

Speech Synthesis Research in Virtual Reality Systems

Miloš Cerňak

KTL FEI STU, Ilkovičova 3, 812 19 Bratislava

Email: mcernak@ktl.elf.stuba.sk

Abstract. *The paper describes a program of speech synthesis research in virtual reality systems, that merges unit selection based speech synthesis, visual speech synthesis and articulatory speech synthesis research with virtual reality application development.*

Keywords. Speech synthesis, Virtual Reality.

1. Introduction

Like speech recognition, speech synthesis is a critical component of a speech-based interface system. Speech recognition and synthesis, together with meaning extraction and speech output generation (what should be spoken) components, are the basic building blocks of a complete speech-based interface system.

As these speech based systems mature, a fertile area of application is the growing number of Virtual Reality (VR) applications. Speech technology can add a great deal of realism and flexibility to these virtual worlds, allow a user to interact more naturally with his virtual surroundings. Because it is such a common mode of human communication and it incorporates another of the human senses, speech is a natural enhancement for VR applications. Speech systems increase the modality of the control and consequently the VR app becomes more useful and more interesting.

In following section we describe some of the intersections of VR and speech technologies, particularly speech synthesis. In section 2 we describe unit selection speech synthesis in speech interface to VR app. Section 3 describes visual speech synthesis in speech interfaces for VR applications and section 4 introduces an articulatory speech synthesis by means of fluid dynamics; a possibility to develop an experimental articulatory synthesizer within the context of VR systems.

2. Unit selection speech synthesis

Unit selection speech synthesis is a concatenative synthesis approach also known as corpus-based speech synthesis [6]. It utilizes encoded speech peculiarities within a recorded speech database (hence corpus) that has been properly linguistically annotated. The entire corpus becomes a kind of acoustic inventory, where best candidates are selected according to the specifications of target units in real time, and consequently the units are concatenated in some optimal way. Speech database consists of whole phrases recorded either for a specific domain [2], or for general text-to-speech. These databases can be also to train speech recognition engines.

There are several attempts to create speech interface to VR applications [4], where unit selection speech synthesis has been successfully used. At the present time, the unit selection approach provides the best intelligibility and naturalness of synthetic speech, compared to other synthesis approaches currently in use (e.g. diphone concatenative synthesis, rule-based formant or articulatory synthesis). Unit selection exploits the exponentially falling cost of computer storage space and will scale well as computational power grows. The unit selection approach is used in AT&T's NextGen TTS system, Microsoft's Whistler, and IBM's trainable speech synthesizer.

Fig. 1 shows visualized molecule in Virtual Reality environment (green RNA chain coming to "feed" the molecules). Speech interface (e.g. dialog modeling system) to the VR application recognizes human commands and generates speech back to the human (e.g. confirmations, questions). Speech synthesis block can successfully use unit selection approach in limited domain, which offer high naturalness of speech.

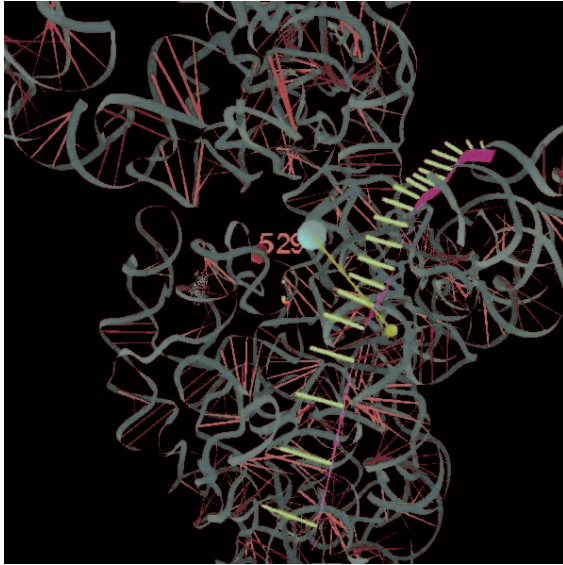


Figure 1. RNA chain in the molecule

3. Visual speech synthesis

Visual speech synthesis falls in the category of multi-modal speech synthesis algorithms and there is currently considerable interest in developing 3D-animated agents (also known as “talking heads”) for use in multi-modal spoken dialog systems [3]. Akin to the offline problem of speech synchronization in facial animation, the development of a real time algorithm for 3D-animated speaking agents has many applications in VR environments. Adding the visual modality to aural speech synthesis within a VR app can qualify the auditory information and provide segmental cues on place of articulation or some extra prosodic information [1]. Visual synthesis also enables the control of VR app by hearing impaired persons, by allowing the auditory cues to be supplemented by means of “reading the lips” of the 3D agent.

Speech synthesis techniques used with talking heads are basically rule-based systems, where speech is generated by means of a set of rules. The rules for such text-to-speech system also generate the parameter tracks for the face.

4. Articulatory speech synthesis

Articulatory speech synthesis research began with W. von Kempelen, who designed a mechanical model of the human vocal tract over two centuries ago. Nowadays, despite the many simplifications made to contemporary models of vocal tract and the movements of human articulators, articulatory speech synthesis is still

computationally expensive, and synthesis results are poor.

The science of fluid mechanics is also known as aeroacoustics. Aeroacoustics was developed in the early 1950s to help reduce noise from jet aircraft engines being developed for civilian use -- a problem concerning sound production and transmission by airflow. Recently, these methods have begun to be applied, via special tools and theoretical framework, to speech production [5]. Combining the expertise from different scientific fields often brings new opportunities and improvements in research and development. Ongoing studies at some VR research labs (e.g. Virtual Reality Applications Center at Iowa State University, USA) are providing new and important insights for the visualization of computational fluid fields that may prove useful in solving these problems.

However, there remain two fundamental problems in applying airflow simulation techniques to articulatory speech synthesis research:

- o Air in the vocal system is not static (moves from the lungs out of the mouth, carrying the sound field along with it)
- o Turbulence is generated

Both of these two types of motion are not acoustic, but have a profound impact on how sounds are produced and how they are transmitted to the ear of the listener [5]. VR systems offer a new tool for visualizing the results of vocal tract modeling, to provide researchers with new insights into how these models can be systematically improved over time. In this case, human vocal tract should be modeled in VR system by means aeroacoustics principles.

Fig. 2 shows an example of visualized velocity magnitude in a jet, taken from similar project [7]. The flow is from left to right (red = high; blue = low), computed from Navier-Stokes solver.

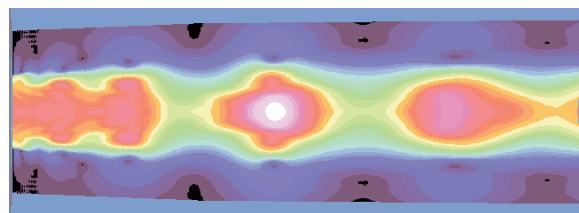


Figure 2. Velocity magnitude in a jet formed behind a small inlet into a larger straight tube

5. Conclusion

In this paper we have given a short introduction to the intersections between speech synthesis research and the application of VR systems. As work progresses, the number of such intersections will rise. Speech technology continues to improve, and as it does its limits will be continually tested by complex applications of these technologies within “non-speech” technology areas, like VR research. A promising avenue for progress lies in close cooperation between different research groups, where expertise from many different areas can be combined.

6. Acknowledgements

I would like to thank to Dr. Adrian Sannier from Virtual Reality Applications Center at Iowa State University for his valuable corrections of this paper.

7. References

- [1] Jonas Beskow, Björn Granström and David House, (2001), A Multi-Modal Speech Synthesis Tool Applied to Audio-Visual Prosody, in In (Keller, E., Bailly, G. Et al., Eds.): *Improvements in Speech Synthesis*, (pp. 372-382). Wiley & Sons.
- [2] Black, A. and Lenzo, K. (2000) *Limited Domain Synthesis* ICSLP2000, Beijing, China.
- [3] Cassel, J., (2000), Nudge nudge wing wing: Elements of face-to-face conversation for embodied conversational agents, In J. Cassel, J. Sullivan, S. Prevost, and E. Churchill (eds), *Embodied Conversational Agents* (pp. 1-27). The MIT Press.[4]
- [4] Cerňak, M., Sannier, A., (2002), Command Speech Interface to Virtual Reality Applications, Technical Report, Virtual Reality Applications Center at Iowa State University of Science and Technology.
- [5] Michael H. Krane, Daniel Sinder, James Flanagan, (1999), Fluid Dynamics Improves Understanding of Speech Production, Acoustical Society of America ASA/EAA/DAGA '99 Meeting, Berlin, Germany.
- [6] Bernd Möbius, (2000), Corpus-Based Speech Synthesis: Methods and Challenges, Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS 6 (4), pp. 87-116.
- [7] Speech Generation From Fluid Dynamic Principles, available from: http://www.speech.kth.se/~olov/art_synth.html [June 2002].