An Anthropomorphic Perspective for Audiovisual Speech Synthesis

Samuel Silva and António Teixeira
DETI — Dep. of Electronics, Telecommunications and Informatics, University of Aveiro, Aveiro, Portugal
IEETA — Institute of Electronics Engineering and Informatics of Aveiro, University of Aveiro, Aveiro, Portugal

Keywords: Audiovisual Speech, Articulatory Synthesis, European Portuguese.

Abstract: In speech communication, both the auditory and visual streams play an important role, ensuring both a certain level of redundancy (e.g., lip movement) and transmission of complementary information (e.g., to emphasize a word). The common current approach to audiovisual speech synthesis, generally based on data-driven methods, yields good results, but relies on models controlled by parameters that do not relate with how humans do it, being hard to interpret and adding little to our understanding of the human speech production apparatus. Modelling the actual system, adopting an anthropomorphic perspective would provide a myriad of novel research paths. This article proposes a conceptual framework to support research and development of an articulatory-based audiovisual speech synthesis system. The core idea is that the speech production system is modelled to produce articulatory parameters with anthropomorphic meaning (e.g., lip opening) driving the synthesis of both the auditory and visual streams. A first instantiation of the framework for European Portuguese illustrates its viability and constitutes an important tool for research in speech production and the deployment of audiovisual speech synthesis in multimodal interaction scenarios, of the utmost relevance for the current and future complex services and applications.

1 INTRODUCTION

Among humans, the most natural form of communication is speech, which makes it a promising modality to foster a natural and transparent interaction with a wide range of devices, from smartphones to intelligent homes. As a result, speech is becoming a common option as an input and output modality. Speech, as a form of communication among humans is not limited to the auditory signal. In fact, the visual features of speech production, such as the movement of the lips, play an important role in adding to the auditory signal, whether by providing redundant information or by complementing it (e.g., conveying a particular emphasis to particular words). Therefore, output that jointly explores both the auditory and visual streams of speech, known as audiovisual speech synthesis, can bring a wide range of advantages over auditory-only speech in providing a richer and more resilient method of communication and adding a more natural feel to human-machine interaction.

Current technologies allow collecting large amounts of data regarding particular biomechanical phenomena (e.g., cerebral, muscular, and facial activity) and a common approach is to use that data to train models that enable its use for particular interaction tasks such as converting a certain myoelectric signal pattern in a particular action. Although this is, indisputably, a very important advance and, in many aspects, the only possible approach considering our knowledge of the involved physiology and the current stage of technology, training a model commonly results in a 'black box' converting input data into specific outputs. This often means that the parameters controlling the model are not actually related with anatomical or physiological elements and hard to conceptualize. Therefore, these models serve their purpose, but provide a limited margin for customization and a minor contribution to our understanding of the phenomena, which continues to be very partial and limited. However, when technology allows a deeper insight into the biomechanics of a phenomena we can move from 'blindly' modelling the transformation of inputs into outputs to modelling the actual system and its components, adopting a set of more anthropomorphic parameters and, in this manner, learn about the phenomena and how to replicate it.

While the literature is prolific in providing several methods for audiovisual synthesis (Schabus et al., 2014), yielding good overall results, these often rely on data-driven/statistical approaches, i.e., the collec-

Silva S. and Teixeira A.
An Anthropomorphic Perspective for Audiovisual Speech Synthesis.
DOI: 10.5220/0006150201630172
In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*, pages 163-172
ISBN: 978-989-758-212-7

tion of big data sets to train a model or provide the grounds for concatenation-based systems. The result is that we end up with systems that: (1) require the collection of large amounts of data; (2) are strongly related with the data they are built from; and (3) do not relate directly with the production system, i.e., the vocal tract and facial movements.

Regarding speech production, technological advances enable studying the static and dynamic aspects of the vocal tract providing enough information to support building first models that can be used to study speech production and enable advancing articulatory speech synthesis (e.g., Birkholz, 2013; Rubin et al., 1981; Teixeira et al., 2008), i.e., speech synthesized by replicating the articulatory patterns performed by humans. The advantage of such a bioinspired system is that it depends on anthropomorphic parameters we can understand, and can be used to assess theoretical aspects, including the importance of particular vocal tract configurations or timings in the production of specific sounds, while also providing a speech synthesis system. Likewise, we consider that an articulatory-based approach to audiovisual speech synthesis, where we build a model replicating the audiovisual speech production system, could provide several advantages, in two domains: (a) it can be used as a highly customizable tool supporting research (Files et al., 2015), enabling experimenting with the different articulators and timings to understand how speech is produced, how the auditory and visual streams relate, and serving educational and clinical applications (Massaro, 2005); (b) it can provide a more versatile system for deploying richer outputs in interaction scenarios. For example, since it does not depend on the collection of novel data, in an articulatory approach the voice can be more easily customized, the visual part completely changed (ranging from a cartoon-like character to a realistically looking avatar), and articulatory limitations introduced (for instance, limit the amplitude of the jaw movement to simulate someone speaking with its jaw movement hindered).

Following this line of thought, we present the conceptualization and first instantiation of a framework to support articulatory-based audiovisual speech synthesis oriented to support both the research on the audiovisual aspects of speech production and its deployment as an interaction modality.

The remainder of this article is organized as follows: section 2 briefly reviews current trends in audiovisual speech synthesis; section 3 discusses the main motivations for pursuing an articulatory approach to audiovisual speech, in the context provided by the state-of-the-art, and presents the basic concepts behind articulatory synthesis; our conceptual view on how to pursue articulatory audiovisual speech synthesis is described in section 4 and section 5 describes a first implementation of the described framework; finally, section 6 presents conclusions and major routes for future work.

2 CURRENT TRENDS IN AUDIOVISUAL SPEECH SYNTHESIS

A wide range of approaches to audiovisual speech synthesis have been presented in the literature. In a recent review, (Mattheyses and Verhelst, 2015) highlight four aspects defining an audiovisual speech synthesizer: 1) the properties of the input information (e.g., text or audio); 2) the properties of the output (e.g., in 2D or 3D); 3) how the visual articulators are defined (e.g., 3D modelling) and animated (e.g., anatomy-based or performance-driven systems (Železný et al., 2015)); and 4) how the different visual configurations that need to be attained are predicted (e.g., rule-based, concatenative).

When performing visual-only speech synthesis, a common approach is to use visemes, a visual representation of the relevant articulators, to represent phonemes. Whenever a certain phoneme occurs, the corresponding viseme is used. Since the visible parts of the vocal tract (e.g., lips) adopt similar configurations for different sounds, and modelling visemes may require laborious time and a digital artist (Serra et al., 2012), it is common to use the same viseme for multiple phonemes. Then, these visemes work as keyframes and are animated by interpolating in-between. In many cases, even though the proposed systems provide a visual speech synthesis that is smooth, it does not have any underlying mechanism to actually implement visual coarticulation. Many phonemes to one viseme mappings, for example, do not account for visual coarticulation (Mattheyses and Verhelst, 2015) and alternatives to minimize this issue have been proposed by adding visemes for diphones and triphones).

Additionally, several authors have addressed coarticulation by proposing models that define how the interpolation between visemes is performed considering, for example, models of facial biomechanics or through a model for visual coarticulation in which interpolation between visemes depends on weights associated with the corresponding phonemes, according to their dominance (e.g., Cohen and Massaro, 1993).

In recent years, data-driven/statistical approaches

have taken the lead (e.g., Schabus et al., 2014). In concatenative visual synthesis, a data collection stage is performed considering a corpus that should be phonetically rich enough to include all the required phones/segments and relevant contexts. So, when synthesizing, the system retrieves the longest possible segments from the database, to minimize the number of concatenations, and synthesizes the desired audiovisual speech. In the case of visual speech based on 3D models a performance-driven animation is required entailing data collections that involve complex settings.

For the Portuguese language, just a few audiovisual synthesis systems have been proposed. In a notable example, Serra et al. (Serra et al., 2012) propose an audiovisual speech synthesis system for European Portuguese (EP) enabling automatic visual speech animation from text or speech audio input. The authors consider a phoneme-to-viseme strategy for animating a 3D model and perform a preliminary evaluation of two different sets of visemes reaching the conclusion that using a small number of visemes (the same viseme is attributed to more phonemes) has a negative impact on the perceived quality of the output.

3 ARTICULATORY SYNTHESIS

In this section, we present the basic concepts providing the grounds for articulatory synthesis and, based on the current state-of-the-art, discuss the main motivations to enrol on an articulatory approach to audiovisual speech synthesis.

3.1 Articulatory Phonology

In articulatory phonology (Browman and Goldstein, 1990; Hall, 2010), the basic unit of speech is not the segment or the feature, but the articulatory gesture. Gestures are a set of instructions that define how a constriction is formed and released in the vocal tract, such as lip closure.

In articulatory phonology, the vocal tract configuration is generally defined by considering five tract variables: lips (LIP), tongue tip (TT), tongue body (TB), velum (VEL) and glottis (GLO). Gestures are specified based on these variables and the constrictions' location (CL) – labial, dental, alveolar, postalveolar, palatal, velar, uvular and pharyngeal – and degree (CD) – closed (for stops), critical (for fricatives) and narrow, mid, and wide (for approximants and vowels). In this context, the gestural specification for the alveolar stop [t], for example, would be: tract variable tongue tip, CD: closed, CL: alveolar. This defines the goal of the gesture: the target.

A gesture is a dynamic action characterized by several phases: the onset of the movement; its progress until reaching the target; release, when it starts moving away from the constriction; and offset, after which the articulator is no longer under the control of the gesture.

Gestures are combined to form larger elements (e.g., syllables and words). This combination is not a simple matter of their sequence in time. Gestures blend with other gestures according to phasing principles: a certain point in the trajectory of one gesture is phased with respect to the trajectory of other gestures. The specification of the different gestures involved in articulating a particular word, along with the time intervals defining the regions of active control for each gesture, is called a gestural score. Given that gestures need to be combined, a gestural score is just the first step for computing articulator's trajectories, which will result from its interpretation and modification.

One Matlab implementation of the articulatory phonology framework described above, named TAsk Dynamics Application (TADA), has been proposed by the Haskins Laboratories (Saltzman and Munhall, 1989; Nam et al., 2006). A gestural model considers the syllables of the input text to generate the gestural score. This includes the specification of the various gestures required and their activation intervals based on models for inter-gestural planning and gestural-coupling. Gestural scores are considered by the task dynamic model to generate the final time functions for articulator trajectories considering the articulators of the CASY vocal tract model (Rubin et al., 1996). These trajectories are considered to configure the vocal tract model and provide the basis to compute the acoustic output (synthetic speech).

3.2 Motivation

Overall, at the current stage of technology, the quality provided by data-driven/statistical approaches to audiovisual speech synthesis surpasses the quality possible with articulatory approaches. However, an articulatory approach serves a wider set of goals beyond the production of the final audiovisual speech output.

The articulatory based approach to speech synthesis starts by not requiring the collection of large datasets for training. Furthermore, in data-driven/statistical systems the synthesized speech depends on the nature of the collected data, e.g., age and gender of the speaker, and creating a new voice involves almost the same amount of effort and re-

sources as the first. An articulatory synthesizer provides enough versatility to be customized to synthesize voices with a wide range of characteristics such as a specific gender or speech disorder. Therefore, while a significant effort might be required to attain a first good quality audiovisual speech synthesis system, new voices will not involve as much effort.

The parameters involved in an articulatory-based speech synthesizer are more intuitive than those controlling a data-driven model since they relate with actual elements of the real system. This means that when a particular synthesis goes wrong one can intuitively make sense of what might have happened or where to perform changes. And this ability to experiment with the different parameters, in search for perceptually better outputs, also contributes to improve our knowledge about speech production.

An approach that favours a biomechanical perspective of the system can also be valuable for educational and clinical purposes. The audiovisual speech synthesis system can be used as an assistant tool for teaching phonetics and therapy, providing an illustration of the different aspects of speech production.

The consideration of visemes for audiovisual speech synthesis, although quite common and providing reasonable results, raises several issues. A particular viseme assumes that all articulators are under the influence of a phone, which is not true, since a phone might be defined by a single gesture (for example, [b]) without influencing all articulators. Even considering a weighted approach, to explicitly define the importance of each viseme, does not address this aspect. Additionally, the use of audiovisual concatenative synthesis for clinical purposes (e.g., speech therapy), to provide someone with an illustration of how to properly articulate a particular token, raises concerns as to whether we are actually providing a proper reference to the patient. An articulatory-based audiovisual synthesis approach would provide more solid ground over which these different aspects can be tackled.

Additionally, since the articulatory-based system models the different articulators and their dynamic behaviour, one can envisage that if it is associated with a technology, such as ultrasound imaging, it might provide instant assessment of speaker performance, e.g., during speech therapy exercises, by comparing speaker performance with the computed trajectories, and provide specific clues on what to improve or support the choice of particular exercises to surpass those difficulties.

Overall, we consider that these aspects support investing in audiovisual speech synthesis approach from an articulatory perspective, favouring further study and understanding of human anatomy, physiology and behaviour, in this context.

4 CONCEPTUAL FRAMEWORK FOR ARTICULATORY AUDIOVISUAL SPEECH SYNTHESIS

In line with the current state-of-the-art and the different motivations presented, establishing the relevance of an articulatory approach to audiovisual speech synthesis, we start by characterizing the envisaged scenarios and establish a set of requirements leading to the conceptualization of a framework supporting R&D in articulatory audiovisual speech synthesis.

4.1 Scenarios and Requirements

We envisage two distinct scenarios for exploring articulatory audiovisual speech synthesis that profit from each other: (1) improve our knowledge of speech production and study the synergies between the audio and visual aspects of speech; and (2) use audiovisual speech synthesis as a rich output modality in multimodal interaction.

4.1.1 Scenario 1: Research in Speech Production

A set of technologies (e.g., Scott et al., 2014) and methods (e.g., Silva et al., 2015) provide an increasing amount of data to characterize different aspects of speech production, regarding, e.g., the configuration of the vocal tract for producing particular sounds, coarticulation and synchronization. The collected data allows a better understanding of the static and dynamic aspects of speech production, but its value can be further established if, based on this data, we are able to postulate rules or even theories that can be tested (simulated) in an experimental setting. Only then we will have confirmation that we fully characterized and understand the observed phenomena and it might potentially trigger our awareness of further aspects that need clarification. Naturally, this sort of research, as previously emphasized, requires modelling the articulatory system and not just its outputs. This rationale also applies to the joint consideration of auditory and visual speech for which an articulatory-based audiovisual speech synthesis system would be an important tool.

Overall, this scenario requires: (1) a system that is able to move from a text input to an audiovisual speech output, modelling all the different stages;

(2) the definition of each language specificity, regarding particular sounds and their articulation; (3) a face model controlled by anthopomorphic parameters, which are easily related with our understanding of a real speaker; and (4) constant evaluation of results to assess output (perceptual) quality and guide future developments.

4.1.2 Scenario 2: Multimodal Interaction

The research in audiovisual speech production should also enable the deployment of a system that can be used as an interaction modality. Research in multimodal interaction envisages complex contexts such as ambient assisted living (AAL) where a heterogeneous set of users experiences with a multitude of interactive devices (Almeida et al., 2016; Teixeira et al., 2016). One concrete example is provided by the ongoing Marie Curie IAPP project IRIS (Freitas et al., 2014) focusing on enhancing communication in a household for a family spanning a wide range of skills, limitations and motivations. As with text-to-speech (Almeida et al., 2014), enabling a customization of the voice and visual appearance of the avatar, eventually personalized to a particular context or user, might also provide improved acceptability. On the subject of avatar versatility, having a 3D model covers a broader range of scenarios, from simple 2D front view up to its insertion in a virtual environment.

Additionally, although the research focus in speech production might favour a particular language, for interaction we are also interested in enabling articulatory audiovisual speech synthesis in multiple languages. This is in line with recent efforts in the literature providing a generic speech modality encapsulating support for multiple languages (Almeida et al., 2014). We also argue that one additional motivation for evolving an articulatory audiovisual speech synthesizer might come from specific requirements deriving from its use in real interaction scenarios. This is only possible if developers are able to integrate it in their applications, as made possible, e.g., by a multimodal interaction architecture and framework (e.g., Teixeira et al., 2016) in which the audiovisual speech synthesis modality should be integrated.

This scenario requires, overall: (1) ability to easily move from our research in audiovisual speech production into its deployment as an interaction modality; (2) methods that exhibit moderate computational costs and use technologies that are compatible with the large variety of devices envisaged; (3) language specific modules sufficiently decoupled and interchangeable to enable support for additional languages; and (4) customizable visual stream, to support personalized 3D avatars.

4.2 Conceptual Framework

Considering the authors' experience in articulatory-based speech synthesis and the motivations previously presented, the corner stone of our proposal is that articulatory-based audiovisual speech synthesis can inherit from the articulatory speech synthesis approach and be driven by the same anthropomorphic parameters.

In what follows, we present the main conceptual aspects of a framework to support R&D in articulatory audiovisual speech synthesis serving the envisaged scenarios and requirements. Our view considers three main blocks: an audiovisual speech synthesis core supported by an evaluation module and integrated as an interaction modality.

Figure 1 depicts these three blocks and puts into evidence the synergies that should exist among them, contributing for advancing the usefulness and quality of articulatory audiovisual speech synthesis, as detailed in what follows.

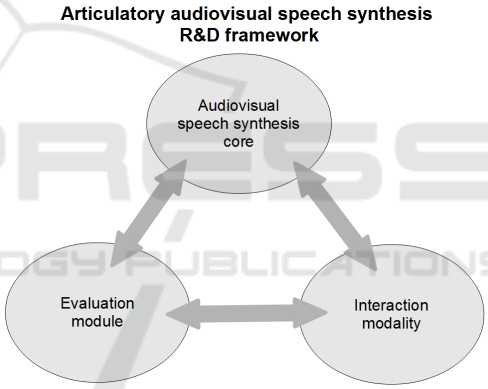

Figure 1: Overview of the conceptual framework for performing research and development in articulatory-based audiovisual speech synthesis.

4.2.1 Audiovisual Synthesis Core

The audiovisual synthesis core should constitute the central element of the framework, dealing with the actual production of the audiovisual output. Figure 2 depicts the main features of the envisaged text-driven synthesis approach.

The language package deals with all aspects that are specific in each language such as the gestural model. The vocal tract static and dynamic modelling deals with moving from the text and language specific data to the definition of the different gestures and their interactions and, finally, to the physiological (anthropological) parameters. These parameters drive a vocal tract model providing the geometry for the speech synthesizer.

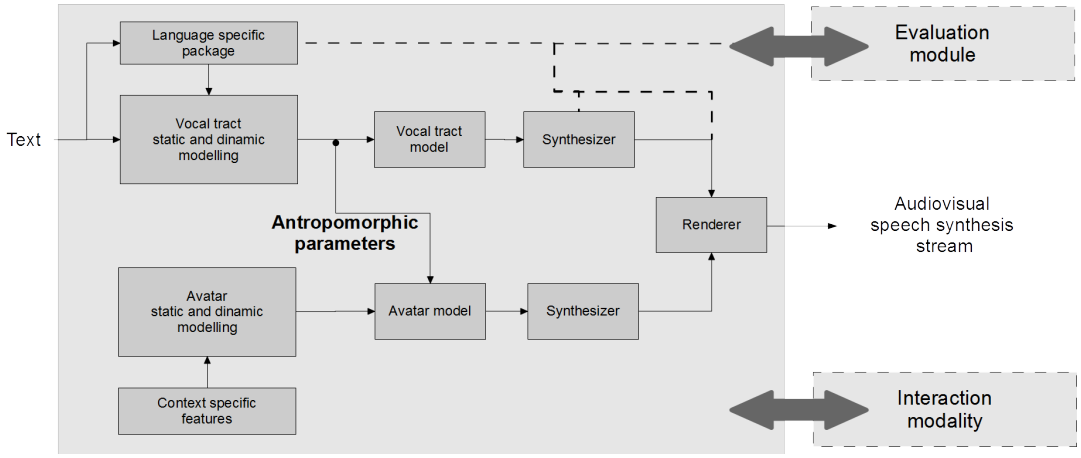

Figure 2: Main aspects of the articulatory based audiovisual speech synthesis core. The highlight is a set of anthropomorphic parameters driving both the auditory and visual synthesis.

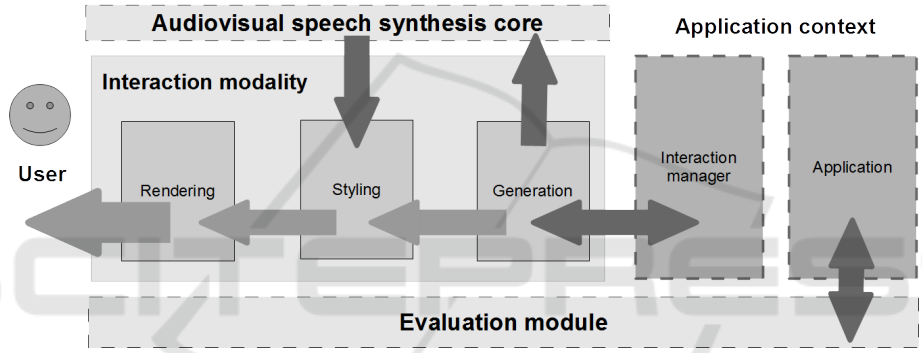

Figure 3: Diagram depicting the main aspects of the conceptual integration between the audiovisual speech synthesis core and an interaction modality made available in a multimodal interaction context.

On the visual synthesis side, context specific requirements, such as specific avatar look (e.g., personalized to a particular user or a character in a film) are provided and a custom avatar model is created supporting a set of controls for animation. One of the notable aspects of this conceptual framework, worth emphasizing, is that we argue that the anthropological parameters (e.g., lip aperture and protrusion), which drive the auditory synthesis, should also serve to configure the model used for visual synthesis. Finally, both streams are rendered to provide audiovisual speech synthesis.

The dashed lines, in figure 2, coming from the evaluation module, are just to emphasize that this module, provided certain assessment data, is envisaged as being able to exert its influence at different levels in the audiovisual synthesizer as further explained ahead.

4.2.2 Interaction Modality

The second aspect we consider relevant is the integration of the audiovisual speech synthesis system into an interaction modality, deployed as part of a multimodal interaction framework. We argue that, without this context of use, research in articulatory audiovisual speech synthesis will fall short of its true potential due to a lack of real application scenarios.

The main aspects regarding this element of the conceptual framework are depicted in figure 3. The main components presented for the interaction modality (i.e., generation, styling and render) derive from the conceptual view of an output component proposed by the W3C for a multimodal interaction architecture (W3C Consortium, 2003).

Articulatory audiovisual synthesis entails a considerable level of complexity, particularly regarding the computation of the articulatory parameters driving the models. Therefore, since the resulting interaction modality can be added to applications running on any kind of device, the core computations, should

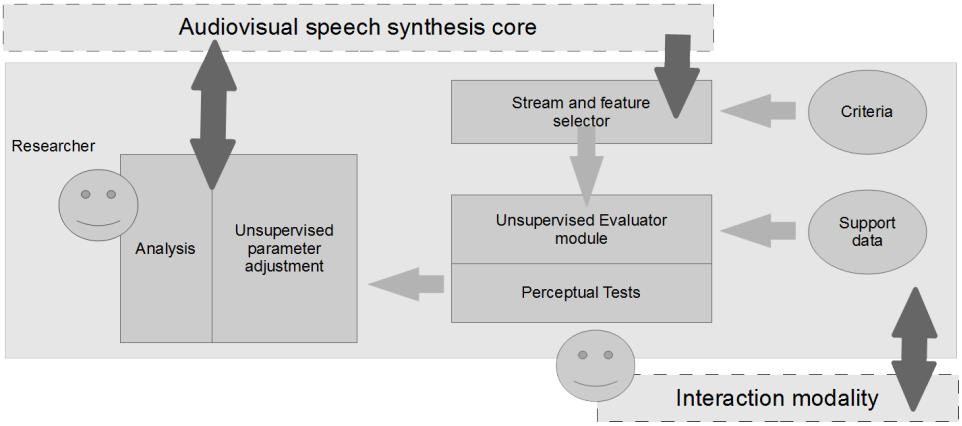

Figure 4: Main features of the conceptualized evaluation module, part of the proposed audiovisual speech synthesis conceptual framework.

occur outside of the device, if needed. This is also aligned with another important requirement. If we want to rapidly move from the research outcomes into the interaction scenario, placing most of the core online, e.g., as a webservice, enables fast deployment of new versions without the need to update the installed applications. The generation module, depicted in figure 3, would be responsible for managing the interface with the online service, providing the input text and requirements. However, it is important to note that, to preserve some versatility on the modality side might entail leaving part of the conceptualized audiovisual speech synthesis core locally, particularly the decisions about how to style and render features. These aspects explain why the interaction modality, although providing an encapsulation of the features provided by the audiovisual synthesis core, is not conceptualized as containing the whole system.

4.2.3 Evaluation

To validate the outcomes of the produced audiovisual speech synthesis and to provide feedback that guides research, evaluation is an important part of the envisaged framework and the overall aspects of the evaluation module are depicted in figure 4. In this regard, perceptual tests are one of the most used options in this domain, providing researchers with an assessment of the adequacy of the outputs regarding, for example, the synthesis of specific sounds or the synchronization between audio and visual synthesis. Nevertheless, even though perceptual tests remain the most important source of 'gold standards', the systematic development of an audiovisual speech synthesis system requires additional evaluation methods enabling a faster and more frequent assessment of different alternatives, additions, and improvements. For example, an unsupervised method that could provide a first evaluation, to guide system evolution before the much more costly perceptual tests, would be a valuable asset. The stream to evaluate, at each time (i.e., auditory, visual or both), the focused parameters, and the text to synthesize should be defined by a set of criteria that can include, e.g., a specific corpora. The unsupervised evaluation might be based on features computed from the considered streams or might require support data to serve as reference, e.g., features extracted from video streams of real speakers.

The perceptual tests can also profit from the deployment of the audiovisual system as an interaction modality. Adding to the more controlled lab settings, data can also be collected from who is using the interaction modality, potentially covering different contexts, and a vital step, we argue, in assessing the acceptability and usefulness of the audiovisual synthesis in real scenarios.

Another important aspect regarding evaluation is how the assessment data might be used, in a feedback loop, to guide system improvement. The first option is that the researcher, based on the evaluation outcomes, makes changes to the system. Another alternative, extending the idea of unsupervised assessment methods, is to enable unsupervised adjustment of system parameters based on the evaluation data. Such an unsupervised evaluation loop would allow automated parameter tuning (e.g., avatar lip aperture adjustment) and a systematic assessment of the influence of each parameter in the overall quality.

5 FRAMEWORK INSTANTIATION

In what follows, we present the first instantiation of the audiovisual speech synthesis conceptual frame-

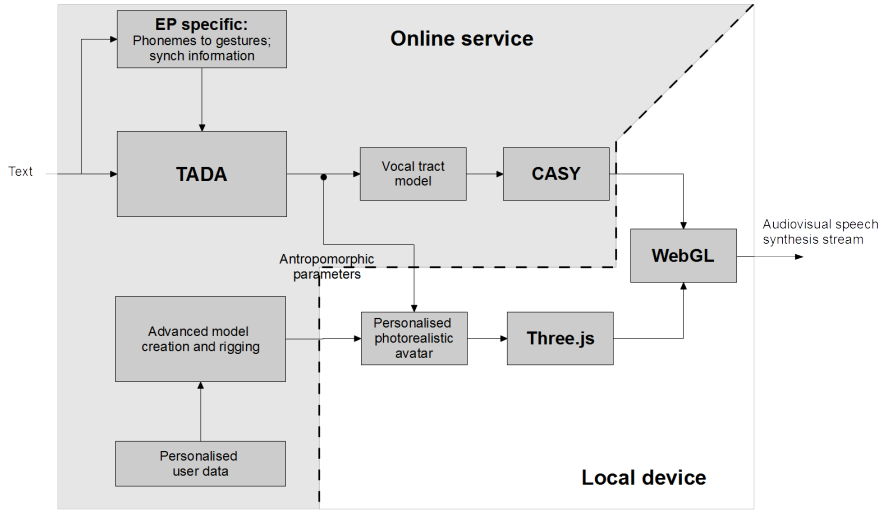

Figure 5: Diagram of the articulatory audiovisual speech synthesis core depicting details regarding the technical options adopted for each of the main blocks.

work proposed in the previous section. At this moment, we have a full first instantiation of the audiovisual speech synthesis core, considerable work regarding its integration as an interaction modality and a very initial sketch of the evaluation module.

5.1 Audiovisual Speech Synthesis Core

Figure 5 depicts the audiovisual speech synthesis core, following the conceptual structure presented earlier, but now providing additional detail regarding the concrete implementation options considered for each of its blocks.

Considering our research interests and previous work, this first instantiation of the framework supports audiovisual synthesis in EP. Note that this is not a limiting aspect to the versatility of this instantiation, since the change of the language specific elements is the key requirement for the framework to support another language. In the language specific block, the main aspect is the data required to derive gestures from words and we considered previous contributions by Teixeira et al. (Teixeira et al., 2008) and Oliveira (Oliveira, 2009) providing the gestural definition of EP sounds. The language specific module also includes the syllabification of the input text, and its conversion to phonemes which, at this time, is still not fully automated and relies on a static dictionary.

For the static and dynamic modelling of the vocal tract we considered TADA that provides the composition of the gestural score and the dynamic model computing the articulator trajectories along time considering coarticulation effects. It is TADA that provides the antropomorphic parameters concerning lip, velar and tongue movement that subsequently drive the auditory and visual synthesis.

Regarding auditory speech synthesis, for the sake of simplicity, we consider, at this stage, the partial implementation of the CASY synthesizer, provided along with TADA (Rubin et al., 1981) that suffers from a few limitations particularly regarding nasality, an important aspect for EP. The envisaged alternative is the consideration of SAPWindows (Teixeira et al., 2002), which was built for EP, but other alternatives are possible, e.g., based on the model and synthesizer proposed by Birkholz (Birkholz, 2013).

As conceptualized, our approach considers the set of anthropomorphic parameters driving auditory speech synthesis (e.g., lip and tongue configurations), as input data for the visual speech synthesis yielding inherent synchronization among both streams.

The 3D model used (see figure 6) has an advanced animation rig provided by FACEINMOTION[1]. It has two distinctive features that make it interesting in this context: (1) it enables controlling aspects, such as lip aperture, with changes propagating smoothly to the face; (2) the rigging is easily transferable across models, meaning that personalized avatars can be created keeping the bones, i.e., the elements in the avatar model controlling parts of the face, common among models.

The articulator trajectories are transformed in animation parameters mostly by performing amplitude adaptation or by using the same articulator data to modify a group of bones. For example, lip protrusion data is used to manipulate bones regarding the upper and lower lip, and the left and right mouth corners since the model does not yet support a higher level

[1] http://www.faceinmotion.com/

Figure 6: The photorealistic avatar considered for the instantiation of the conceptual framework.

control for this gesture.

Since animation relies on the articulator trajectories it is basically independent from the language or articulatory model considered in TADA, as long as articulator trajectories are generated in the same value ranges.

5.2 Interaction Modality

To move into the envisaged scenario of deploying the audiovisual speech synthesis system as an interaction modality, it has been spread between local and online services.

First, TADA has been optimized to run faster and relevant parts converted to C language, overall reducing the required computation time by half. Then, and to move it away from its native Matlab environment, not fitted for integration into an interaction modality, it was encapsulated inside a RESTful service returning articulators' trajectories for given sentences. This also fulfils the goal of putting the computational weight outside of the local device.

Regarding the rendering of the audiovisual synthesis, we considered that at least part of it needed to be on the client side to provide some level of control and customization (e.g., a personalized avatar) on how the synthesis is integrated and used. Therefore, the final rendering is performed on the client. Figure 5 depicts the part of the core that is actually performed by the client device.

For rendering the audiovisual stream, on the client side, we considered WebGL in conjunction with Three.js, the later used to manipulate the model according to the received anthropomorphic parameters. Using WebGL has the advantage of not requiring the installation of any additional libraries since it is natively supported by most web browsers, allowing for out-of-the-box multi-platform, multi-device support. To make it clear, this does not invalidate the adoption of another forms of rendering, e.g., including the avatar in a complex virtual scene. A simple example video can be found at http://sweet.ua.pt/sss/resources/visualspeech/files/opapaestanotrabalho2.mp4

5.3 Evaluation

At this time, the audiovisual synthesis core already provides output streams that can be used for perceptual tests. However, regarding the unsupervised assessment, we are still at an early stage, studying how the reference data should be gathered and used.

6 CONCLUSIONS

This article argues in favour of an anthropomorphic (articulatory) perspective for audiovisual speech synthesis and proposes the main aspects of a conceptual framework that should enable its evolution. This framework considers articulatory audiovisual speech synthesis as a research tool in speech production, supported by an evaluation module and tightly connected to its use in multimodal interaction systems. A first instantiation of the conceptual framework is also presented demonstrating the viability of the proposal. It already provides audiovisual synthesis of simple sentences in EP. At it current stage, it already enables moving from working on ground work such as gestural models up to having a speaking avatar in a web browser.

Considering the proposed conceptual framework, our first instantiation still does not fully address evaluation. While at its current stage the system can enable the execution of perceptual tests, we have yet to deploy a systematic quantitative evaluation module to cover all stages of the framework.

ACKNOWLEDGEMENTS

Samuel Silva is funded by grant SFRH/BPD/108151-/2015 from FCT. Research partially funded by IEETA Research Unit funding (UID/CEC/00127/2013.) and Marie Curie Actions IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP).

REFERENCES

Almeida, N., Silva, S., and Teixeira, A. (2014). Design and development of speech interaction: A methodology. In *Proc. of HCII, LNCS 8511*, pages 370–381, Crete, Grece.

Almeida, N., Silva, S., Teixeira, A. J. S., and Vieira, D. (2016). Multi-device applications using the multimodal architecture. In Dahl, D., editor, *Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything, (to appear)*. Springer, New York, NY, USA.

Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, 8(4):1–17.

Browman, C. P. and Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18:299–320.

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.

Files, B. T., Tjan, B. S., Jiang, J., and Bernstein, L. E. (2015). Visual speech discrimination and identification of natural and synthetic consonant stimuli. *Frontiers in psychology*, 6.

Freitas, J., Candeias, S., Dias, M. S., Lleida, E., Ortega, A., Teixeira, A., Silva, S., Acarturk, C., and Orvalho, V. (2014). The IRIS project: A liaison between industry and academia towards natural multimodal communication. In *Proc. Iberspeech*, pages 338–347, Las Palmas de Gran Canária, Spain.

Hall, N. (2010). Articulatory phonology. *Language and Linguistics Compass*, 4(9):818–830.

Massaro, D. W. (2005). *The Psychology and Technology of Talking Heads: Applications in Language Learning*, pages 183–214. Springer Netherlands, Dordrecht.

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

Nam, H., Goldstein, L., Browman, C., Rubin, P., Proctor, M., and Saltzman, E. (2006). *TADA manual*. New Haven, CT: Haskins Labs.

Oliveira, C. (2009). *From Grapheme to Gesture. Linguistic Contributions for an Articulatory Based Text-To-Speech System*. PhD thesis, University of Aveiro (in Portuguese).

Rubin, P., Baer, T., and Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *The Journal of the Acoustical Society of America*, 70(2):321–328.

Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., and Browman, C. (1996). CASY and extensions to the task-dynamic model. In *Proc. Speech Prod. Seminar*, pages 125–128.

Saltzman, E. L. and Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological psychology*, 1(4):333–382.

Schabus, D., Pucher, M., and Hofer, G. (2014). Joint audiovisual hidden semi-markov model-based speech synthesis. *J. of Selected Topics in Signal Proc.*, 8(2):336–347.

Scott, A. D., Wylezinska, M., Birch, M. J., and Miquel, M. E. (2014). Speech mri: Morphology and function. *Physica Medica*, 30(6):604 – 618.

Serra, J., Ribeiro, M., Freitas, J., Orvalho, V., and Dias, M. S. (2012). A proposal for a visual speech animation system for european portuguese. In *Proc. IberSPEECH*, pages 267–276, Madrid, Spain. Springer.

Silva, S., Almeida, N., Pereira, C., Martins, A. I., Rosa, A. F., e Silva, M. O., and Teixeira, A. (2015). Design and development of multimodal applications: A vision on key issues and methods. In *Proc. HCII, LNCS*.

Teixeira, A., Oliveira, C., and Barbosa, P. (2008). European Portuguese articulatory based text-to-speech: First results. In *Proc. PROPOR, LNAI 5190*, pages 101–111.

Teixeira, A., Silva, L., Martinez, R., and Vaz, F. (2002). SAPWindows - towards a versatile modular articulatory synthesizer. In *Proc. of IEEE Workshop on Speech Synthesis*, pages 31–34.

Teixeira, A. J. S., Almeida, N., Pereira, C., e Silva, M. O., Vieira, D., and Silva, S. (2016). Applications of the multimodal interaction architecture in ambient assisted living. In Dahl, D., editor, *Multimodal Interaction with W3C Standards: Towards Natural User Interfaces to Everything, (to appear)*. Springer, New York, NY, USA.

W3C Consortium (2003). W3C multimodal interaction framework - technical note (accessed oct 2016).

Železný, M., Krňoul, Z., and Jedlička, P. (2015). *Analysis of Facial Motion Capture Data for Visual Speech Synthesis*, pages 81–88. Springer International Publishing, Cham.