

Audiovisual speech synthesis: An overview of the state-of-the-art

Wesley Mattheyses^{a,*}, Werner Verhelst^{a,b}

^a *Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium*

^b *iMinds, Gaston Crommenlaan 8, Box 102, 9050 Ghent, Belgium*

Received 21 February 2014; received in revised form 4 November 2014; accepted 4 November 2014

Available online 11 November 2014

Abstract

We live in a world where there are countless interactions with computer systems in every-day situations. In the most ideal case, this interaction feels as familiar and as natural as the communication we experience with other humans. To this end, an ideal means of communication between a user and a computer system consists of audiovisual speech signals. Audiovisual text-to-speech technology allows the computer system to utter any spoken message towards its users. Over the last decades, a wide range of techniques for performing audiovisual speech synthesis has been developed. This paper gives a comprehensive overview on these approaches using a categorization of the systems based on multiple important aspects that determine the properties of the synthesized speech signals. The paper makes a clear distinction between the techniques that are used to model the virtual speaker and the techniques that are used to generate the appropriate speech gestures. In addition, the paper discusses the evaluation of audiovisual speech synthesizers, it elaborates on the hardware requirements for performing visual speech synthesis and it describes some important future directions that should stimulate the use of audiovisual speech synthesis technology in real-life applications.

© 2014 Elsevier B.V. All rights reserved.

Keywords: Audiovisual speech synthesis; Visual speech synthesis; Speech synthesis

Contents

1. Introduction to audiovisual speech synthesis	183
1.1. Motivation for speech synthesis	183
1.2. The multimodality of speech	184
1.3. Synthetic audiovisual speech for human–machine interaction	184
1.4. The generation of synthetic speech	186
2. The synthesis of audiovisual speech	186
3. Input requirements	187
3.1. Text-driven systems	187
3.2. Speech-driven systems	187
3.3. Speaker cloning	188
4. Output modality	188
4.1. Single-phase or two-phase synthesis	188
4.2. Literature overview	188

* Corresponding author.

E-mail addresses: wmatthey@etro.vub.ac.be (W. Mattheyses), wverhels@etro.vub.ac.be (W. Verhelst).

5.	Output dimensions	189
5.1.	Synthesis in 3D	189
5.2.	Synthesis in 2D	190
5.3.	Other approaches	190
6.	Static photorealism	190
6.1.	Photorealism in 3D synthesis	190
6.2.	Photorealism in 2D synthesis	191
7.	Definition of the visual articulators and their variations	192
7.1.	3D speaker modelling	192
7.1.1.	Terminal-analog systems	192
7.1.2.	Anatomy-based systems	192
7.1.3.	Performance-driven animation	193
7.2.	2D speaker modelling	194
7.2.1.	Modelling using photographs	194
7.2.2.	Modelling using video sequences	195
7.2.3.	2D image parameterization	195
7.2.4.	2D graphical modelling	195
7.3.	Standardization: FACS and MPEG-4	195
7.3.1.	Facial Action Coding System	196
7.3.2.	MPEG-4	196
8.	Prediction of the target speech gestures	197
8.1.	Coarticulation	197
8.2.	Rule-based synthesis	198
8.2.1.	3D rule-based synthesis	199
8.2.2.	2D rule-based synthesis	201
8.2.3.	Other rule-based synthesizers	201
8.3.	Concatenative synthesis	201
8.3.1.	3D concatenative synthesis	202
8.3.2.	2D concatenative synthesis	203
8.4.	Synthesis based on statistical prediction	204
9.	Summary	205
9.1.	Input requirements	206
9.2.	Output modality	206
9.3.	Output dimensions	206
9.4.	Static photorealism	206
9.5.	Definition of the visual articulators and their variations	206
9.6.	Prediction of the target speech gestures	206
10.	Evaluating the quality of a visual speech synthesizer	206
10.1.	Definition of speech quality	206
10.2.	Quality measures	207
10.2.1.	Subjective measures	207
10.2.2.	Perceptual measures	207
10.2.3.	Objective measures	207
11.	Hardware consequences	208
11.1.	3D-based synthesis	208
11.2.	2D-based synthesis	209
11.3.	Today's issues	209
12.	Future directions	209
12.1.	Audiovisual coherence	210
12.2.	Adding emotions and expressions	210
12.3.	Standardisation	211
	References	212

1. Introduction to audiovisual speech synthesis

1.1. Motivation for speech synthesis

At the present time, cars, heavy machinery, vending machines, medical devices, and even the simplest home

appliances such as fridges and central heating systems are computer controlled. For every such device a human–machine interface is needed for enabling users to control the machine and to make it possible for the device to give feedback to its users. The ultimate goal should be that the computer systems that surround us are perfectly integrated

Download English Version:

<https://daneshyari.com/en/article/6961208>

Download Persian Version:

<https://daneshyari.com/article/6961208>

[Daneshyari.com](https://daneshyari.com)