Speech is the richest and most ubiquitous modality of communication used by human beings. Voice production is one of the most expressive instruments of the human body and people have always been fascinated for artificial speech and singing production. Even though vocal behavior and expression is a very complex mechanism, we realize a highly interactive and social process. Until now, artificial voice production has been based on the Text-To-Speech (TTS) technologies, converting a static text into an intelligible and natural waveform, lately with great success. However the relevance of speech production is not only based on these properties, it is also linked to the context, to an ongoing process of interaction between speakers and to their social environment and background. Through this interdisciplinary redefinition of expressive speech, we see that the nature of voice production is primarily a real-time, dynamic gesture, involving the vocal organs, face and body. Speech is a performance, a gestural phenomenon that transmits messages with both information and emotions.

1. The MAGE Synthesis Platform

As these new trends in understanding expressivity in speech are being explored, one might notice that a real solid platform for performative speech synthesis is missing. Indeed TTS, as a platform, has been tackling and greatly solving the problem of text reading, but not the problem of the artificial speaker. In the text reading paradigm, a significant amount of text is required in advance to be processed into sound as a whole target. Moreover, during this text to speech conversion process any external influence is rather limited, since it results in sound quality degradation. On the other hand, an artificial speaker enables interactive communication by inferring on speech outputs at various production levels and time scales. In fact, such a system has totally different requirements. It needs to have a reactive programming architecture, and to be both listener-specific and context-aware. To our current knowledge, MAGE is the first platform for reactive programming of speech synthesis solutions. It was released recently as a C/C++ open-source project [1, 2].

The MAGE\(^1\) speech synthesis engine is based on HTS [3], the open-source HMM-based statistical parametric synthesis algorithm from Tokuda et al [4]. While still not reaching the intelligibility and naturalness of non-uniform unit selection (NUU) algorithms, HMM-based approaches are quickly improving and, more importantly, they are based on a highly flexible well-defined architecture. Indeed both HMM-based trajectory generation and MFCC-based sound synthesis can be deeply modified. MAGE is a complete architectural redesign of HTS, streaming the speech sound in real-time, according to synthesis parameters that are sent on-the-fly. MAGE has a high-quality output while rendering incoming labels, durations, pitch and vocal tract length parameters with only one phonetic-label delay.

2. Implementation and Integration

MAGE provides an API for reactive programming in C and C++, aimed at being included in realtime audio softwares. It is thread safe and independent from the actual HTS-based speech engine. In this version, we release MAGE with the pHTS engine, a performative modification of HTS. Fast and easy prototyping is possible with MAGE, since it can be easily imported into other platforms. It can be simply combined with OSC-enabled sensors, allowing both context and prosody user control. Contextual control is implemented by parsing on-the-fly small groups of phonemes, that we can “chunks” into streams of phonetic labels. Prosody control is based on altering on-the-fly pitch curves, duration or vocal tract parameters from gestures.

MAGE comes as a consequential implementation, following the idea of performative speech synthesis, as a way of looking beyond TTS. It results from discussions across different disciplines, such as speech processing, linguistics and human-computer interaction (HCI), attempting to bring a new platform for addressing their problems. In the area of computer music and new interfaces for musical expression, MAGE is targeted to combine simple prototyping with meaningful gestural control, to bring synthetic speech on stage and suggest new art forms.

3. Discussion

Our goal is to build a new framework for understanding long-term questions in speech production, such as degrees of coarticulation, speech motor control, speech planning, intonation, voice quality, speech time scales, etc. through gestural control and interactive interfaces, mainly through mobile and social computing. At this point it is important that we have feedback for this project from actual users. Users that set to work this library, have questions, suggestions and comments. Convening the expectations of the users towards more interactive use of speech synthesis, will eventually bring the next release of the platform to address this practical feedback and become a better tool for engineers, linguists and artists.

4. References


\(^1\)The MAGE project is funded through two PhD grants by the University of Mons (numediart, grant 716631) and Acapela Group S.A.