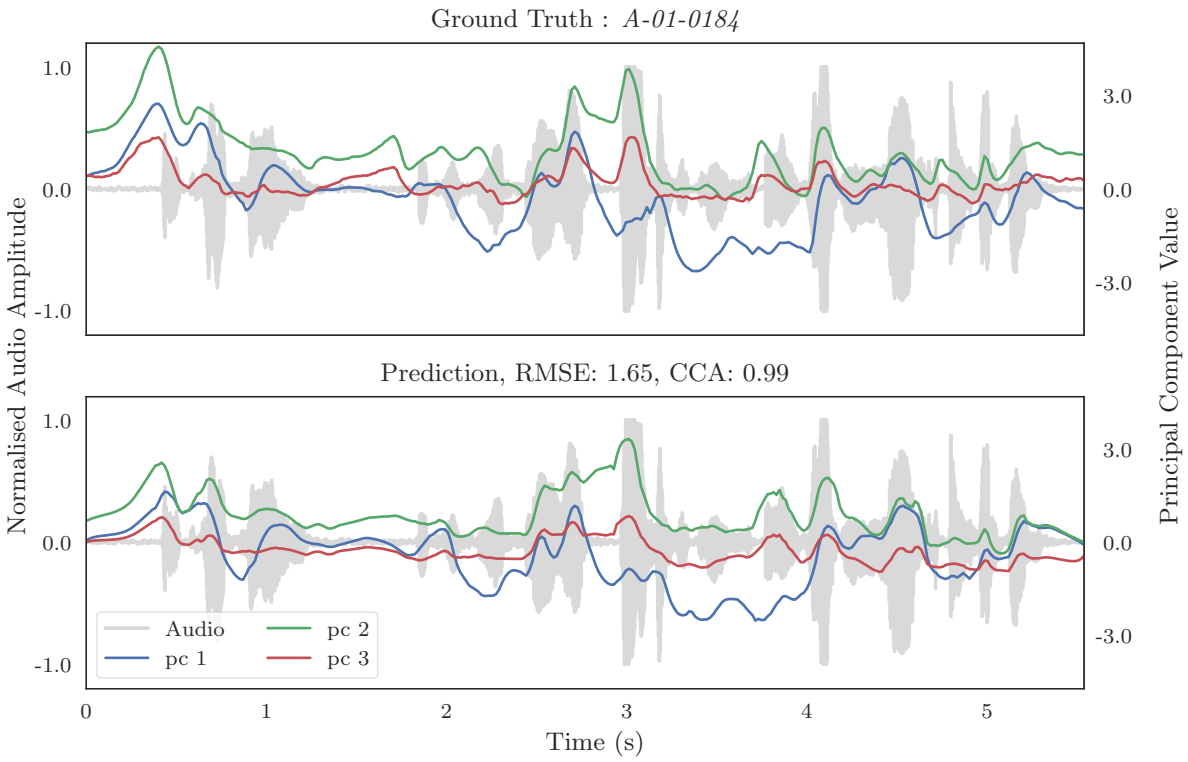


Figure 8.14: The ground truth and prediction of the first three Principal Components from the joint learning model. The first three components are largely associated with the orofacial area. We show CCA for the components plotted in PCA space. We can see clearly how the components are modulated by the audio. Qualitatively, one can observe how closely the component values in the prediction follow the ground truth. The model has been trained starting with the weights from the model in Section 8.3, and shows a small improvement in CCA.

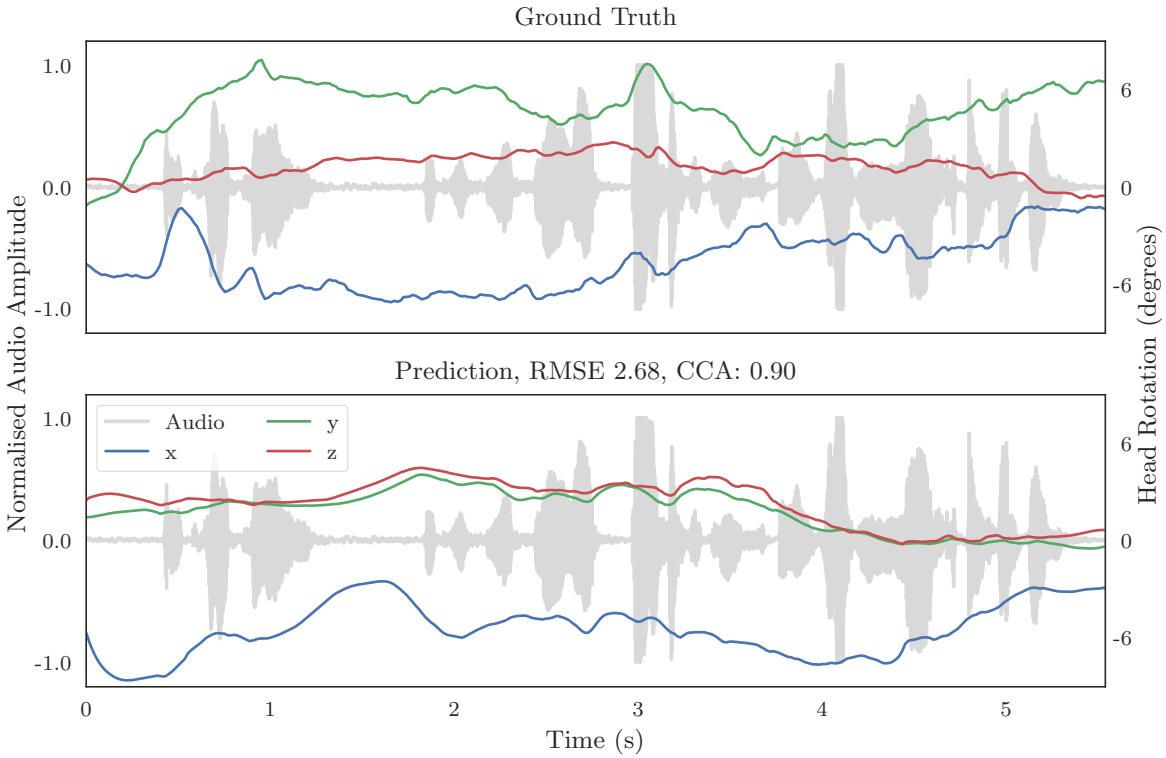


Figure 8.15: The ground truth and prediction of the rigid head pose angles from te joint learning model. We observe how head pose angle is also modulated by the audio, but has somewhat more diverse expectation. Quantitatively, for the three rotation axes we show CCA and RMSE (degrees).

We used SDT [Macmillan and Creelman, 2004] to calculate the sensitivity index $d\prime$, the distance between the mean of the stimulus and the mean of the noise in dimensionless units of standard deviation. Here, the noise is the ground truth 'Real' example and the stimulus, or signal, is the predicted example which we dub 'Fake'. We regard this as a suitable test as it is not effected by bias, eg. a user selecting the same answer repeatedly. We use the terms real and fake in deference to GAN terminology [Goodfellow et al., 2014], but here, our discriminator is a human viewer.

We show the raw results of the data collection in Figure 8.16, with the x axis showing the observed probability of all the answers. We see that R1, the first ground truth example, is

Table 8.10: Results of our user study for joint predictions from this chapter. A 'Hit' represents an answer of 'Fake' when the example is a prediction. A 'Correct Reject' is to answer 'Real' for a ground truth example. When a user responds 'Fake' to a real example, we report 'False Alarm', and finally, a 'Miss' is a 'Real' response to a prediction.

	Probability
Correct Reject	0.71
False Alarm	0.29
Hit	0.36
Miss	0.64

Table 8.11: We calculate the sensitivity index $d\prime$ to gain insight to viewer acceptance. In dimensionless units of standard deviation of the distribution of our examples, $d\prime$ gives a measure of how strongly viewers discriminated between real and predicted examples. Small values, less than one, indicate plausible animation predictions. We also show Z score for Hit and False Alarm, where $Z(p), p \in [0, 1]$, is the inverse of the cumulative distribution function of the normal distribution.

	$d\prime$	zH	zFA
Score	**0.192**	-0.366	-0.558

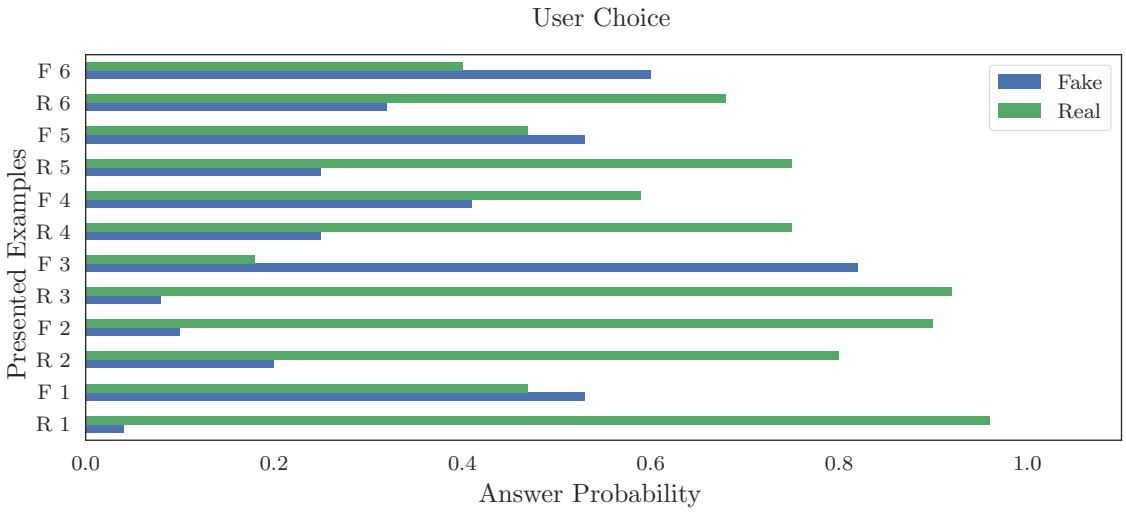


Figure 8.16: The data collected from a user study. We show the raw data as the percentage of responses to the question 'Real or Fake', when the user was shown an example from the corpus, or a counterpart prediction respectively.

selected as 'True' with $p \approx 0.95$ whereas the predicted counterpart, F1 is selected as 'True' with $p \approx 0.48$. Table 8.10 shows the Hit and Miss scores for the collected data. A 'Hit' represents an answer of 'Fake' when the example is a prediction. A 'Correct Reject' is to answer 'Real' for a ground truth example. When a user responds 'Fake' to a real example, we report 'False Alarm', and finally, a 'Miss' is a 'Real' response to a prediction. We use the equal variance model [Wickens, 2002] and the distance between the two distributions is calculated as the difference of the Z scores, $Z(Hit) - Z(FalseAlarm)$. Finally, we report the Z scores in Table 8.11 to calculate the *Sensitivity Index*, $d\prime$, which gives us a measure in standard deviations, of the distance between the noise and signal distributions.

8.7 Comparison with other methods

Three recent works: 'Synthesizing Obama: Learning Lip Sync from Audio Output Obama Video' [Suwajanakorn et al., 2017], 'Audio-driven facial animation by joint end-to-end learning of pose and emotion' [Karras et al., 2017], and 'A deep learning approach for generalized speech animation' [Taylor et al., 2017] *arguably* represent the state of the art for speech animation.

To make direct comparison with any of these works is difficult as they all have different goals and present different methods of evaluation. The first work [Suwajanakorn et al., 2017] is perhaps easiest to remove from our comparison. It is a 2.5D method for replacing the lip motion with new speech for one targeted speaker, Barack Obama. The only evaluation provided are videos of predictions depicting Obama speaking, clearly we can not compare in an objective way our results with theirs. Karras et al. [2017] provide results from subjective testing, for their method that uses performance capture as ground truth. Curiously, they could easily provide empirical measurements, for example reconstruction loss to the ground truth mesh, but they do not. Taylor et al. [2017] provide both subjective comparison to their ground truth, and, error measure to the AAM parameters of their ground truth. The useful comparison here is the pixel loss. In their Figure [Taylor et al., 2017, Figure 10],

they indicate pixel MSE as ≈ 15. We can convert our reconstruction loss in millimetres for the facial features (The first row in Table 8.9) to pixel loss by counting pixels in a known dimension. We use the diameter of the iris, which is close to 12mm for adults, and count 30 pixels across.

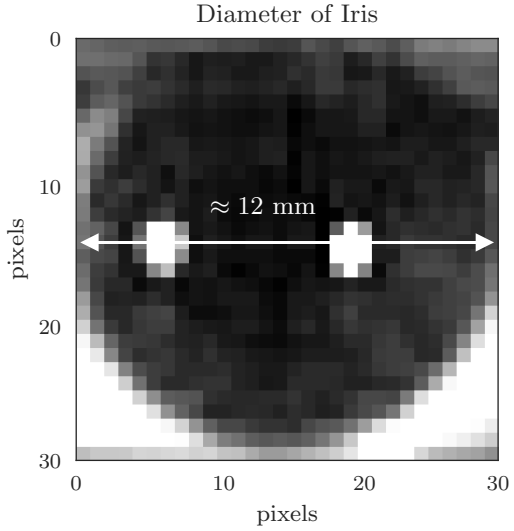


Figure 8.17: Estimate of pixel loss using iris diameter. We count 30 pixels for 12mm.

Table 8.12: Comparing the results of our method and Taylor et al. [2017]. We covert our results to pixel error, and Taylor et al. [2017] to millimetres to enable a direct comparison.

Method	Pixels	Millimetres
Taylor et al. [2017]	3.9	1.5
Our Method	4.4	1.7

We show both Taylor et al. [2017] and our own method converted to pixels, along with Taylor et al. [2017] converted to mm in Table 8.12. A difference of 0.2 mm between the methods is certainly small, and well within the tolerance of human ability to place a marker on a frame of video.

For subjective tests, Karras et al. [2017] and Taylor et al. [2017] provide a comparison with their method and the ground truth used for model training, as do we. We can interpret our

results shown in Table 8.10 in a slightly different way by accumulating correct rejection and hit as a preference for the prediction, and conversely, false alarm and miss as a preference for the ground truth. If we then present all three results as a probability of choice we can make the comparison shown in Table 8.13.

Table 8.13: Comparing the normalised results of our user study and other recent methods. Note that users prefer the prediction over ground truth for Taylor et al. [2017], suggesting some perceptual noise in the comparison. Karras et al. [2017] have the worst result, but the highest fidelity model, allowing users to see minor anomalies and hence identify the prediction.

Method	Truth	Prediction
Taylor et al. [2017]	0.46	0.54
Karras et al. [2017]	0.77	0.23
Our Method	0.53	0.46

Table 8.13 makes an interesting comparison, and suggests that subjective testing *within the domain* of facial animation is probably only useful for measuring the ability of an authors own method. Karras et al. [2017] use a high resolution mesh for the performance capture of their ground truth. The higher resolution likely makes it easier for users to discriminate between real and predicted examples. There is no doubt that the examples they show are impressive, and this comparison does not do full justice to their method. Our method and Taylor et al. [2017] have the most similar methods. We both use AAMs to track video performance, but Taylor et al. [2017] train their model on AAM parameters in 2D, whereas we derive a 3D shape model for training. Although Taylor et al. [2017] have better results than our method, we must allow a margin for noise as they show results for their method *better* than ground truth. Note that Taylor et al. [2017] only predict the orofacial area.

In light of the comparisons made in this section, we claim near parity to Taylor et al. [2017]. In addition, our model predicts the *complete* facial expressions *and* rigid head pose.

8.8 Discussion

Future work involves seeking a generalisation of our method so we can not only train on multiple speakers from within our corpus, but also predict speakers from outside our corpus. Our technique works equally well for each speaker individually, so we are optimistic regarding that goal. We have already shown in Section 8.4 that we can predict the PCA expression values from phones rather than audio features, so the concept of speaker normalisation for animation appears achievable. The one drawback of a phoneme based model is the requirement for an aligned transcript extracted from the audio, either by annotation, or prior processing. We now suggest that this step could be achieved by training a large audio data base to learn phoneme alignment, and using its lower layers as a front end to our model here, thereby providing speaker normalisation while retaining convenient microphone input.

A general system for character animation would need to drive any reasonable character, which may or may not have human-like features, *and* co-exist within an industry standard production pipeline. We have already mentioned rig re-targeting [Taylor et al., 2017] as a technique for sampling a deforming mesh to learn animator friendly blend-shape weights, thus merging the pipeline forward of our parametrised shape model.

The concept of forcing a model to learn via an intermediate modality presents new ideas for tackling other visual modes that are under independent control in addition to head pose. We arrive at the final experiment of the work with a data driven model for predicting a complete character head animation solely from audio speech features input, with model predictions that include accurate lip sync, animation of all the facial features, *and* rigid head pose rotations and translations.

9 Discussion

In this thesis we developed algorithms to predict the head pose of a human speaker. Specifically, we developed algorithms to predict the rigid head pose of a speaker, solely from the speech audio. As we pursued our goals we learnt we could apply our methods not just to the speaker, but also to the listener in a conversation. Our work culminates in methods to predict the complete facial expression, lip sync and rigid head pose of the speaker, simultaneously, from only the audio.

The goal of this thesis was to develop models that can *learn* visual actions from data without semantic labelling, and then, provide plausible speech animation from easily recorded sound.

Our path to achieving this goal was to develop a corpus that represented the *style* of speech we wanted to model; expressive, prosodic, natural speech. This allowed us to recognise that a significant part of the motion, or gesture, during speech is not deterministic. This key observation greatly informed the direction of the work, and we include models that can predict a great variety of head pose trajectories from a single utterance.

Speech animation, the process of animating a human like model to give the impression it is talking, still relies on the work of skilled animators, or performance capture. Both of these approaches are time consuming, expensive, and lack scalability. We are exploring an area of speech animation that is content driven; providing a sparse input to predict rich output. Applications for this technology seem most obviously connected to the entertainment industry; films, television, and games all require a high volume of realistic animation. There are also compelling arguments for HCI, perhaps for psychological therapy or medical diagnoses. Virtual and augmented reality is currently enjoying unprecedented investment,

and we will want to project our presence, with ever increasing realism, into these worlds. Worlds where the physical limits of time and proximity are no barrier to social interaction. If we consider our work as a general prediction of human action, the scope of applications expands considerably; just one example use case are autonomous road going vehicles.

9.1 Contributions

Much of the previous work on predicting the rigid head pose of the speaker involved clustering of the motion, or labelling trajectories in some way. We completely dispense with this idea and present models that learn motion only from data. Where we can compare our work with other recent authors, our empirical measurements are far ahead of theirs (Table 6.13) . Going beyond our contemporaries, we are able to show that the BLSTM model is able to predict head pose from phoneme features, removing much of the speaker dependent attributes of audio.

We then go further by introducing, for the first time, *generative* predictions of speaker head pose. We can predict, not just one, but a very large quantity of plausible head pose trajectories. We remain unique in this regard.

For the head pose of the listener in conversation, even some of the most recent work in the literature retains the idea of a rule based system, whereby a listener's response is generated when the pitch of the speaker's voice rises, or, a period of time has elapsed, for example [Schröder et al., 2015]. Again, we generate listener responses by *learning* from the data. We predict listener head pose actions directly from the speaker's voice.

Finally, we turn our attention to modelling the entire facial expression of the speaker. There are relatively few examples of modelling the 3D facial expression of a speaker, predicted just from the speaker's voice. Of the very recent examples we are able to consider, our work compares very favourably using the same measures that those authors present (Section 8.7). However, not only do we predict the facial expression, including accurate lip sync, our

model also predicts the speaker's rigid head pose. Table 8.12 shows we are predicting facial landmarks with comparable accuracy to Taylor et al. [2017]. To the best of our knowledge, we are not aware of a model from another author that makes predictions of facial expression, with this fidelity, *and* rigid head pose.

9.2 Future Direction

There are clear avenues of pursuit for improving our work. Within this thesis, we have already recognised the potential for developing audio input that does not rely on hand engineered features. Although standard audio features have been used to great effect for speech recognition for example, our task is somewhat different. We do not agree that feature extraction that performs well in alternative domains are the only choice, and remain committed to developing models that accept raw time domain audio as input. We have shown that our technique works well for phoneme input, suggesting a path toward an audio front end, trained on a large speech corpus, to predict phonemes. Collecting, or finding, this data is a far simpler task than collecting a sufficiently large body of multi-modal data. This approach gives us the opportunity to remove the last layer and re-purpose the objective, leaving a much smaller set of parameters to train for our specific task. We would gain not only the ideal speech representation for our task, but also, robust speaker normalisation.

Collecting multi-modal data was a substantial part of our work, certainly in terms of hours spent. Much of this time was consumed tracking the video, and this perhaps highlights why performance capture is such an active research area. There have been some very interesting recent developments in performance capture. All of these recent methods use models trained on high resolution ground truth 3D meshes, and therein lies the difficulty. Large scale multi-view stereo, a typical route to obtaining such high resolution ground truth data, is the preserve of the most well funded organisations. To counter this difficulty, we

propose generating appropriate *synthetic* data to create a model to track our more sparsely recorded live data.

The most significant part of the data we did not take advantage of for training, are the combined appearance values across the multiple views. Here too, the recording of dynamic multi-view textures is a very active research area, and we must enter this area to extract this information from our model.

What do we hope to gain by increasing the density and detail of the model? We can gain access to the minutiae of expression, the subtleties, that human viewers are so attuned to recognising when meeting other people, or viewing representations of them. It is worth noting, that despite almost unlimited resources in, for example, the motion picture industry, a truly convincing virtual human has not yet been seen.

We believe the near future of our work is much closer to the raw data. Our long term goal is content driven speech animation indistinguishable from human performance.

10 Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., Bengio, Y., Bergeron, A., Bergstra, J., Bisson, V., Bleecher Snyder, J., Bouchard, N., Boulanger-Lewandowski, N., Bouthillier, X., de Brébisson, A., Breuleux, O., Carrier, P.-L., Cho, K., Chorowski, J., Christiano, P., Cooijmans, T., Côté, M.-A., Côté, M., Courville, A., Dauphin, Y. N., Delalleau, O., Demouth, J., Desjardins, G., Dieleman, S., Dinh, L., Ducoffe, M., Dumoulin, V., Ebrahimi Kahou, S., Erhan, D., Fan, Z., Firat, O., Germain, M., Glorot, X., Goodfellow, I., Graham, M., Gulcehre, C., Hamel, P., Harlouchet, I., Heng, J.-P., Hidasi, B., Honari, S., Jain, A., Jean, S., Jia, K., Korobov, M., Kulkarni, V., Lamb, A., Lamblin, P., Larsen, E., Laurent, C., Lee, S., Lefrancois, S., Lemieux, S., Léonard, N., Lin, Z., Livezey, J. A., Lorenz, C., Lowin, J., Ma, Q., Manzagol, P.-A., Mastropietro, O., McGibbon, R. T., Memisevic, R., van Merriënboer, B., Michalski, V., Mirza, M., Orlandi, A., Pal, C., Pascanu, R., Pezeshki, M., Raffel, C., Renshaw, D., Rocklin, M., Romero, A., Roth, M., Sadowski, P., Salvatier, J., Savard, F., Schlüter, J., Schulman, J., Schwartz, G., Serban, I. V., Serdyuk, D., Shabanian, S., Simon, E., Spieckermann,

S., Subramanyam, S. R., Sygnowski, J., Tanguay, J., van Tulder, G., Turian, J., Urban, S., Vincent, P., Visin, F., de Vries, H., Warde-Farley, D., Webb, D. J., Willson, M., Xu, K., Xue, L., Yao, L., Zhang, S., and Zhang, Y. (2016). Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688.

Alexander, O., Rogers, M., Lambeth, W., Chiang, M., and Debevec, P. (2009). The digital emily project: photo real facial modeling and animation. In *ACM SIGGRAPH 2009 Courses*, page 12. ACM.

Allwood, J., Nivre, J., and Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of semantics*, 9(1):1–26.

Anderson, R., Stenger, B., Wan, V., and Cipolla, R. (2013). Expressive visual text-to-speech using active appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3382–3389. IEEE.

Andor, D., Alberti, C., Weiss, D., Severyn, A., Presta, A., Ganchev, K., Petrov, S., and Collins, M. (2016). Globally normalized transition-based neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2442–2452.

Bäckström, T. and Magi, C. (2006). Properties of line spectrum pair polynomials—a review. *Signal processing*, 86(11):3286–3298.

Bailenson, J. N. and Blascovich, J. (2004). Avatars. In *Encyclopedia of Human-Computer Interaction, Berkshire Publishing Group*, pages 64–68.

Bailenson, J. N., Blascovich, J., Beall, A. C., and Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, 29(7):819–833.

Baker, S. and Matthews, I. (2001). Equivalence and efficiency of image alignment algorithms. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–1090–I–1097 vol.1.

Ben Youssef, A., Shimodaira, H., and Braude, D. A. (2013). Articulatory features for speech-driven head motion synthesis. *Proceedings of Interspeech, Lyon, France*.

Bengio, Y., Laufer, E., Alain, G., and Yosinski, J. (2014). Deep generative stochastic networks trainable by backprop. In *Proceedings of The 31st International Conference on Machine Learning*, pages 226–234.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: a cpu and gpu math expression compiler. In *Proceedings of the Python for scientific computing conference (SciPy)*, volume 4, page 3. Austin, TX.

Bevacqua, E., De Sevin, E., Hyniewska, S. J., and Pelachaud, C. (2012). A listener model: introducing personality traits. *Journal on Multimodal User Interfaces*, 6(1-2):27–38.

Birdwhistell, R. L. (1952). Introduction to kinesics : an annotation system for analysis of body motion and gesture.

Black, M. and Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions usinglocal parametric models of image motion. pages 374–381.

Boersma, P. and Weenik, D. (1996). Praat: a system for doing phonetics by computer. report of the institute of phonetic sciences of the university of amsterdam. *Amsterdam: University of Amsterdam.*

Borshukov, G., Piponi, D., Larsen, O., Lewis, J. P., and Tempelaar-Lietz, C. (2005). Universal capture-image-based facial animation for the matrix reloaded. In *ACM Siggraph 2005 Courses*, page 16. ACM.

Bouguet, J.-Y. (2002). Camera calibration tool-box for matlab. *http://www. vision. caltech. edu/bouguetj/calib_doc/.*

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2016). Generating sentences from a continuous space. *CoNLL 2016*, page 10.

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools.*

Brand, M. (1999). Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 21–28. ACM Press/Addison-Wesley Publishing Co.

Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 353–360. ACM Press/Addison-Wesley Publishing Co.

Brenton, H., Gillies, M., Ballin, D., and Chatting, D. (2005). The uncanny valley: does it exist. In *Proceedings of conference of human computer interaction, workshop on human animated character interaction*. Citeseer.

Busso, C., Deng, Z., Grimm, M., Neumann, U., and Narayanan, S. (2007). Rigid head motion in expressive speech animation: Analysis and synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1075–1086.

Busso, C., Deng, Z., Neumann, U., and Narayanan, S. (2005). Natural head motion synthesis driven by acoustic prosodic features. *Journal of Visualization and Computer Animation*, 16(3-4):283–290.

Butterworth, B. and Beattie, G. (1978). Gesture and silence as indicators of planning in speech. In *Recent advances in the psychology of language*, pages 347–360. Springer.

Butterworth, B. and Hadar, U. (1989). Gesture, speech, and computational stages: a reply to mcneill.

Cao, Y., Tien, W. C., Faloutsos, P., and Pighin, F. (2005). Expressive speech-driven facial animation. *ACM Transactions on Graphics (TOG)*, 24(4):1283–1302.

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M., et al. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696.

Cassell, J. et al. (2000). Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. *Embodied conversational agents*, 1.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994). Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pages 413–420. ACM.

Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Vaucelle, C., and Vilhjálmsson, H. (2002). Mack: Media lab autonomous conversational kiosk.

Cassell, J., Vilhjálmsson, H. H., and Bickmore, T. (2001). Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 477–486. ACM.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*.

Choi, J. D. (2016). Dynamic feature induction: The last gist to the state-of-the-art. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 271–281.

Chollet, F. et al. (2015). Keras. `https://github.com/fchollet/keras`.

Cohen, M. M. and Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. In *Models and techniques in computer animation*, pages 139–156. Springer.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685.

Cootes, T. F. and Taylor, C. J. (2001). Statistical models of appearance for medical image analysis and computer vision. In *Medical Imaging 2001: Image Processing*, volume 4322, pages 236–249. International Society for Optics and Photonics.

Dai, Y., Li, H., and He, M. (2014). A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, 107(2):101–122.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., and Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, 380:499 EP –.

Dautenhahn, K. and Werry, I. (2004). Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition*, 12(1):1–35.

De Cheveigné, A. and Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930.

Deng, L., Hinton, G., and Kingsbury, B. (2013a). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603. IEEE.

Deng, L., Li, J., Huang, J.-T., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., et al. (2013b). Recent advances in deep learning for speech research at Microsoft. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8604–8608. IEEE.

Deng, Z., Narayanan, S., Busso, C., and Neumann, U. (2004). Audio-based head motion synthesis for avatar-based telepresence systems. In *Proceedings of the 2004 ACM SIGMM workshop on Effective telepresence*, pages 24–30. ACM.

Ding, C., Xie, L., and Zhu, P. (2014). Head motion synthesis from speech using deep neural networks. *Multimedia Tools and Applications*, pages 1–18.

Ding, C., Zhu, P., and Xie, L. (2015). Blstm neural networks for speech driven head motion synthesis. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Edwards, G. J., Cootes, T. F., and Taylor, C. J. (1998). Face recognition using active appearance models. In *European conference on computer vision*, pages 581–595. Springer.

Ekman, P. and Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98.

Ekman, P. and Friesen, W. V. (1978). Facial action coding system: a technique for the measurement of facial movement.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752.

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of speech and hearing research*, 11 4:796–804.

Freud, S. (1955). The 'uncanny' (1919) standard edition 17: 217-256 London.

Gehring, J., Miao, Y., Metze, F., and Waibel, A. (2013). Extracting deep bottleneck features using stacked auto-encoders. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3377–3381. IEEE.

Gold, T. and Pumphrey, R. J. (1948). Hearing. i. the cochlea as a frequency analyzer. *Proceedings of the Royal Society of London B: Biological Sciences*, 135(881):462–491.

Goldin-Meadow, S., McNeill, D., and Singleton, J. (1996). Silence is liberating: removing the handcuffs on grammatical expression in the manual modality. *Psychological review*, 103(1):34.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

Graf, H. P., Cosatto, E., Strom, V., and Huang, F. J. (2002). Visual prosody: facial movements accompanying speech. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 396–401.

Graves, A. (2012). Supervised sequence labelling. In *Supervised sequence labelling with recurrent neural networks*, pages 5–13. Springer.

Graves, A. (2013). Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850.

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *ICML*, volume 14, pages 1764–1772.

Greenwood, D., Laycock, S., and Matthews, I. (2017a). Predicting head pose from speech with a conditional variational autoencoder. *Proc. Interspeech 2017*, pages 3991–3995.

Greenwood, D., Laycock, S., and Matthews, I. (2017b). Predicting head pose in dyadic conversation. In *International Conference on Intelligent Virtual Agents*, pages 160–169. Springer.

Greenwood, D., Matthews, I., and Laycock, S. (2018). Joint learning of facial expression and head pose from speech. In *Proc. Interspeech 2018*, pages 2484–2488.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232.

Haag, K. and Shimodaira, H. (2016). Bidirectional lstm networks employing stacked bottleneck features for expressive speech-driven head motion synthesis. In *International Conference on Intelligent Virtual Agents*, pages 198–207. Springer.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech*, pages 223–227.

Hartley, R. I. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition.

Heinzel, G., Rüdiger, A., and Schilling, R. (2002). Spectrum and spectral density estimation by the discrete fourier transform (dft), including a comprehensive list of window functions and some new at-top windows.

Heylen, D., Poel, M., Nijholt, A., et al. (2001). Generation of facial expressions from emotion using a fuzzy rule based system. In *Australian Joint Conference on Artificial Intelligence*, pages 83–94. Springer.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580.

Hjortsjö, C.-H. (1969). *Man's face and mimic language*. Studen litteratur.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hofer, G. and Shimodaira, H. (2007). Automatic head motion prediction from speech data. In *Eighth Annual Conference of the International Speech Communication Association*.

Honnibal, M. and Johnson, M. (2015). An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, pages 321–377.

Huang, J.-T., Li, J., and Gong, Y. (2015). An analysis of convolutional neural networks for speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4989–4993. IEEE.

Hubal, R. C., Fishbein, D. H., Sheppard, M. S., Paschall, M. J., Eldreth, D. L., and Hyde, C. T. (2008). How do varied populations interact with embodied conversational agents? findings from inner-city adolescents and prisoners. *Computers in Human Behavior*, 24(3):1104–1138.

Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57(S1):S35–S35.

Jentsch, E. (1997). On the psychology of the uncanny (1906) 1. *Angelaki: Journal of the Theoretical Humanities*, 2(1):7–16.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM.

Jordan, M. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Eighth Annual Conference of the Cognitive Science Society, 1986*, pages 513–546.

Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

Karras, T., Aila, T., Laine, S., Herva, A., and Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph.*, 36(4):94:1–94:12.

Kasi, K. and Zahorian, S. A. (2002). Yet another algorithm for pitch tracking. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 1, pages I–361. IEEE.

Kendon, A. (1972). Some relationships between body motion and speech. *Studies in dyadic communication*, 7:177.

Kendon, A. (1983). Gesture and speech: How they interact. *Nonverbal interaction*, 11:13–45.

Kendon, A. (1994). Do gestures communicate? A review. *Res. Lang. Soc. Interact.*

Kim, T., Yue, Y., Taylor, S., and Matthews, I. (2015). A decision tree framework for spatiotemporal sequence prediction. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 577–586. ACM.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *international conference on learning representations*.

Klinger, E., Bouchard, S., Légeron, P., Roy, S., Lauer, F., Chemin, I., and Nugues, P. (2005). Virtual reality therapy versus cognitive behavior therapy for social phobia: A preliminary controlled study. *Cyberpsychology & behavior*, 8(1):76–88.

Konda, K., Bouthillier, X., Memisevic, R., and Vincent, P. (2015). Dropout as data augmentation. *stat*, 1050:29.

Kunin, M., Osaki, Y., Cohen, B., and Raphan, T. (2007). Rotation axes of the head during positioning, head shaking, and locomotion. *Journal of Neurophysiology*, 98(5):3095–3108.

Kuratate, T., Munhall, K. G., Rubin, P., Vatikiotis-Bateson, E., and Yehia, H. (1999). Audio-visual synthesis of talking faces from speech production correlates. In *EuroSpeech*.

Kuratate, T., Yehia, H., and Vatikiotis-Bateson, E. (1998). Kinematics-based synthesis of realistic talking faces. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*.

Laine, S., Karras, T., Aila, T., Herva, A., Saito, S., Yu, R., Li, H., and Lehtinen, J. (2017). Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, page 10. ACM.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lewis, J. (1991). Automated lip-sync: Background and techniques. *Computer Animation and Virtual Worlds*, 2(4):118–122.

Lewis, J. P. and Parke, F. I. (1986). Automated lip-synch and speech synthesis for character animation. *SIGCHI Bull.*, 17(SI):143–147.

Li, B., Xie, L., Zhu, P., and Fan, B. (2013). Head motion generation for speech-driven talking avatar. *Journal of Tsinghua University (Science and Technology)*, 6:035.

Li, X., Wu, Z., Meng, H. M., Jia, J., Lou, X., and Cai, L. (2016). Expressive speech driven talking avatar synthesis with dblstm using limited amount of emotional bimodal data. In *INTERSPEECH*, pages 1477–1481.

Linde, Y., Buzo, A., and Gray, R. M. (1980). An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28(1):84–95.

Lisetti, C. L. (2008). Embodied conversational agents for psychotherapy. In *Proceedings of the CHI 2008 conference workshop on technology in mental health*, pages 1–12.

Logan, B. et al. (2000). Mel frequency cepstral coefficients for music modeling. In *ISMIR*.

Maatman, R., Gratch, J., and Marsella, S. (2005). Natural behavior of a listening agent. In *International Workshop on Intelligent Virtual Agents*, pages 25–36. Springer.

Macmillan, N. A. and Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.

Markel, J. E. and Gray, A. (1982). Linear prediction of speech.

Matthews, I. and Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164.

Mattheyses, W. and Verhelst, W. (2015). Audiovisual speech synthesis: An overview of the state-of-the-art. *Speech Communication*, 66:182–217.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (2017). Montreal forced aligner: trainable text-speech alignment using kaldi.

McCarthy, M. (2003). Talking back:" small" interactional response tokens in everyday conversation. *Research on language and social interaction*, 36(1):33–63.

McCullough, K. (1992). Visual imagery in language and gesture. In *Annual Meeting of the Belgian Linguistic Society*. Brussels, Belgium.

McDonagh, S., Klaudiny, M., Bradley, D., Beeler, T., Matthews, I., and Mitchell, K. (2016). Synthetic prior design for real-time face tracking. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 639–648. IEEE.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* University of Chicago Press.

Melis, G., Dyer, C., and Blunsom, P. (2018). On the state of the art of evaluation in neural language models. *international conference on learning representations.*

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Moore, R. K. (2012). A bayesian explanation of the 'uncanny valley'effect and related psychological phenomena. *Scientific reports*, 2.

Morency, L.-P., de Kok, I., and Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. In *Intelligent Virtual Agents*, pages 176–190. Springer.

Mori, M. (1970). The uncanny valley. *Energy*, 7(4):33–35.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: head movement improves auditory speech perception. *Psychological science : A journal of the American Psychological Society / APS*, 15(2):133–137.

Natrella, M. (2010). *NIST/SEMATECH e-Handbook of Statistical Methods.* NIST/SEMATECH.

Noll, A. M. (1967). Cepstrum pitch determination. *The journal of the acoustical society of America*, 41(2):293–309.

Ochshorn, R. M. and Hawkins, M. (2017). Gentle forced aligner. Available at `https://github.com/lowerquality/gentle`.

of America Standards Institute, A. N. S. I. U. S. ([1973]). *American national standard psychoacoustical terminology*. New York : The Institute, [1973] ©1973.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society.

Prechelt, L. (1998). Early stopping-but when? *neural information processing systems*, pages 55–69.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Ramakrishna, V., Kanade, T., and Sheikh, Y. (2012). Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision*, pages 573–586. Springer.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1278–1286.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, CORNELL AERONAUTICAL LAB INC BUFFALO NY.

Sadoughi, N. and Busso, C. (2017). Joint learning of speech-driven facial motion with bidirectional long-short term memory. In Beskow, J., Peters, C., Castellano, G., O'Sullivan, C., Leite, I., and Kopp, S., editors, *Intelligent Virtual Agents*, pages 389–402, Cham. Springer International Publishing.

Sainath, T. N., Kingsbury, B., Mohamed, A.-r., Dahl, G. E., Saon, G., Soltau, H., Beran, T., Aravkin, A. Y., and Ramabhadran, B. (2013). Improvements to deep convolutional neural networks for lvcsr. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 315–320. IEEE.

Sainath, T. N., Kingsbury, B., Mohamed, A.-r., Saon, G., and Ramabhadran, B. (2014). Improvements to filterbank and delta learning within a deep neural network framework. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6839–6843. IEEE.

Saygin, A. P., Chaminade, T., and Ishiguro, H. (2010). The perception of humans and robots: Uncanny hills in parietal cortex. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 2716–2720.

Schröder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., ter Maat, M., McKeown, G., Pammi, S., Pantic, M., Pelachaud, C., Schuller, B. W., de Sevin, E.,

Valstar, M. F., and Wöllmer, M. (2015). Building autonomous sensitive artificial listeners (extended abstract). In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 456–462.

Schroeder, M. R. (1968). Period histogram and product spectrum: New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America*, 43(4):829–834.

Schüssler, H. W. (1976). A stability theorem for discrete systems. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(1):87–89.

Schuster, A. (1898). On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Journal of Geophysical Research*, 3(1):13–41.

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. D. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Simons, N., Krämer, N. C., and Kopp, S. (2007). The effects of an embodied agent's nonverbal behavior on user's evaluation and behavioral mimicry. pages 1–41.

Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 3483–3491. Curran Associates, Inc.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. 15:1929–1958.

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13.

Taylor, S., Kim, T., Yue, Y., Mahler, M., Krahe, J., Rodriguez, A. G., Hodgins, J., and Matthews, I. (2017). A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):93.

Taylor, S. L., Mahler, M., Theobald, B.-J., and Matthews, I. (2012). Dynamic units of visual speech. In *Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation*, pages 275–284. Eurographics Association.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2).

Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, 1(1):19–22.

Tobias, J. (2012). *Foundations of modern auditory theory.* Elsevier.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Zafeiriou, S., et al. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204. IEEE.

Trutoiu, L. C., Carter, E. J., Matthews, I., and Hodgins, J. K. (2011). Modeling and animating eye blinks. *ACM Transactions on Applied Perception*, 8(3):1–17.

van der Maaten, L. and Hendriks, E. (2012). Action unit classification using active appearance models and conditional random fields. *Cognitive processing*, 13(2):507–518.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Walker, J., Doersch, C., Gupta, A., and Hebert, M. (2016). An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pages 835–851. Springer.

Wang, L., Han, W., Soong, F. K., and Huo, Q. (2011). Text driven 3d photo-realistic talking head. In *Twelfth Annual Conference of the International Speech Communication Association.*

Ward, N. and Tsukahara, W. (2000). Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207.

Wickens, T. D. (2002). *Elementary signal detection theory.* Oxford University Press, USA.

Yehia, H., Kuratate, T., and Vatikiotis-Bateson, E. (2000). Facial animation and head motion driven by speech acoustics. In *5th Seminar on Speech Production: Models and Data*, pages 265–268. Kloster Seeon, Germany.

Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting*, pages 567–578.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (1997). *The HTK book*, volume 2. Entropic Cambridge Research Laboratory Cambridge.

Zen, H., Tokuda, K., and Kitamura, T. (2007). Reformulating the hmm as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech & Language*, 21(1):153–173.

Zhang, Z. (1999). Flexible camera calibration by viewing a plane from unknown orientations. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 00(c):666–673 vol.1.