

Talk Show

Byungjun Kwon, Björn Erlach and Luc Döbereiner

Amsterdam, August 2007



1 Introduction

Our goal was to build a system for the creation of music, text and synthesized speech using the *VR Stamp*¹ voice recognition module. The aimed result was a music piece or installation situated in between of text and sound, speech and music. Speech recognition techniques offer possibilities for working at these boundaries, as they can provide a textual interpretation of sound, yet their sensing capabilities are not limited to speech, but can equally well be applied to other sounds.

1.1 The VR Stamp

The *VR Stamp* module is designed to be embedded in industrially fabricated products, which require voice recognition and speech output. It is programmable and comes with a range of speech recognition functionalities and a C compiler, which allows us to develop our own applications, using its recognition capabilities.

In comparison to software systems, which often provide immense flexibility, its synthesis and processing capabilities are quite limited. The sound output is restricted to simple sound file playback. We experimented with non-speech sounds, but realized that the chip's built-in technology is optimized for speech-sounds, so we decided to use speech-related sound material. However, we chose the *VR Stamp*, because of its limitations. Its constraints provide us with a framework, which can be explored in the ten days of our residency.

¹See http://www.sensoryinc.com/products/vr_stamp_toolkits.html

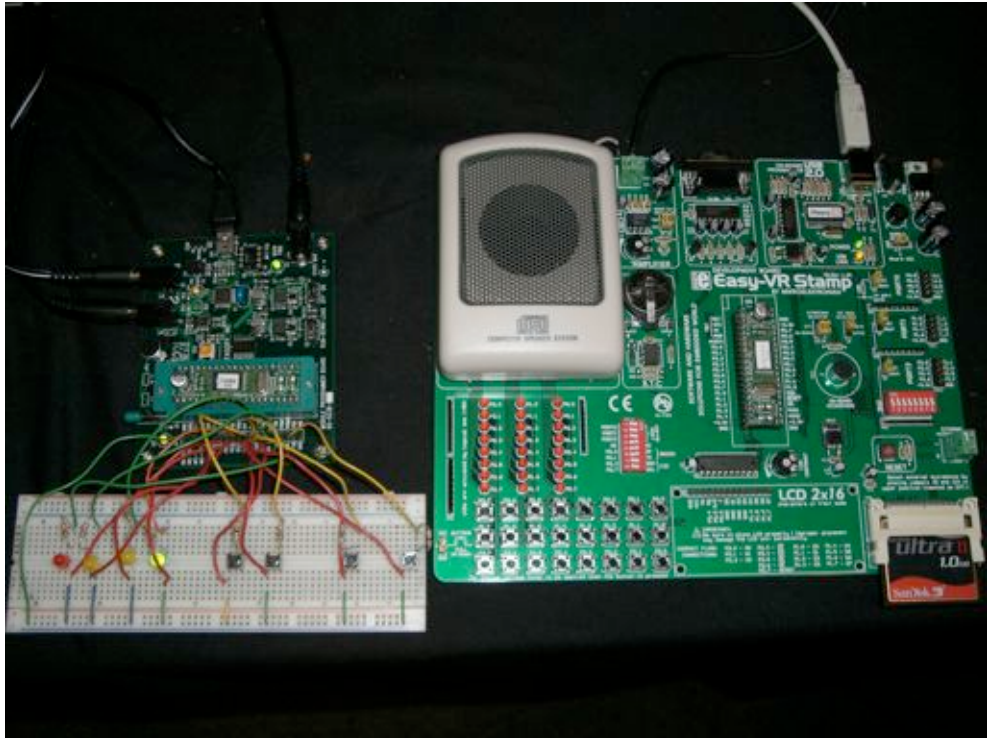


Figure 1: *VR Stamp Toolkit* (left) and *Esay VR Stamp* (right)

2 Developing a Set-Up

After a phase of several days in which we became acquainted with the *VR Stamp*, its development environment and functionalities, we started developing several installative set-ups. On a general level a set-up consists of a set of trained sounds and a mapping of these sounds to responses. The responses are utterances consisting of several phonemes and possibly silence in between, these sequences of phonemes and silences are built up from smaller constituent patterns, probabilities for silence, consonant/vowel ratio and other parameters depend on the recognized input.

Firstly, we trained the system to respond to a set of Latin and Chinese phonemes and secondly we developed response rules, which are used to create sounds by concatenation of these phonemes depending on the previously recognized input. The current set-up consists of two autonomous *VR Stamps*, one with a Latin and one with a Chinese language model, trained to react to each others output. The communication between the two devices functions purely acoustically, thus they can also react to other sounds in the room. The set-up could be described as a Markov-process, but due to the acoustic communication, there is a degree of indeterminacy.

3 Future Work

There are several ways in which we want to extend the project further. Due to technical and time constraints we were only able to use the chips speaker dependent voice recognition capabilities. In the future we want to switch to a speaker independent model, which may also allow us to train the system to non-speech sounds.

Secondly we plan to use other language models, then only Latin and Chinese and built a *state* into the system to allow gradual change of the system's behavior, possibly triggered by certain sounds, which would result in more variations.

Finally we plan to built an independent hardware using the programmed chips, instead of the development boards.