AnTon: an Animatronic Model of a Human Tongue and Vocal Tract

Robin Hofe, Roger K. Moore
Department of Computer Science, University of Sheffield, Sheffield, UK
R.Hofe@sheffield.ac.uk

Abstract
This paper describes AnTon, the first animatronic model of a human tongue and vocal tract which is being developed using biomimetic design principles. The present model consists of movable tongue and jaw models that are connected to a fixed hyoid bone and a skull. AnTon’s ability to reproduce speech gestures has been investigated using MRI scans of human subjects producing standard vowels. In the future, the animatronic vocal tract will be used to study the relation between speech variation and energy investment, specifically in the context of the Lombard reflex.

Index Terms: speaking robots, vocal tract modelling, speech variation, speech energetics, Lombard speech

1. Introduction
Contemporary speech technology struggles with the amount of variation that occurs in natural speech [1]. This means that, in order to estimate the parameters of statistical models, researchers collect larger and larger databases of recorded speech. Despite this effort, speech technology applications are still a long way away from human performance levels, and it is doubtful whether the accumulation of more and more speech data alone will be sufficient to solve the problem [2]. It can therefore be argued that it might be beneficial to reach a better understanding about the general principles that govern variation [3].

Lindblom’s theory of hypo- and hyperspeech (H&H) explains speech variation qualitatively as a result of speakers investing more or less energy into their speech output [4]. The governing parameters are the intelligibility they want to achieve, and their communicative environment. In order not to waste valuable energy during speech production, only so much is invested as is necessary to maintain the intelligibility required. The result is variation along a continuum from hypo- to hyper-articulated speech. A prominent example of such behaviour is the Lombard reflex [5]. Speakers who are subject to noise in their conversational environment increase their vocal effort to remain intelligible.

There is, however, a lack of quantitative data to support the claims made by the H&H theory, e.g. it is not known how much energy is actually invested to make up for a given amount of noise. The reason for this shortfall lies in the difficulties involved in acquiring quantitative data from human subjects [6]. One solution might be to collect data from an artificial speech generator. For this purpose, an animatronic model of a human tongue and vocal tract is currently being developed in the Department of Computer Science at the University of Sheffield – ‘AnTon’ (see figure 1). This paper introduces the AnTon project and presents first results from speech gesture imitation tasks.

The following section describes the project’s aims and the animatronic model. Section 3 introduces AnTon in its current stage of development. Results of experiments to mimic articulatory gestures are presented in section 4.

2. AnTon: an overview

2.1. Aim of the project
A number of researchers are trying to find appropriate models for the different sources of speech variation in order to increase the performance of applications when presented with spontaneous speech. Amongst the sources of variation that receive a large amount of attention are:

- emotional state [7, 8];
- medical conditions [9, 10];
- Lombard speech [11, 12];
- coarticulation [13];

Instead of developing many different models to treat the various sources of variation, one could argue that a unified framework might be useful. The H&H theory provides such framework. H&H formulates general principles that govern
speech variation, namely speech energetics and intelligibility requirements. In order to apply these principles, the collection of quantitative data is a prerequisite.

Speech variation can be defined as deviations from ‘neutral’ speech in a given feature space. Within that space, a variation vector $V$ can be found, having a magnitude, a direction, and an orientation (see figure 2). In order for an H&H framework to be applicable, the features used have to relate to speech energetics.

The AnTon project’s aim is to evaluate the H&H theory quantitatively by formulating the magnitude of $V$ as a function of an objective and controllable influence. Lombard speech is the result of such an influence (a noise signal), and for that reason, the animatronic tongue and vocal tract model will be used in Lombard reflex experiments. The hypothesis tested is that an increase in noise requires an increase in vocal effort in order to maintain a specified level of intelligibility.

The direction and orientation of $V$ relate to the qualitative aspects of variation. The specification of individual types of variation within an H&H framework will set these parameters.

2.2. The animatronic vocal tract

AnTon is an animatronic model of a human tongue and vocal tract. The ultimate aim is that is will eventually learn to produce speech sounds using auditory feedback as well as gather information about the energy costs involved in producing specific speech gestures. Human anatomy provides the guideline for its construction; functional considerations are only made where it is impossible or not feasible to mimic the biological system. The artificial vocal tract’s functionality is thus derived solely from its anatomy, i.e. a ‘biomimetic’ design principle [14].

A functional design (e.g. as in the ‘Waseda Talker’ [15]) would have resulted in a model that might have been much more energy efficient than the biological system. This is due to the fact that speech is not the primary function of the human vocal tract [16]. Rather, the evolved organ is the result of joint optimisation for a number of different tasks. This compromise has to be present in a model the output of which should be comparable with human behaviour, hence the biomimetic approach.

The finished robot will comprise actuated models of the tongue, jaw, hyoid bone, and soft palate. Passive structures will include facial skin, lips, and pharynx. Active lip movement and a subglottal system might be added to further refine the model in the future.

The animatronic model will have access to auditory feedback to perceive and judge its own utterances. It will also receive feedback about the energy cost of its articulatory gestures, obtained from measuring the motors’ power consumption.

3. State of development

AnTon currently consists of movable tongue and jaw models that are actuated by servo motors. The tongue is attached to the jaw and rests on a hyoid bone that is fixed in its position. The jaw is connected to the skull by two joint capsules enclosing the mandibular condyles [17]. Bones in the model consist of hard plastic material, whereas the tongue is cast from a very soft silicone.

Muscles are represented by filaments which connect to servo motors. The servo motors are driven by a controller board connected to a PC’s USB port. The filaments are connected to the motors either individually, or in symmetric pairs. There are currently 16 servo motors available, eleven of which are used at present. The capacity can easily be extended to allow for 32 or more motors in the future.

The four muscles associated with jaw movements during speech [18] are represented by seven filaments. They follow the lines of action found in the biological antetype [19]. Five of the muscles forming the tongue are represented by eleven muscle filaments which follow realistic muscle fibre orientations [20, 21]. The filaments of the tongue model connect to a system of meshes embedded in the silicone tongue body. Figure 3 illustrates which muscles are incorporated in the tongue model. The palatoglossus will be added together with the soft palate.

4. Performance

AnTon’s performance was compared in pre-vocalisation experiments with midsagittal MRI scans of a human subject articulating the cardinal vowels /a/, /a/, and /i/ [22].

4.1. Test procedure

It is not straightforward to obtain a midsagittal tongue contour from the animatronic model. The solution was to use pieces of thermoplastic. The thermoplastic applied is deformable at temperatures above 60°C and returns to a solid state when it cools down. Strands of warm thermoplastic were used to obtain the midsagittal contour (see figure 4). This procedure was repeated three times for each tongue position, in order to detect possible errors in single casts. Despite the delicate procedure to apply the thermoplastic, the series of contours obtained were consistent.

The method is rather slow and it takes a long time to imitate a gesture by successively drawing closer to it. For this reason, it was tried to match the target gesture qualitatively rather than exactly.
4.2. Results

The results for the vowels /a/ and /u/ are presented in figure 5. The tongue contour for the /a/ gesture was closer to the target than the one for the /u/. The reason is that the animatronic model was capable of adopting the /a/ position using mainly one pair of muscles, the styloglossi, which are represented by two muscle filaments in the model. This made the step-by-step convergence towards the MRI target easier. For the /u/ sound, the genioglossus and the pair of hyoglossi, represented by three and four filaments, respectively, worked together to produce the shape. The convergence in that case was much slower and involved some amount of try-and-error.

The vowel /i/ could not be imitated as the current model lacks a palatoglossus muscle to pull the tongue upwards. The fixed hyoid bone also restricts upward movement. As mentioned above, these limitations will be overcome in the next step of development.

Overall, AnTon was able to reproduce those gestures that were expected at its current stage. A video of the animatronic model in motion can be downloaded from the project website: http://www.dcs.shef.ac.uk/~robin/anton/anton.html.

5. Conclusion

This paper describes the first biomimetic animatronic model of a human tongue and vocal tract – AnTon. It was designed to research fundamental properties of variation in speech. It is intended to serve as a tool to investigate speech energetics, and to evaluate Lindblom’s H&H theory quantitatively.

The present stage of development and future improvements of the model have been discussed, and functional tests and a comparison with typical vowel gestures have been presented. The tests confirmed the ability of the model to adopt the basic articulatory positions that can be expected given the present construction progress.

Regular updates on the development and experiments with AnTon will be available on the project website: http://www.dcs.shef.ac.uk/~robin/anton/anton.html.

6. Acknowledgements

The MRI data is by courtesy of Pierre Badin and Antoine Serrurier, CNRS, Grenoble, France [22].

7. References


Figure 5: Comparison of articulator shapes for the articulatory gestures forming (a) /a/, and (b) /u/. MRI data by courtesy of Pierre Badin [22]. Black: soft tissue articulator boundaries, MRI; grey: teeth and bone, MRI; white: Tongue contour obtained from AnTon.