Can Anybody Read Me?
Motion Capture Recordings for an Adaptable Visual Speech Synthesizer

Simon Alexanderson, Jonas Beskow

KTH Speech Music and Hearing
Royal Institute of Technology, Stockholm, Sweden
{simonal,beskow}@kth.se

1. Introduction

Speech produced in noise exhibits not only increased loudness, but also larger articulatory movements [1]. According to Lindblom's theory of Hyper-Hypo articulation [2], speakers tend to economize their speech production with the goal to make them self understood in a particular communicative situation. For an animated virtual character to function well in different environments, the ability to adapt articulatory effort seems like a useful trait. Below we describe our work towards a visual speech synthesizer capable of simulating articulatory motions for such a character, applicable to different listening conditions. To this end we have used motion capture to record a target speaker trying to make himself understood by a listener, under different conditions: in quiet, in noise and while whispering. The data will later be used to train data-driven articulatory control models for the animated character.

2. Data recording

The speaker was a male Swedish actor who was seated face to face with a listener, and was instructed to read short sentences and words from a monitor, and make sure that the listener understood what was being said. Both listener and speaker wore headphones, where they could hear the their own speech as picked up by a common omni-directional microphone, at a level that was pre-adjusted to roughly to compensate for the attenuation of the headphones. An optional stationary brown noise signal was also fed to the headphones, at different levels throughout the recording (see below).

The speaker's facial movement were recorded by a 10-camera NaturalPoint OptiTrack optical motion capture system operating at 100 frames/sec. The speaker was equipped with 37 reflective facial markers + 4 on the head. In addition, HD-video was captured using a JVC GZ-1 video camcorder. Speech was recorded via a Studio Projects C1 large diaphragm condenser microphone and an RME FireFace 800 external sound card. In order to synchronize the motion capture and the audio, a custom device was constructed, featuring three switchable IR LEDs. When switched on, the LEDs would show up as markers in the motion capture system, at the same time producing an electrical pulse in one input channel of the sound card. The same sync pulses were fed to the external-mic input of the camcorder, thus allow for precise and fully automated post synchronization of all data streams.

The recorded material consisted of 180 short Swedish sentences and 63 nonsense VCV-words (21 Swedish consonants in three different vocalic contexts). A set of 180 English sentences were also recorded.

The full Swedish sentence set was recorded under three different conditions: *Quiet*, *Noisy* and *Whispered*. Quiet is the baseline condition, where no noise was presented in the headphones. In the *Noisy* condition, brown noise at the level of 80 dB SPL was presented in the headphones of both speaker and listener. In the *Whispered* condition, no noise was presented, but the speaker was instructed to keep his voice to a whisper, and still try to make himself understood to the listener. This was done in an attempt to elicit exaggerated lip movements. A reduced set consisting of the 40 first sentences was recorded for two additional noise levels: 70 dB SPL and 90 dB SPL. VCV-words were only recorded in the *Quiet* condition. The English sentence set was only recorded in the conditions *Quiet* and *Noisy* (80 dB).

3. Preliminary data analysis

The motion capture data was sorted and labeled, and cut into segments based on the sync-signal injected by the switched LEDs. A first analysis was made by studying the distance between upper and lower lip. Figure 1 shows this distance for one Swedish sentence, for the quiet and noisy (80 dB) conditions. As expected, the lip movements exhibit much larger amplitude in the noisy conditions than in the quiet.

Table 1 shows that the average inter-lip velocity is lowest in the quiet condition and increases with noise level. The whispered condition exhibits almost twice the velocity as the in the quiet case.

Table 1: Average inter-lip velocity

	Quiet	*70 dB*	*80 dB*	*90 dB*	*Whisper*
Speed mm/s	16.6	26.1	34.0	49.8	32.0

4. Acknoledgement

This work was supported by EU Lifelong Learning Programme, project LipRead (http://www.lipread.eu)

5. References

[1] Fitzpatrick, M., Kim, J. & Davis, C. (2011): "The effect of seeing the interlocutor on auditory and visual speech production in noise", In *AVSP-2011*, 31-35.

[2] Lindblom, B. (1990): *Explaining phonetic variation: A sketch of the H & H theory,* In W. J. Hardcastle & A. Marchal: "Speech Production and, 403-439, Kluwer Academic Publishers, Dordrecht.

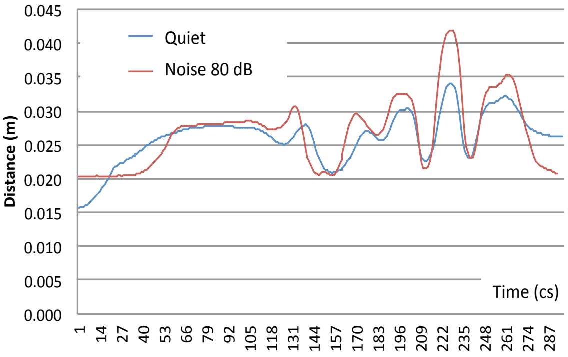

Figure 1: Intrer-lip distance for the Swedish sentence *Dom flyttade möblerna.*