The MMASCS multi-modal annotated synchronous corpus of audio, video, facial motion and tongue motion data of normal, fast and slow speech

Dietmar Schabus[1,2], **Michael Pucher**[1], **Phil Hoole**[3]

[1]Telecommunications Research Center Vienna (FTW), Vienna, Austria
[2]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria
[3]Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität, Munich, Germany
schabus@ftw.at, pucher@ftw.at, hoole@phonetik.uni-muenchen.de

Abstract

In this paper, we describe and analyze a corpus of speech data that we have recorded in multiple modalities simultaneously: facial motion via optical motion capturing, tongue motion via electro-magnetic articulography, as well as conventional video and high-quality audio. The corpus consists of 320 phonetically diverse sentences uttered by a male Austrian German speaker at normal, fast and slow speaking rate. We analyze the influence of speaking rate on phone durations and on tongue motion. Furthermore, we investigate the correlation between tongue and facial motion. The data corpus is available free of charge for research use, including phonetic annotations and a playback software which visualizes the 3D data, from the website `http://cordelia.ftw.at/mmascs`

Keywords: multi-modal speech corpus, articulatory data, facial motion

1. Introduction

For data-driven speech technology research, training corpora of speech data are an essential asset that is often created and used by research groups when required, but less often made available for the general research community. The creation of high-quality annotated corpora is a highly time-consuming and hence expensive task. This is true to an even larger extent when multiple modalities are recorded simultaneously, because of the additional requirement of synchronization between the different modalities. Furthermore, it can be argued that corpora including data acquired using special hardware, like motion capturing and especially Electro-Magnetic Articulography (EMA), have an even greater value because the equipment itself and the know-how to operate it are required for recording. To our knowledge, there are only two corpora of EMA data available free of charge, both from the University of Edinburgh (Wrench, 1999; Richmond et al., 2011). As far as speech motion capture data is concerned, there is for example a corpus of 10 speakers in affective dyadic interaction in American English (Busso et al., 2008), and we have also previously created a corpus of read speech from three Austrian German speakers with facial motion capturing (Schabus et al., 2012a)[1]. Recently, we have recorded a new interesting corpus of speech data where both EMA and motion capturing data were recorded at the same time, which we would like to share with the research community.

This new corpus differs from the existing ones in several aspects. Most importantly, it combines facial motion capture data with intra-oral EMA data. In comparison to optical motion capturing only, this has the obvious advantage of also providing tongue motion data, which is impossible to capture optically. In comparison to EMA data only, it has the advantage of providing a larger number of tracked points on the lips, eyelids, eyebrows and other areas of the face. While it is in principle possible to use EMA coils also on the face surface, the inexpensive and easy-to-attach optical markers are much less intrusive for the speaker than the EMA coils with their cable connection (one cable per coil) to the articulograph. Another difference is that our data is for Austrian German speech. One can imagine that it might be interesting to investigate inter-lingual differences in speech motion, once a larger number of corpora (of EMA and/or facial motion data) in various languages is available (of course speaker-specific effects would need to be accounted for). Finally, our data is different in that it comprises data of speech at three different speaking rates (normal, fast and slow).

In addition to general analytic usages, this corpus could be useful for other fields of research. We have been investigating 3D facial speech motion synthesis based on facial motion capturing data (Schabus et al., 2011; Schabus et al., 2012b; Schabus et al., 2013; Schabus et al., 2014), where the additional tongue data could be used to train an additional synthesizer for tongue motion, as in (Beskow et al., 2003). Cross-modality control models for speech synthesis, which have been investigated using EMA data and speech (Ling et al., 2008; Ling et al., 2009) and using facial motion data and speech (Hollenstein et al., 2013) could benefit from the usage of all three modalities in combination. Finally, we have used speech data at normal and fast speaking rates before to create ultra-fast synthetic speech via interpolation (Pucher et al., 2010). Incorporating additionally face and tongue motion data into such a system for ultra-fast speech could improve modeling and hence synthesis results.

The remainder of this paper is organized as follows. In Section 2., we describe the recording procedure and the result-

[1] That corpus is available from the website at `http://cordelia.ftw.at/fmsc`

ing data corpus. In Section 3. we give information about the form of the release, as well as about a 3D data visualization software, which is part of the release package. Section 4. provides some statistical information and data analysis results for the corpus. Finally, Section 5. summarizes and concludes this paper.

2. Recordings

We have recorded a 30-year old male native speaker of Austrian German reading 320 phonetically diverse sentences off a computer screen. The recordings took place at the premises of Ludwig-Maximilians-Universität in Munich, inside a Studio Box Premium (Studiobox, 2014) recording booth. 223 sentences of the recording script are from a well-known German text corpus (100 "Berlin" sentences, 100 "Marburg" sentences, 16 "Buttergeschichte" sentences, 7 "Nordwind und Sonne" sentences). The remaining 97 sentences were selected automatically from a large newspaper text corpus based on the improvement caused when added to the selection with respect to the representation of the diphone occurrence distribution in the entire large corpus.

Facial movement was recorded using a NaturalPoint OptiTrack Expression system (Naturalpoint, 2014) using seven FLEX:V100R2 infrared cameras. This system records the 3D position of 37 reflective markers glued to the speaker's face at 100 Hz. Four additional markers on a headband allow the removal of rigid head motion. Additionally, the system also records frontal-view gray scale video footage, also at 100 Hz (synchronized), at a resolution of 640×480 pixels. The system's standard 37 marker layout was used, as depicted in Figure 1.

Articulatory movement was recorded with a Carstens Medizinelektronik Articulograph AG501 (Carstens Medizinelektronik, 2014) EMA system. In contrast to its predecessor AG500, the AG501 does not feature an acrylic glass cube around the speaker's head, which rendered simultaneous optical marker recording impossible. The AG501 produces alternating magnetic fields, thus inducing currents in the sensor coils attached to the speaker's tongue and mouth. The currents are transmitted via a cable from each sensor to the measurement unit, where they are measured and recorded. From these measurements, the system's software computes the 3D position of each sensor coil at 250 Hz. Articulatory sensors were placed on the back, middle and tip of the tongue, on the gums above the incisors and on the nasal bridge (all five on the mid-sagittal plane). Two more sensors were placed behind the ears, and finally an eighth sensor was placed on the lower lip, between the central lower lip and right lower lip markers of the OptiTrack system. Figure 1 shows the position of most EMA sensors. Using the sensors on the nasal bridge, above the incisors and behind the ears, rigid head motion can be removed from the data. The EMA data was filtered using a finite impulse response low pass filter (Kaiser window) with cutoff frequencies of 40 Hz (tongue tip), 20 Hz (tongue middle, tongue back, lower lip), and 5 Hz (behind ears, upper incisors, nasal bridge).

Audio was recorded with a Sennheiser ME66 supercardioid microphone, with a John Hardy M1 pre-amplifier. The microphone signal as well as the synchronization signals from the EMA and OptiTrack systems were captured with a National Instruments Compact DAQ system at 25600 Hz. Audio is encoded as 32-bit floating point PCM.

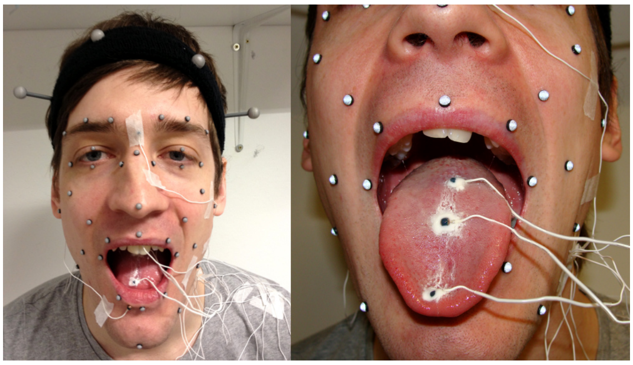
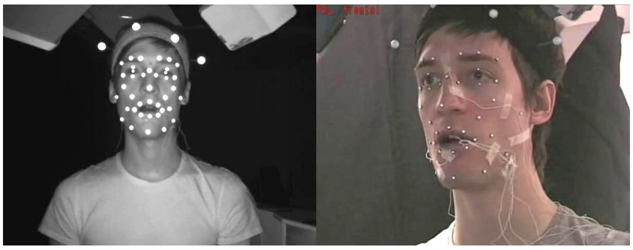


Figure 1: Placement of facial markers and EMA coils



(a) (b)

Figure 2: Example still images from the gray scale video from the OptiTrack system (a) and the color video from the camcorder (b).

Additional video footage was recorded with a Sony DSR-PD100AP digital camcorder at 25 frames per second (50 fields interlaced) and from an almost frontal view. Figure 2 shows example frames from the two kinds of videos.

All 320 sentences were first recorded at a normal speaking rate, then again at a fast speaking rate and then again at a slow speaking rate, in direct succession with short breaks. Unfortunately, one of the tongue coils disengaged during the slow part, and the recordings had to be aborted after 130 slow sentences.

3. Release and Playback Software

For the release, the data has been synchronized and cut into separate files per utterance, in all modalities (audio, video, EMA data, facial movement data). Phone borders were determined by a flat-start forced alignment procedure using HTK (Young et al., 2006) and the resulting quin-phone full-context HTK label files and mono-phone label files are part of the release. Tracking errors, which are common in optical motion capturing (like marker swaps, trajectory gaps, etc.) have been manually corrected to a large extent. EMA data and facial marker data have been aligned in coordinate space based on the position of the markers on the nasal bridge of the two systems, after rigid head motion has been removed from both 3D data streams.

The facial motion and EMA data are provided in the form of text files containing matrices that represent spatial coordinates of markers/coils per row, with one column per time

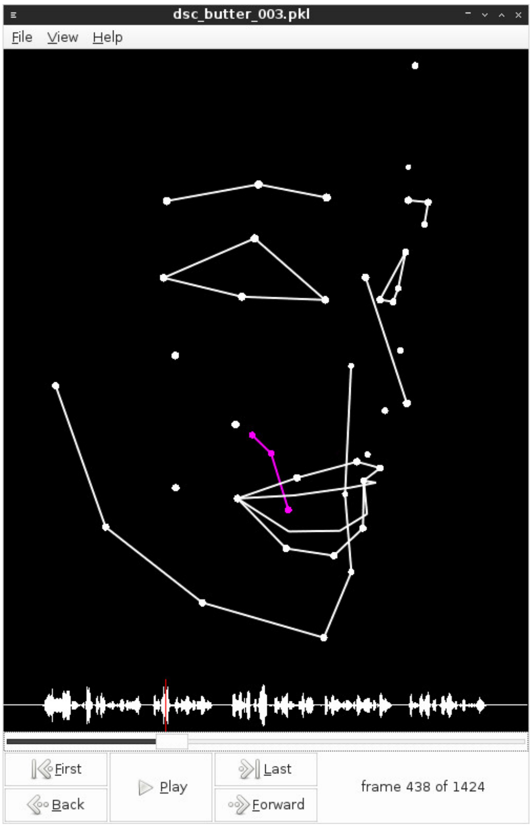


Figure 3: Screen shot of 3D data visualization software included in the corpus release

frame. Audio data is provided in the form of RIFF wave audio files, mono channel, 25600 Hz, 16-bit signed integer PCM encoding. Video data is provided in the form of H264/AAC MPEG-4 video files.

The release also contains a playback software implemented in Python using OpenGL, which visualizes the 3D data (facial markers and tongue coils) and also simultaneously plays back the corresponding audio. Figure 3 shows a screen shot of this software.

The corpus is available free of charge for research purposes from the website `http://cordelia.ftw.at/mmascs`.

4. Data Analysis

To get a better understanding of the data we have recorded, this section presents some statistics and analysis results on the corpus. As already mentioned, we have 320 sentences for both normal and fast speaking rate, and 130 sentences for slow speaking rate. For symmetry, all analytics in this section are based on the 130 sentences which we have available in all three speaking rates.

Figure 4 shows the distributions of the utterance durations for the three speaking rates as boxplots, disregarding initial and terminal silences and intra-utterance pauses. As the same 130 sentences were used, the figure shows that there is a significant difference in duration between the three speaking rates.

To quantify the different speaking rates in more depth, we have looked at the phone durations as determined by the flat-start forced alignment procedure. Figure 5 shows boxplots of the phone durations, excluding all silences and

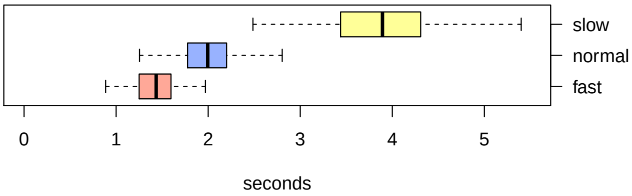


Figure 4: Boxplots of utterance durations for the three speaking rates (outliers not shown).

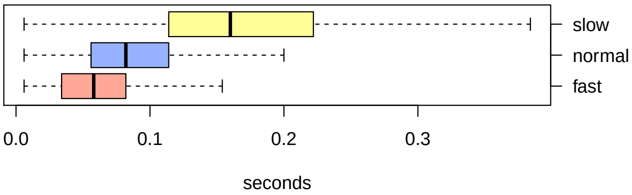


Figure 5: Boxplots of phone durations for the three speaking rates (outliers not shown).

pauses. The median phone durations for the slow, normal and fast speaking rate data are 160 ms, 82 ms and 58 ms, respectively, which are equivalent to 375, 732 and 1034 phones per minute, respectively. In addition to the occurrence of longer phones, the data also show a larger variability in phone duration with decreasing speaking rate. Furthermore, when we partition this data by phone, we can observe that the change of duration between speaking rates is larger for long vowels and diphthongs than for short vowels and stops. This is illustrated in Figure 6, which shows the duration distributions for some common short and long vowels.

To achieve a faster speaking rate, i.e., to articulate the same sequence of phones in a shorter time, three factors can be modified: 1) the velocity of the articulator movements can be increased, 2) the distance between the target articulator positions can be reduced, and/or 3) the duration of phases with stable articulator position can be shortened. Given the data of our corpus, the first two are straightforward to assess, and shall be investigated in the following.

Regarding the first factor, we have computed the movement velocities for the three tongue sensors based on the distance traveled between every two consecutive frames of the EMA trajectories. Figure 7 shows the distributions of peak velocities (greatest velocity within a phone) for the three speaking rates. Although this data may contain some noise, the increase in tongue motion velocity from slow to normal and from normal to fast speaking rate is clearly visible. The same data, but partitioned by phone, is shown in Figure 9. Again, this data is not completely reliable due to possible problems in the automatic alignment and possible tracking errors, and due to the fact that some phones do not occur very often in the corpus. Nevertheless, it is interesting to see that the order of phones is quite similar across the three speaking rates when sorted by median (as in Figure 9). In particular, phones near the close/front corner of the IPA vowel chart ([i], [iː], [y], [yː], [ɪ], [ʏ], [eː], [øː]) and certain fricatives ([s], [ʃ], [ç]) exhibit low peak velocities (and thus appear close to the bottom of Figure 9), whereas

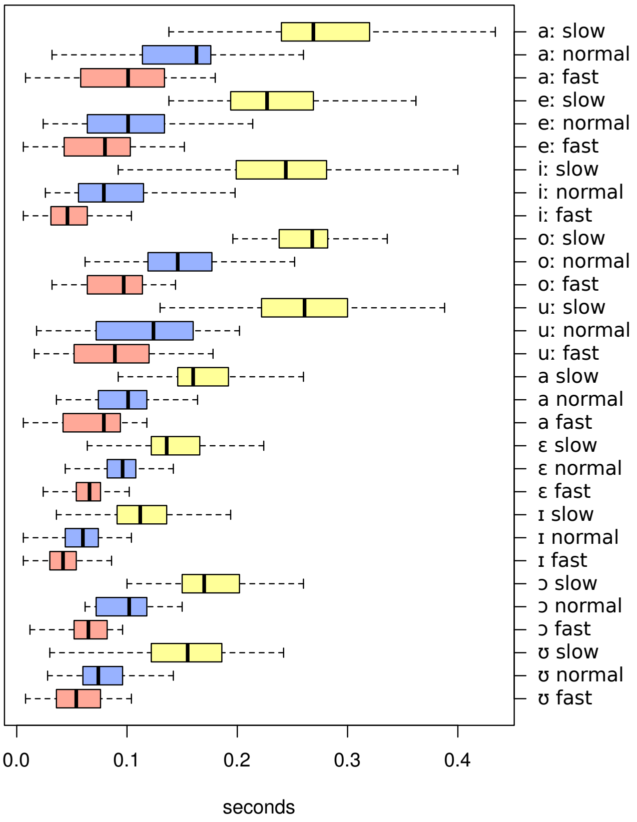


Figure 6: Boxplots of phone durations of some common short and long vowels, for the three speaking rates (outliers not shown).

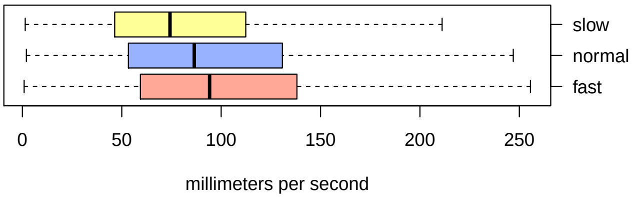


Figure 7: Boxplots of peak movement velocities of the three tongue sensors (outliers not shown).

vowels far from the close/front corner ([uː], [ʊ], [oː], [ɔ], [a]) and diphthongs exhibit high peak velocities (and thus appear close to the top of Figure 9).

Regarding the second factor, i.e., the influence of speaking rate on tongue target positions, we have gathered for each of the three tongue sensors (back, middle and tip of the tongue) the deviation from its average position, as shown in the boxplots of Figure 8. The figure shows each of the x, y and z coordinates separately, which correspond to the left/right, up/down and front/back directions from the speaker's point of view. A slight decrease in positional variability can be seen for increased speaking rate, suggesting that tongue movement needs to be reduced for faster speech.

These findings are in line with, e.g., (Flege, 1988), where increased speaking rate is reported to result from a combination of both increased movement velocity and decreased divergence of the tongue from a "centroid" or "rest" position.

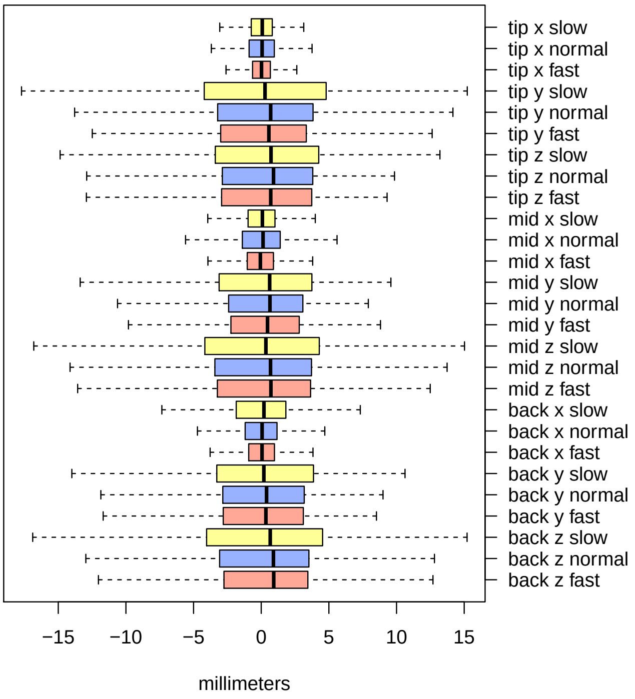
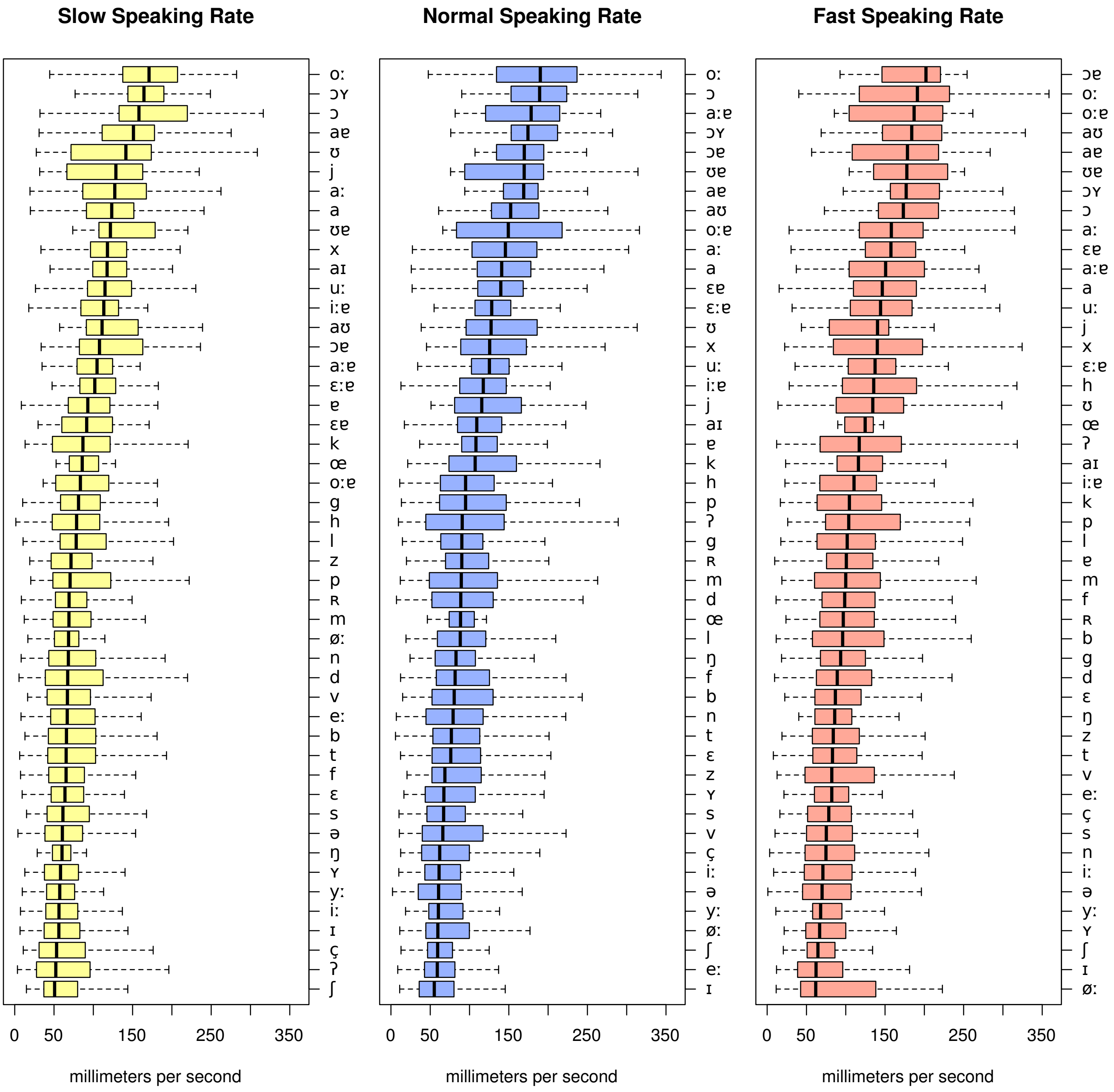


Figure 8: Boxplots of mean-normalized spatial coordinates of the three tongue sensors (outliers not shown).

Similar to (Yehia et al., 1998; Jiang et al., 2000; Beskow et al., 2003), we have also looked at how well the tongue motion data can be predicted from the facial motion data and vice versa. In a 10-fold cross validation setup, we have computed a linear regression to predict one tenth of the tongue (face) data from the corresponding face (tongue) data, where the other nine tenths of the data are used to estimate the predictor. Then Pearson's correlation coefficients are computed between the predicted and the originally recorded tongue (face) data. Note that we excluded the face markers on the eyebrows and eyelids for this analysis step because their movement can be expected to be unrelated to phone articulation. The average correlation coefficients resulting from this procedure are shown in Table 1. The results are comparable to the ones of the "Sentences, 3 coils" condition in (Beskow et al., 2003) (tongue from face: 0.525, face from tongue: 0.357), which is the condition most similar to our setup. It can be seen that prediction of tongue motion from face motion is more successful than prediction in the opposite direction. There does not seem to be a clear influence of speaking rate on the values in Table 1.

Speaking Rate	Face from Tongue	Tongue from Face
Slow	0.234	0.445
Normal	0.226	0.558
Fast	0.279	0.523

Table 1: Average Pearson's correlation coefficients between measured and predicted marker coordinates



(a) (b) (c)

Figure 9: Boxplots of peak movement velocities of the three tongue sensors per phone (outliers not shown), for (a) slow, (b) normal, and (c) fast speaking rates.

5. Summary and Conclusion

This paper presented the new MMASCS multi-modal annotated synchronous corpus of speech, which consists of simultaneously recorded facial motion capture data, electromagnetic articulography data, audio and video, of an Austrian German speaker reading the same corpus at normal, fast and slow speaking rate. The data is released free of charge for research purposes including phonetic labels, documentation as well as a 3D visualization software to play back the 3D data.

Our own future work with this corpus will include audio-visual speech synthesis with tongue modeling, as well as investigating the benefit of face and tongue data for (audio-only) synthesis of fast speech. We hope the corpus will prove useful also for other applications in the speech research community.

6. Acknowledgements

This research was funded by the Austrian Science Fund (FWF): P22890-N23. The Competence Center FTW Forschungszentrum Telekommunikation Wien GmbH is funded within the program COMET Competence Centers for Excellent Technologies by BMVIT, BMWFW, and the City of Vienna. The COMET program is managed by the FFG.

7. References

Jonas Beskow, Olov Engwall, and Björn Granström. 2003. Resynthesis of facial and intraoral articulation from simultaneous measurements. In *Proc. ICPhS*, pages 431–434, Barcelona, Spain.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359. http://sail.usc.edu/iemocap/.

Carstens Medizinelektronik. 2014. http://www.articulograph.de.

James E. Flege. 1988. Effects of speaking rate on tongue position and velocity of movement in vowel production. *Journal of the Acoustical Society of America*, 84(3):901–916.

Jakob Hollenstein, Michael Pucher, and Dietmar Schabus. 2013. Visual control of hidden-semi-Markov-model based acoustic speech synthesis. In *Proc. AVSP*, pages 31–36, Annecy, France.

Jintao Jiang, Abeer Alwan, Lynne E. Bernstein, Patricia Keating, and Ed Auer. 2000. On the correlation between facial movements, tongue movements and speech acoustics. In *Proc. ICSLP*, pages 42–45, Beijing, China.

Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang. 2008. Articulatory control of HMM-based parametric speech synthesis driven by phonetic knowledge. In *Proc. Interspeech*, pages 573–576, Brisbane, Australia.

Zhen-Hua Ling, Korin Richmond, Junichi Yamagishi, and Ren-Hua Wang. 2009. Integrating articulatory features into HMM-based parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(6):1171–1185.

Naturalpoint. 2014. http://www.naturalpoint.com.

Michael Pucher, Dietmar Schabus, and Junichi Yamagishi. 2010. Synthesis of fast speech with interpolation of adapted HSMMs and its evaluation by blind and sighted listeners. In *Proc. Interspeech*, pages 2186–2189, Makuhari, Japan.

Korin Richmond, Phil Hoole, and Simon King. 2011. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Proc. Interspeech*, pages 1505–1508, Florence, Italy. http://www.mngu0.org.

Dietmar Schabus, Michael Pucher, and Gregor Hofer. 2011. Simultaneous speech and animation synthesis. In *ACM SIGGRAPH 2011 Posters*, pages 8:1–8:1, Vancouver, British Columbia, Canada.

Dietmar Schabus, Michael Pucher, and Gregor Hofer. 2012a. Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis. In *Proc. LREC*, pages 3313–3316, Istanbul, Turkey.

Dietmar Schabus, Michael Pucher, and Gregor Hofer. 2012b. Sepaker-adaptive visual speech synthesis in the HMM-framework. In *Proc. Interspeech*, pages 979–982, Portland, OR, USA.

Dietmar Schabus, Michael Pucher, and Gregor Hofer. 2013. Objective and subjective feature evaluation for speaker-adaptive visual speech synthesis. In *Proc. AVSP*, pages 37–42, Annecy, France.

Dietmar Schabus, Michael Pucher, and Gregor Hofer. 2014. Joint audiovisual hidden semi-Markov model-based speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):336–347.

Studiobox. 2014. http://www.acousticbooth-studiobox.com.

Alan Wrench. 1999. The MOCHA-TIMIT articulatory database. http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html.

Hani Yehia, Philip Rubin, and Eric Vatikiotis-Bateson. 1998. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(12):23–43.

Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. 2006. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.