Lip Synchronization by Acoustic Inversion

Citation for published version:

Digital Object Identifier (DOI):
10.1145/1836845.1836857

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
ACM SIGGRAPH 2010 Posters

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Lip Synchronization by Acoustic Inversion

Gregor Hofer*
University of Edinburgh

Korin Richmond†
University of Edinburgh

Michael Berger‡
University of Edinburgh

Figure 1: The word "how" as generated from our model (left) and the rendered equivalent sample frames (right). The red contours indicate the final frame, with frames going back in time represented with progressively lighter grayscale contours.

Keywords: lip synchronization, facial animation, neural networks

1 Introduction

Talking computer animated characters are a common sight in video games and movies. Although doing the mouth animation by hand gives the best results it is not always feasible because of cost or time constraints. Therefore producing lip animation automatically is highly desirable. The problem can therefore be phrased as mapping from speech to lip animation or in other words as an acoustic inversion. In our work we propose a solution that takes a sequence of input frames of speech and maps it directly to an output sequence of animation frames. The key point is that there is no need for phonemes or visemes which cuts one step in the usual lip synchronization process.

2 Speech to motion mapping method

At the heart of our mapping model is the mixture density network (MDN). The MDN can be considered as combining a trainable regression function (typically a non-linear regressor such as an artificial neural network) with a probability density function. In our work, we have been using a multilayer perceptron (MLP) as a trainable non-linear regressor and a Gaussian mixture model (GMM). An illustration is shown in Fig. 2. The role of the MLP is to take an input vector in one domain ($x$, acoustic features in this case) and map to the control parameters (priors, means and variances) of the pdf over the domain of the target parameters ($t$, motion features). In this way, the MDN offers a model of probability density over the target domain conditioned on the input domain, $p(t|x)$. Once trained, we input a sequence of acoustic feature vectors for an utterance and get as output a sequence of pdfs over the static motion features and their delta and deltadeltas. Trajectories can be generated in real time from the speech frames using a small amount of context frames. The model was trained on the first 4 PCA components of 28 tracked points around the mouth. The total amount of training data was 207 utterances ($\sim$ 25 min.). The input speech features were the first 25 mel cepstrum coefficients, which are standard speech recognition features. A short perceptual evaluation was carried out with 17 subjects judging 5 utterances. No significant difference between the preference for animation from the original data (46%) and animation from the model (54%) was found, which is a promising result to further develop MDNs for lip synchronization.

References


* e-mail: ghofer@inf.ed.ac.uk
† e-mail: korin@cstr.ed.ac.uk
‡ e-mail: m.a.berger@sms.ed.ac.uk