MASSY Speaks English: Adaptation and Evaluation of a Talking Head

Sascha Fagel

Department of Language and Communication, Berlin Institute of Technology, Germany

sascha.fagel@tu-berlin.de

Abstract

This paper presents an extension of a talking head that was implemented to synthesize audiovisual speech utterances in German. A mapping of English to German phonemes and a subsequent refinement of the co-articulation model enables the system to generate English utterances. A modified rhyme test was carried out for evaluation in terms of intelligibility. Results show that the visualization increases the identification of isolated words from 27 % to 50 % when played with audio.

Index Terms: talking head, intelligibility, evaluation

1. Introduction

Speech production is a physiological process that can be heard and seen. When speech is perceived, audition and vision provide partly complementary information that results (among others) in enhanced intelligibility of audiovisual speech compared to speech that is only audible [1]. This advantage was shown to be replicable using synthetic visual speech as long as audible and visible speech are coherent [2].

2. System Description and Adaptation

The basic system [3] consists of modules for phonetic transcription, audio synthesis, visual articulation, and the virtual face. A plain text serves as system input. The phonetic module generates a phone chain as well as phone and pause durations by embedding the program txt2pho [4]. The audio synthesis module generates the audio signal from the phonetic and prosodic data. The audio is rendered by the MBROLA speech synthesizer [5]. The visual articulation module generates motion information in terms of control parameters for virtual articulators. These articulatory parameters are generated with a simplified dominance model adapted from [6]. This model was automatically adjusted to optimally reconstruct (in an RMS sense) articulatory data from EMA measurements of VCV style utterances. Once formed, it is used to predict articulatory movements of arbitrary German phone chains. The face module moves the virtual articulators of the virtual head according to the control parameters and adds the audio signal to create the complete audiovisual speech output. The tongue, the velum and a part of the pharyngeal wall was manually designed on the basis of midsagittal MRI. The articulation parameters of the virtual head used in the present paper are lip width, jaw height, lip height, lower lip retraction, tongue tip height, tongue back height, tongue advance, and velum height.

The English phonemes were mapped to German phonemes based on articulatory considerations. The tongue positions of the interdentals /T,D/ (which do not exist in German) were adjusted manually. The high-level synthesis of the festival speech synthesizer [7] was embedded in the system for the English phonetic transcription. The American English voice us1 of the mbrola speech synthesizer was used for the audio synthesis.

3. Evaluation

The modified rhyme test (MRT) from [8] without carrier sentence was carried out as a preliminary evaluation of the adapted system in terms of intelligibility. The test consisted of six lists of 50 monosyllabic words each. 12 subjects with normal hearing and normal or corrected to normal vision participated in the test. All word lists were used in the test but distributed over the subjects. One of the six word lists was presented to a subject audiovisually, another one was presented audible alone (blocked conditions). Hence, a higher intelligibility of audiovisual compared to audio alone presentation – if present – cannot be produced by a possible learning effect. After each presented spoken word the subject was requested to select the most probable one from six alternatives.

All 12 subjects benefit from the additional synthetic visual speech. The recognition rate increased from 27 % (std. 6.1 %) correct answers in audio alone condition to 50.2 % (std. 12.6 %) in audiovisual condition (at a chance level of 16.7 %). This gain is highly significant (p < .001). The error reduction due to the synthetic face added to the audio presentation – as called “audiovisual benefit” by [1] – is 31.7 %.

4. Conclusions

The speech visualization of the presented English TTavS system shows a significantly enhanced intelligibility compared to audio alone presentation in an evaluation experiment of isolated word recognition. The gain in intelligibility (23.2 %) is nearly the same as that one of the original German system [3] (23.6 %). However, the error reduction – which is less dependent from the SNR – is lower for the English system (31.7 %) than for the German system (41.8 %). Due to the differing evaluation schemes the comparability of the results on the English and the German system is very limited. Despite the clear benefit of the visualization an evaluation on utterance level is desirable.

5. References