Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

Recording speech articulation in dialogue: Evaluating a synchronized double electromagnetic articulography setup

Christian Geng[a,*], Alice Turk[b], James M. Scobbie[c], Cedric Macmartin[b], Philip Hoole[d], Korin Richmond[e], Alan Wrench[f,c], Marianne Pouplier[d], Ellen Gurman Bard[b], Ziggy Campbell[b], Catherine Dickie[b], Eddie Dubourg[b], William Hardcastle[c], Evia Kainada[g], Simon King[e], Robin Lickley[c], Satsuki Nakai[c], Steve Renals[e], Kevin White[b], Ronny Wiegand[b]

[a] *Department Linguistik, Universität Potsdam, Germany*
[b] *Linguistics and English Language, The University of Edinburgh, UK*
[c] *Speech Science Research Centre, Queen Margaret University, Edinburgh, UK*
[d] *Institut für Phonetik und Sprachverarbeitung, LMU München, Germany*
[e] *Centre for Speech Technology Research, The University of Edinburgh, UK*
[f] *Articulate Instruments, Edinburgh, UK*
[g] *Technological Educational Institute of Patras, Greece*

ARTICLE INFO

Article history:
Received 9 October 2012
Received in revised form
24 June 2013
Accepted 8 July 2013

ABSTRACT

We demonstrate the workability of an experimental facility that is geared towards the acquisition of articulatory data from a variety of speech styles common in language use, by means of two synchronized electromagnetic articulography (EMA) devices. This approach synthesizes the advantages of real dialogue settings for speech research with a detailed description of the physiological reality of speech production. We describe the facility's method for acquiring synchronized audio streams of two speakers and the system that enables communication among control room technicians, experimenters and participants. Further, we demonstrate the feasibility of the approach by evaluating problems inherent to this specific setup: The first problem is the accuracy of temporal synchronization of the two EMA machines, the second is the severity of electromagnetic interference between the two machines. Our results suggest that the synchronization method used yields an accuracy of approximately 1 ms. Electromagnetic interference was derived from the complex-valued signal amplitudes. This dependent variable was analyzed as a function of the recording status – i.e. on/off – of the interfering machine's transmitters. The intermachine distance was varied between 1 m and 8.5 m. Results suggest that a distance of approximately 6.5 m is appropriate to achieve data quality comparable to that of single speaker recordings.

1. Introduction

Within the fields of speech science and speech technology there exists a tension between demands for data with a high degree of ecological validity and data reflecting the physiological reality of speech: Real language typically takes place in unscripted dialogue, but this kind of dialogue is hard to record experimentally. Considerable progress has been made in the development of techniques to elicitate spontaneous speech that allows the scientific study of linguistic phenomena without sole reliance on read speech (Anderson et al., 1991; Gravano, Benus, Chávez, Hirschberg, & Wilcox, 2007; Van Engen et al., 2010). Studies that simultaneously use such elicitation techniques in conjunction with methods used for the measurement of physiological aspects of speech production are, however, at best rare. In part, this is due to the fact that physiological methods measuring the behavior of the vocal tract during speech present higher administrative costs than do acoustic recordings, and that these administrative costs increase when several participants are to be recorded simultaneously.[1] Still, in our view, such an approach is tractable and data from such a combination have the potential to have strong contributions in heterogeneous disciplines such as speech pathology, speech technology, linguistics and psychology.

[*] Corresponding author. Tel.: +49 331 977x2578; fax: +49 331 9772087.
E-mail addresses: christian.geng@uni-potsdam.de, geng@uni-potsdam.de (C. Geng).

[1] Note that there also exists the possibility of a setup in which one speaker is recorded physiologically using EMA while engaged in a spontaneous dialogue with another speaker for whom perhaps only audio data exist. While probably sufficient for many research aims including most speech synthesis and recognition applications, this approach has its limitation for research topics like for example rhythmical entrainment between speakers or cross-speaker accommodation.

0095-4470/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2013.07.002

Currently, standard acoustic modeling for *automatic speech recognition* uses very little of available speech production knowledge. An increasing body of evidence suggests that knowledge of speech production mechanisms affords simple explanations for many phenomena observed in speech that cannot be easily analyzed from the acoustic signal or phonetic transcription alone. While appropriate machine learning methods for incorporating speech production systems into recognition systems are available (for an overview see King et al., 2007), few usable corpora containing acoustic and oral movement data exist: The X-ray Microbeam database (Westbury, 1994), the MOCHA-TIMIT corpus (Wrench & Hardcastle, 2000), and, more recently the mngu0 corpus (Richmond, Hoole, & King, 2011). Recent research on *speech errors* has revealed heretofore unknown articulatory properties of errors which may go undetected by acoustic or auditory evaluation; these have contributed to theories of the relationship between cognitive utterance planning and articulation (Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007; Pouplier & Goldstein, 2010; Pouplier & Hardcastle, 2005). Similarly, for research on *speech disfluencies*, electromagnetic articulography (EMA) data have the potential to uncover covert error and repairs, even during silence.

In this paper, we describe the setup of the Edinburgh Speech Production Facility (ESPF), and address three issues that potentially affect any multi-machine facility built for the purpose of acquiring speech data from multiple participants: (1) Communication among participants and experimenters, (2) synchronization, and (3) inter-machine interference. Some aspects of our approach to these issues are applicable to multi-machine, multi-participant speech data acquisition in general, while others are specific to facilities containing two Carstens' AG500 machines. For example, for labs involving an alternative system for electromagnetic tracking, such as the Wave system by Northern Digital (Berry, 2011), many aspects of the synchronization issues we address here will be identical, while others will differ slightly, since the synchronization between audio and articulation is key-frame based in the Wave system, in contrast to the Carstens binary coding of recording status using dedicated hardware. Issues of electromagnetic interference are also relevant for Wave users, but our approach is not directly transferable. This is because the data structure outputs by the Wave are very different from those outputs by the Carstens systems. However, our treatment of this topic will hopefully remind future researchers of the fact that resolving this issue is essential for the success of synchronized articulography research. And finally, some aspects of our experimental setup and protocols reflect our recording strategy and the stimulus materials we were aiming to acquire, in our particular recording context. For example, our decision to separate data into separate files grouped by task reflects our wish to acquire manageable chunks of data. And the fairly complex audio setup we describe here was required to elicit a broad cross-section of speaking styles within a single session, while avoiding electromagnetic interference.

The description and evaluation of such a setup comprise several steps: The following section (Section 2) gives a general overview of the facility installation as a whole and elaborates the need for a flexible audio capturing system including the possibility to manipulate the mutual audibility between participants as well as options for the experimenters to speak to participants. We refer to such a system as a talkback system that was implemented in addition to the participant's audio capture used for acoustic analysis. The subsequent sections deal with problems specific to the acquisition of synchronized articulography. A first section empirically evaluates the temporal synchronization of the EMA machines empirically by exemplifying data acquisition that simulates the recording situation by starting and stopping the EMA devices (Section 3). After that, Section 4 motivates the need to evaluate the electromagnetic interference between the two articulographs. The final section concludes.

2. Electromagnetic articulography and facility architecture

The objective of simultaneously recording articulatory data and the acoustic waveform of two speakers engaged in a cross-section of speech styles to a large extent dictates the general architecture of a laboratory such as the Edinburgh facility. Electromagnetic articulography (EMA) uses alternating magnetic fields generated at different frequencies by six transmitter coils. These fields induce alternating currents in up to 12 sensors. The amount of induced current is proportional to sensor–transmitter distances. This operation principle allows the calculation of sensor positions in a three-dimensional Cartesian coordinate system and two additional sensor orientations. The electromagnetic operation principle of the AG500 as just described imposes specific constraints on the design of a facility whose purpose is to simultaneously record articulatory data from two participants. Both machines generate electromagnetic fields at identical carrier frequencies and these magnetic fields must be guaranteed to not interfere with each other since this would compromise the quality of the measurement data. This problem can only be accommodated by placing the machines at an appropriate distance from each other. Both ourselves and the manufacturer had made estimations of the minimum distance necessary to obtain high quality data prior to project onset. The variability in these estimates was regarded as high and it was therefore decided that a more systematic exploration of the distance/interference function would be necessary. At the same time, this constraint of a minimum distance between the two AG500, together with the placement of the participants in separate booths due to acoustic reasons, makes it necessary to amplify the acoustic signal of participants in order for them to be mutually understandable, i.e. the setup calls for the implementation of a sophisticated talkback system. This requirement contrasts with the acoustics-only experimental setup realized in early Map Task studies (Anderson et al., 1991). The solution adopted by the current project was to separate the participants and place them in separate booths; therefore the developed talkback system required headphones for both participants and experimenters. Such a talkback system not only requires that participants can communicate with each other, but also that they can hear instructions given by the experimenter in the control room at the same time, and that they can also talk to the control room themselves. It was hoped that this move would allow smooth operation of experimental sessions, but it was also made for scientific reasons: such a flexible architecture allows for experimental designs which manipulate mutual audibility.

In fact, the materials acquired during the production of the Edinburgh ESPF tap the full potential of this possibility.

- Monologue tasks like story reading (Honorof, McCullough, Somerville, & last retrieved June 24, 2013), Wellsian lexical sets (Wells, 1982) and diadochokinetic tasks were acquired. These tasks require both participants to be mutually inaudible. Of course it would be possible to record the monologue passages one after the other. However, sensors glued to the speech organs have a limited lifetime, i.e. are subject to detachment after a certain period of time, and therefore an effective procedure is essential.
- The other extreme where the speech tasks require mutual understanding of participants is dialogue. In the context of the current project we recorded Spot the Difference picture (Van Engen et al., 2010; Van Engen, Baker, Choi, Kim, & Bradlow, 2007), Story-recall and Map Tasks (Anderson et al., 1991).
- In addition, the data collection undertaken in the context of the present paper also comprised asymmetric recording situations. For example there is the possibility of combining story recall and shadowing (Marslen-Wilson, 1973) by means of such asymmetric settings: speaker A tells a story; speaker B shadows speaker A, but is not audible to speaker A.

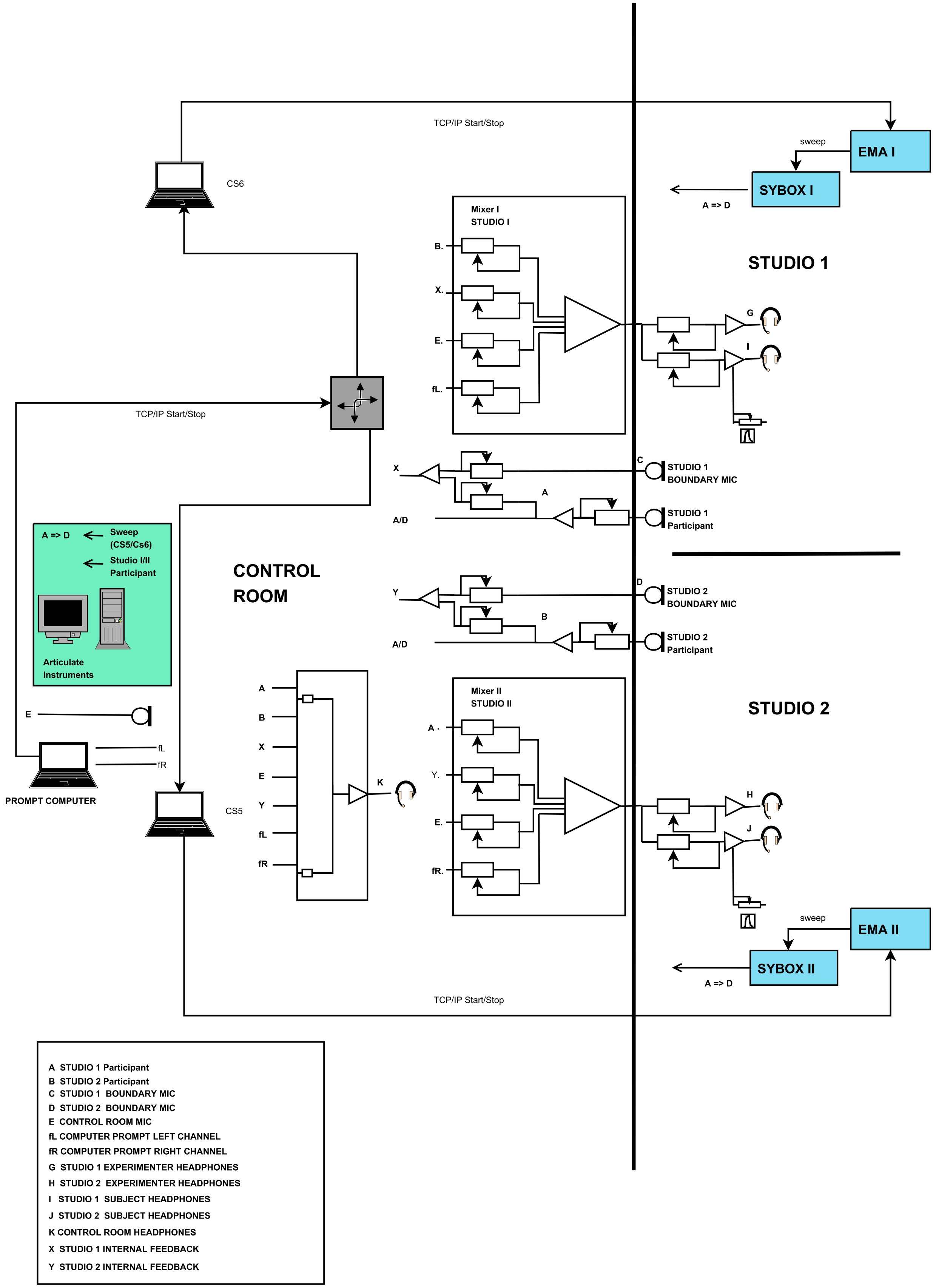

Fig. 1. Schematic of the speech production facility. Basic control flow. The setup consists of an audio talkback system and a synchronized system of two parallel Carstens AG500 units.

2.1. Talkback system

The full setup – omitting only representations of prompting screens and devices for experimental monitoring – is depicted in Fig. 1. In that figure, the part to the left of the bold vertical dividing line represents the control room area; this control room is spatially and acoustically separated from each of the booths. The recording booths are shown on the right part of the figure and are separated by a bold horizontal line that represents their spatial and acoustical separation. The signal in each of these booths is picked up by two types of microphone, (i) directional microphones (Studio 1 Participant (A), Studio 2 Participant (B)) and (ii) omnidirectional microphones (Studio 1 Omnidirectional Mic (C)/Studio 2 Omnidirectional Mic (D)). The directional microphone signals are directly fed into the A/D and are primarily used for further scientific analysis. In addition they are added to a mix containing the signals picked up by the omnidirectional microphones which primarily pick up the studio booth ambience for the talkback system but serve no further scientific purpose. This mix is referred to as “internal feedback” and labeled X and Y in Fig. 1 for Studio 1 and Studio 2 respectively. In addition to participant microphones, the microphone for the experimenter seated in the control room is labeled E. The final sound source is the acoustic prompt signal of the computer prompt, with left channel being labeled as fL and the right channel as fR. There is one (sub-)mixer per studio located in the control room. Fig. 1 shows them as “Mixer Studio I” and “Mixer Studio 2” respectively. These mixers serve the purpose of generating the desired mix for each experimental condition—also in consultation with the participants. As an example “Mixer Studio I” receives the following signals: The participant's signal A, the Studio 2's internal feedback Y, the signal of the control room microphone A and the one channel of the prompting computer's (mono) signal fR. Mixer Studio 2 is set up equivalently; it receives signals from the participant in Studio 2, the internal feedback signal of the other studio, the signal from the control room microphone and one channel of the prompting computer's (mono) signal (A, Y, E, fR respectively). In addition to the mixers for the two studios, there is a (master-)mixer in the control room. This mixer receives the microphone signals from both studios (A and B), the Internal Feedback signals from both studios (X and Y), the control room microphone (E), as well as both channels of the acoustic prompt and outputs this signal to the experimenters’ headphones.

The same functionality can be implemented in hardware by selecting from a wide range of available audio equipment. In order to give a detailed account of the recording hardware used in the setup of the ESPF, all essential pieces of equipment are listed in a separate Appendix A.

This setup allows the experimenters in the control room to arbitrarily route signals from any source – control room, participant, the experimenter herself – to any destination. Such flexibility turned out to be vital for the design of our study in several respects:

- Consider for example situations where participants need standard instructions for a dialogue task to be carried out. In this situation, it is often helpful to provide both participants with a standardized set of instructions. This can be achieved by routing one experimenter's audio signal to all possible destinations. Once task instruction is completed, the signal from the experimenter's microphone is no longer necessary, and even has the potential to disturb the participants. The setup just described can flexibly adapt to the new situation by control room experimenters subtracting their own audio signal from the participant headphones. The Control Room Microphone (E) is represented in Fig. 1 in both studio-specific mixers (Mixer I STUDIO I and Mixer II STUDIO II) pointing to this possibility of addressing the participants in each studio separately.
- Also different speech tasks may require different settings concerning the mutual audibility of participants. These heterogeneous demands were already mentioned above. In Fig. 1, this possibility to manipulate inter-booth audibility is reflected as “B Studio 2 Participant” in Mixer I STUDIO I and as “A Studio 1 Participant” in Mixer II STUDIO II.

2.1.1. Piezoelectronic headphones

The magnetic coils used to move the speaker diaphragms in standard headphones pose a risk of electromagnetic interference when used within the EMA cubes. We therefore replaced the moving-coil speakers in a standard Phillips closed headset with a Piezoelectric Murata VSB50EWH0301B sounder. These speakers use the Piezo principle, whereby an electric charge is applied across a thin layer of piezoelectric material (in this case quartz) which causes the material to contract; the alternation of charge creates alternating shrinkage and expansion of the material which in turn drives the alternating movement of the speaker diaphragm. Known disadvantages of these speakers (low amplitude, poor frequency response) were partially offset by a headphone amplifier with tone control. The amplifier boosted the voltage used to drive the speakers, and the tone control boosted the amplitude of selected frequencies in order to improve intelligibility.

3. Synchronization of the EMA machines

Fig. 2 sketches the control flow during a parallel EMA experiment. This sketch effectively is a subset of the full laboratory setup already shown in Fig. 1 limited to aspects relevant for machine synchronization issues.

A central prompting computer issues commands which tell two specialized computers managing the AG500 recording procedure (control servers, labeled CS5 and CS6 in Fig. 1, one for each EMA machine) to change the recording status of the two EMA machines. This control side of the system is implemented via the TCP/IP protocol, with the signal traveling from the prompt computer to the EMA systems (EMA I and EMA II in Figs. 1 and 2) via a router and the control servers mentioned above. Apart from the TCP/IP streams the AG500 also comes with a synchronization device called “SYBOX”. The function of the SYBOX is to emit the system's timing and status information (trigger- and pretrigger signals, recording status information). They are shown as SYBOX I and SYBOX II in Figs. 1 and 2. They are the key synchronization devices as they allow us to determine the exact start and stop times of both AG500 machines. There are at least three sources of latencies conceivable on the control side of the setup: First, the prompt computer cannot send the commands to change the recording status to both machines absolutely simultaneously. Rather, the functionality provided by the manufacturer consists of a two-step procedure that minimizes the latencies between the machines. The first step is to prepare both machines separately to receive a recording status change command from the prompt computer (by sending `click` via TCP/IP); the change status command is executed separately in a second step for both machines by sending `go`. After successfully changing the recording state, both connections can be closed.

Second, latencies can also be generated by the network itself. As shown in Fig. 2, the prompt computer communicates with the control servers via a TCP/IP router which is part of a local subnet of the intranet. Third, the network hardware was not explicitly designed to minimize network latencies. Apart from diagnosing the differences in the *relative* timing of the two AG500 machines, there is another question which deserves to be answered: It is not clear whether the internal clocks of two EMA machines have the tendency to diverge over the course of the long trials that can be anticipated when recording dialogue speech. We aim to (i) specify which of these issues are solved by the setup approach taken and (ii) to give a quantitative account of the severity of the remaining problems. The empirical approach by which the data are analyzed here is to capture the sweep signal generated by the

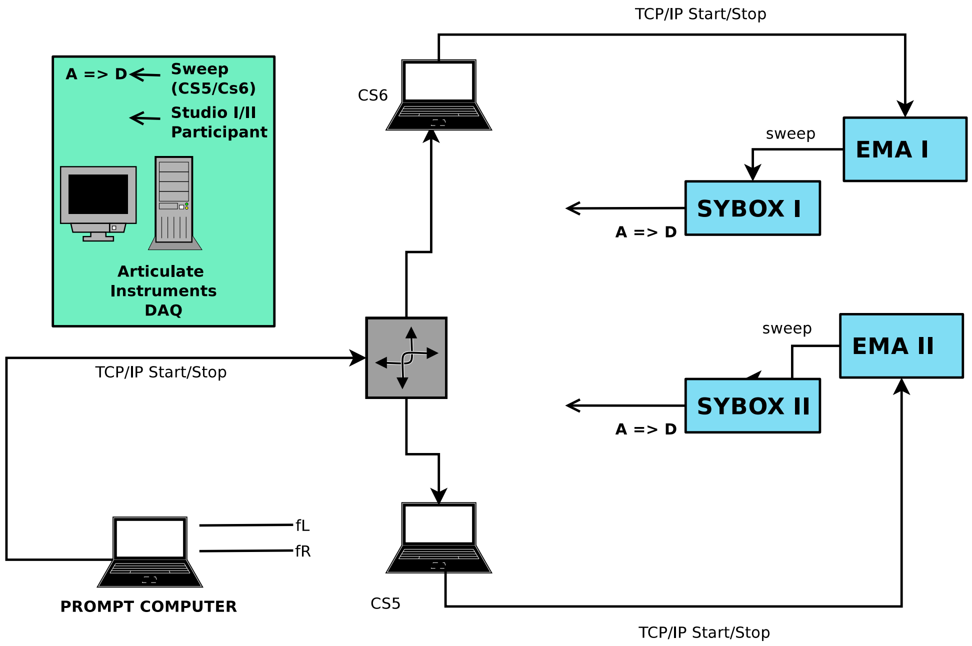

Fig. 2. Setup of the EMA facility: Basic control flow of the EMA subsystem.

EMA machine's central unit, the LIDA (Linux Integrated Data Acquisition), and emitted through the SYBOX: The sweep signal is a rectangular pulse that indicates whether the AG500 machines are recording or not, effectively encoding binary recording status by TTL voltages. The sweep signal was captured by means of an Articulate Instruments data acquisition (DAQ) system: On the hardware side, the cables carrying sweep signals were connected to an 8+4 Channel Analogue/Video Breakout Box (BRK1) manufactured by Articulate Instruments. The actual A/D conversion was carried out by an ADLINK DAQ-2213 8-channel, 16-bit differential input data acquisition A/D card mounted in a standard PC. The same system was also used to capture the speech acoustics from both speakers (see Fig. 1).

The sampling frequency was set to 32 kHz. The captured data in turn are used to extract the rising and falling flanks of the sweep signal synchronization impulse emitted by the SYBOXes, and allow determination of the exact start and stop times of both AG500 machines.

3.1. Machine speeds

First, the stability of the relative timing of the AG500 was evaluated. For this purpose, a single sweep of maximum duration was recorded and captured by the method described in the previous paragraph. Note that the AG500 currently is capable of recording a maximum of 65,535 samples (approximately 328 s) at 200 Hz sample rate. One of the relevant aspects is to check whether after completion of the simultaneous sweeps, very similar durations are reported for both tracks of the synchronization data captured by the data acquisition system. In the case of significant differences, it would have to be concluded that both EMA machines run at different internal speeds. To check whether these problems are present, (a) the number of samples recorded by the AG500 units with the recording duration set to the maximum and (b) the corresponding duration of the synchronization data were compared. Here, the maximum number of 65,535 AG500 samples recorded within corresponded to 65,535.8812 and 65,535.8187 AG500 samples in the extracted synchronization data. We consider the difference of 0.0625 AG500 samples (=0.3125 ms) over the maximum trial duration as negligible and that therefore both EMA machines run at fairly consistent speeds. A related, second question concerns the comparison of the machine speeds of the DAQ system and the EMA machines. In order to understand this analysis, consider the acquisition of one second of EMA data using the setup in Fig. 2. Given the sample rate of 200 Hz, this ideally should amount to 200 EMA samples and 32,000 samples of data acquired by the DAQ system. However, if the hardware clocks of the EMA machines and the DAQ system are different, there will in practice be divergences that are linearly increasing as acquisition time increases. Conceptually, this kind of desynchronization can be seen as a linear stretch or compression of the time axis of one of the data modalities relative to the other. In practice, this stretching/compression of the time axis can be corrected by replacing the nominal sample rate by an empirically justified one accounting for this divergence.

We demonstrate this linearity in Fig. 3 by showing typical patterns for one machine in a dual recording carried out during the run time of the current project. Correlations and *R*-Squares of 1 verify that the linear adjustment of sample rate is well motivated in the context of our setup.[2]

3.2. Quantification of relative onsets asynchronies

The second question deals with the quantification of the *relative* onsets of rising and falling flanks of the sweep pulses. The aim of this section is to demonstrate that these temporal misalignments are tiny, negligible, and unimportant. This issue is, at least at first glance, more closely related to likely research questions of the current project than the characterization of hardware speeds dealt with in Section 3.1. For example, if timing between speakers is controlled – like in turn-taking (e.g. Wilson & Wilson, 2005) or rhythmical entrainment (e.g. Cummins, 2009) – then timing problems between AG500 machines

[2] As an anonymous reviewer points out, some of the points in the figure are slightly off-diagonal. However, this is not a graphing problem, also regression analysis and plotting are done with full numerical precision (32 bit floats), and the value for the correlation is in fact 1. We are treating these deviations as residual system inaccuracy due to unknown factors. Note that the maximum residuum of the linear regression plotted is in the microsecond range (3.474165e−05 s) and meaningless in practice. Also note that for our purpose we only need to show that (i) the drift is linear, (ii) the residual is not correlated with the total recording duration, and (iii) the residual is practically meaningless. In the current example, taking the EMA sample rate as the gold standard, the DAQ sample rate would have to be adjusted to 32,000.43 Hz.

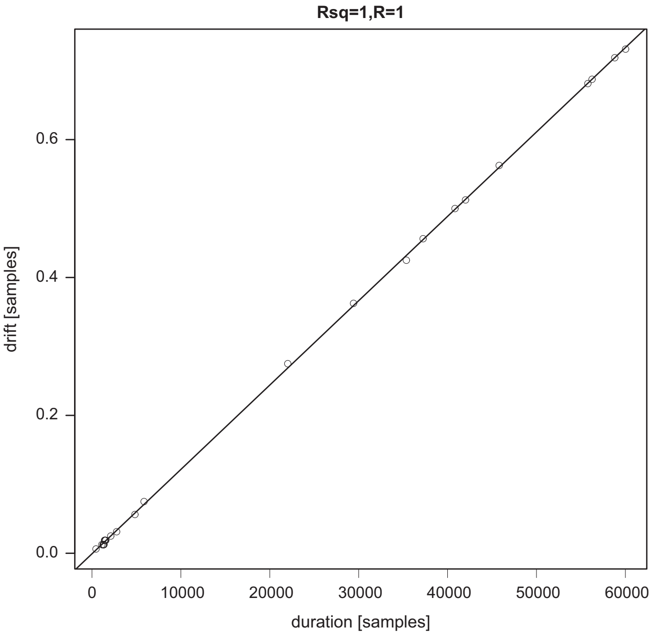

Fig. 3. Desynchronization of AG500 LIDA and DAQ systems as a function of acquisition time. Both abscissa and ordinate are expressed in EMA samples (1/200 s=5 ms).

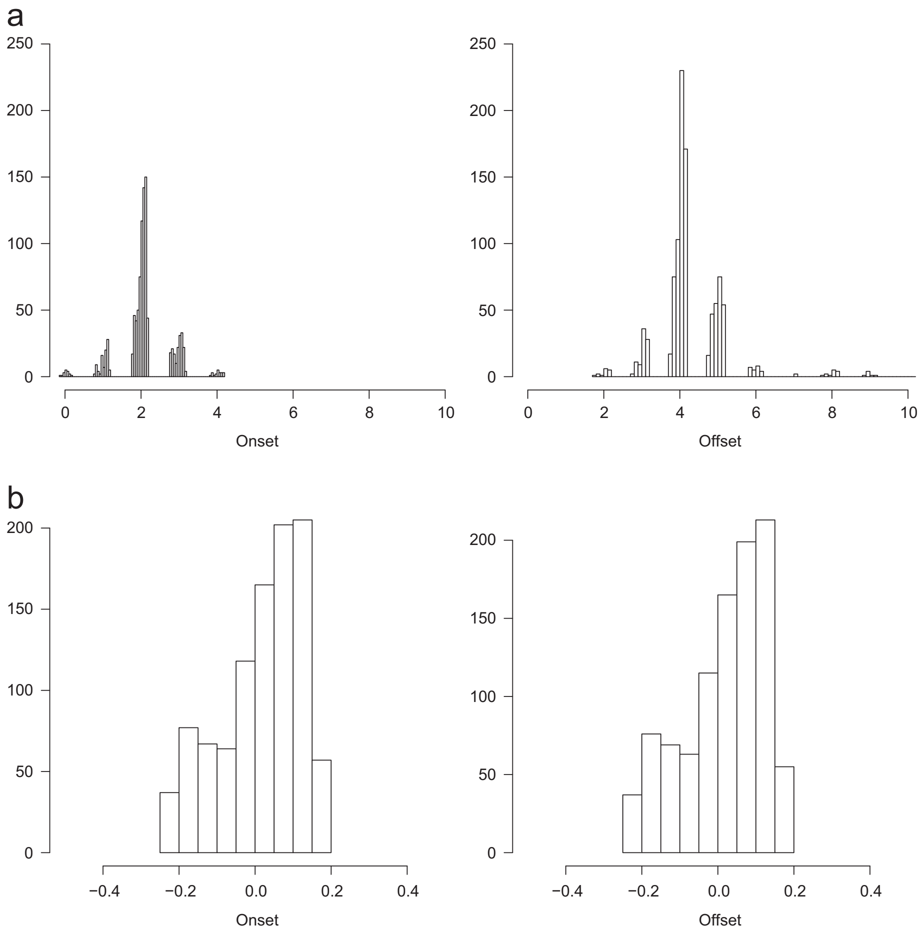

Fig. 4. Summary of relative latencies of the synchronization impulses as measured by the DAQ-device. Top panels: Histograms of raw latencies expressed in EMA samples (eq. 1/200 s=5 ms). Bottom panels: Histograms of latencies after the removal of whole sample contribution. Left panels: Data for trial onset pulses. Right panels: trial offset impulses. (a) Synch. Latencies [in EMA samples] and (b) Synch Latencies after correction [in EMA samples].

would directly result in measurement noise of the dependent variable. Therefore it seemed to be advisable to also collect data on the relative onset asynchronies between the AG500. For this purpose both AG500 machines were started and stopped simultaneously. We recorded 1000 trials between 1 and 10 s, i.e. 100 trials each. The extraction of the synchronization information is equivalent to the one used in the previous section.

Fig. 4, top panels, shows the histograms of the relative lags in start times for the two machines. The unit on the abscissa corresponds to the duration of one EMA sample. First there are considerable lags between the time of starting/stopping the first machine and starting/stopping the second one. A second observation is that these lags are considerably larger for the stopping commands (median: 4.09 AG500 samples) than they are for the starting commands (median=2.06 AG500 samples). The most striking observation though is that the lags are clustered around integer-valued EMA sample durations, but that the variances within these clusters are relatively small, i.e. there is no overlap between the integer-valued durations. This semi-quantized pattern suggests that there are several heterogeneous sources for the intermachine asynchronies, and that the largest part of the variance by far originates in full-sample misalignments of the start/stop pulses of the two AG500s. It is likely that these larger misalignments of EMA sample magnitude originate in the software-based subsystem: As already discussed and shown in Fig. 2, the prompt computer sends TCP/IP commands to change the recording status to the two AG500 units via the router and the control server notebooks. If this is correct, it should be legitimate to correct for these misalignments by padding leading and trailing chunks of speech where necessary. The effect of such a padding is shown in the lower panels of Fig. 4. The most striking result is that the probability densities look almost identical for the pulses starting the EMA systems and those stopping the systems. Without having a causal hypothesis, this makes it likely that they originate from the same underlying mechanism. Apart from that, it is noteworthy that the temporal misalignment after this whole sample correction is rather negligible with a mean of 0.0146(0.073 ms) and a median of 0.0375(0.1875 ms) AG500 samples. The worst case was a misalignment of a little more than 20% of an EMA sample (0.225, or 1.125 ms). In sum, it seems justified to apply a whole sample padding to the data. In a first step, for each file, we determined *n*, the number of samples mismatch between the machine started first and the machine started second. In a second step we have made *n* copies of the first data sample of the machine started second and appended it to the beginning of the file. An equivalent (respecting sample rate) procedure was applied to the audio data.

4. Electromagnetic interference

The AG500 system consists of six transmitter coils arranged spherically. These six transmitters are driven by different carrier frequencies ranging from 7.5 to 13.75 kHz (7.5, 8.75, 10.0 , 11.25, 12.5 and 13.75 kHz respectively). Each of these transmitters electromagnetically induces a current in up to 12 sensor coils. The voltage measured at the sensors varies as a function of the distance from the transmitter coils and the sensors' orientation in the field. The AG500 quantizes these induced voltage values (aka “amplitudes”) at 16-bit resolution and quantifies them as a pseudo unit termed “digit” (dig). The estimation of Cartesian sensor positions and rotations utilizes the proportionality between induced current (“amplitude”) and distance from the transmitter by means of nonlinear optimization (e.g. Hoole & Zierdt, 2010) or other tracking techniques like Particle, or Kalman filters. However, the most essential point to emphasize for the present purpose is that the carrier frequencies of the transmitter coils – in contrast to the predecessor machine, the AG200 – cannot be adjusted. This gives rise to the possibility that each machine in fact measures a mixture of its own transmitters' amplitudes and those of the other, interferent machine. As mentioned, these amplitudes form the basis for the estimation of the desired positional and rotational parameters. Therefore intermachine electromagnetic interferences have the potential to pose a serious threat for the reliability of the data measured by the facility. Note that it would in principle be possible to overcome the problem of electromagnetic interference between the machines by using a heterogeneous setup, i.e. using different motion capture systems for each of the speakers. While such an alternative second system is commercially available at the time of writing – the *Wave* system by Northern Digital (see Berry, 2011; Kröger et al., 2008) – we currently have insufficient knowledge about its principles of operation. Also, in the particular case of the Edinburgh facility, the *Wave* was not available at the time when it was established.[3] In the following we will aim to quantify the magnitude of this intermachine interference. The next sections present the measurements and the procedures that were made at an attempt of an evaluation (Sections 4.1 and 4.2), thereafter the analysis and results of this evaluation are presented (Section 4.3).

4.1. Experimental setup

As shown above, The Edinburgh facility was designed to have separate recording studios housing one of the two AG500 each, and a control room for the coordination of activities in the studios. The evaluation of the severity of interference was carried out in the facility itself, by varying the distance between the machines in the studios. A sketch of its geometry is shown in Fig. 5.

The two studios (STUDIO 1 and STUDIO 2 in Fig. 5) are of almost identical size (480 cm × 280 cm). They are separated by a wall of 100 cm cross section.

The AG500 carrier units are approximately quadratic and 80 cm wide (see the inset at the bottom of Fig. 5). Therefore, the maximum distance between the machines that can be achieved when moving them along the long side of the wall in each of the booths in theory would amount to 900 cm (=480 cm+480 cm+100 cm)−(80 cm+80 cm). The minimum distance between the machines is determined by the separating wall and amounts to 100 cm. Preliminary estimates of the mutual influence of the two machines that were provided by the Carstens Medizinelektronik at the time of purchase suggested a substantial amount of interference at 5 m, and a small amount at 8 m and 10 m distances. In order to arrive at a more comprehensive picture, we decided to analyze a dataset comprising the range of possible distances between the two machines. This intermachine distance serves as the main independent variable and was manipulated in five steps. The guiding principle of the analysis is to measure the signal amplitudes generated by one machine with the receiver unit of the other thus having one machine generating interferences measured by the other – and vice versa. The AG500 system offers the (undocumented) possibility to change the transmission status for all the transmitter coils simultaneously between on and off.

The dependent variable that will be analyzed in the following section is derived from the so-called complex amplitudes which are an intermediate product in the processing chain: The AG500 system generates its signal amplitudes from raw data by demodulation: Each of the six transmitter coils emits a “carrier-” signal in the VLF (Very Low Frequency) range which is modulated by movements of the receiver coils in the measurement field. In order to simultaneously use multiple transmitters at high temporal resolution, the system permanently emits six different carrier frequencies. The contributions of the six transmitters are extracted by a demodulation method which results in signal amplitudes. These amplitudes are complex at first, contain both real and imaginary parts corresponding to amplitudes and phases, and it is these complex amplitudes in the *z*-plane that serve as the basis for any further analysis of intermachine interference. The advantage of using complex amplitudes instead of the real part of the amplitude only is

[3] However, a setup consisting of heterogeneous hardware is disadvantageous due to other reasons: protocols for data post processing would have to be established for different kinds of devices independently. In addition, the choice of EMA machine should not affect the data, but in practice it is plausible that it does, for example due to coil and wire differences and machine specifications.

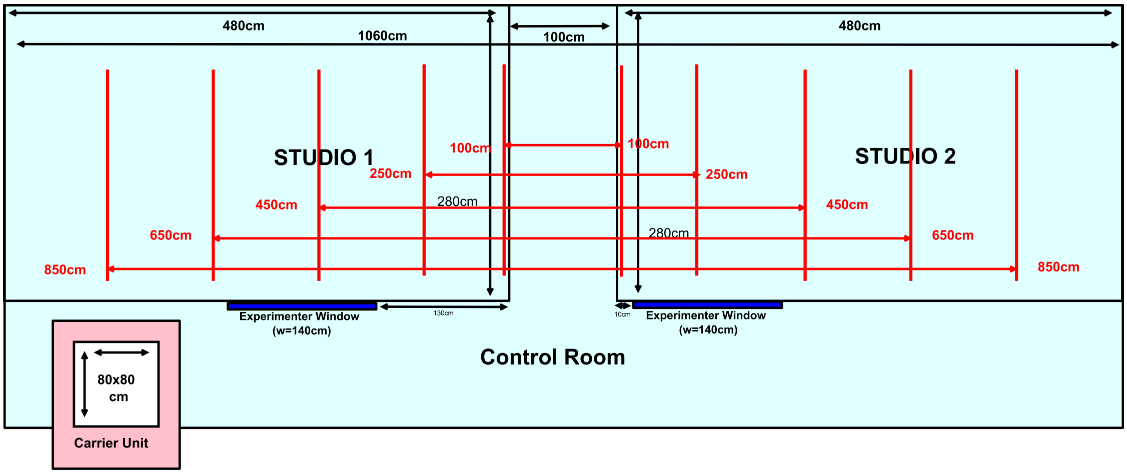

Fig. 5. The figure sketches the studio geometry and the positioning of the machines relative to each other in the experiments evaluating the severity of electromagnetic interference.

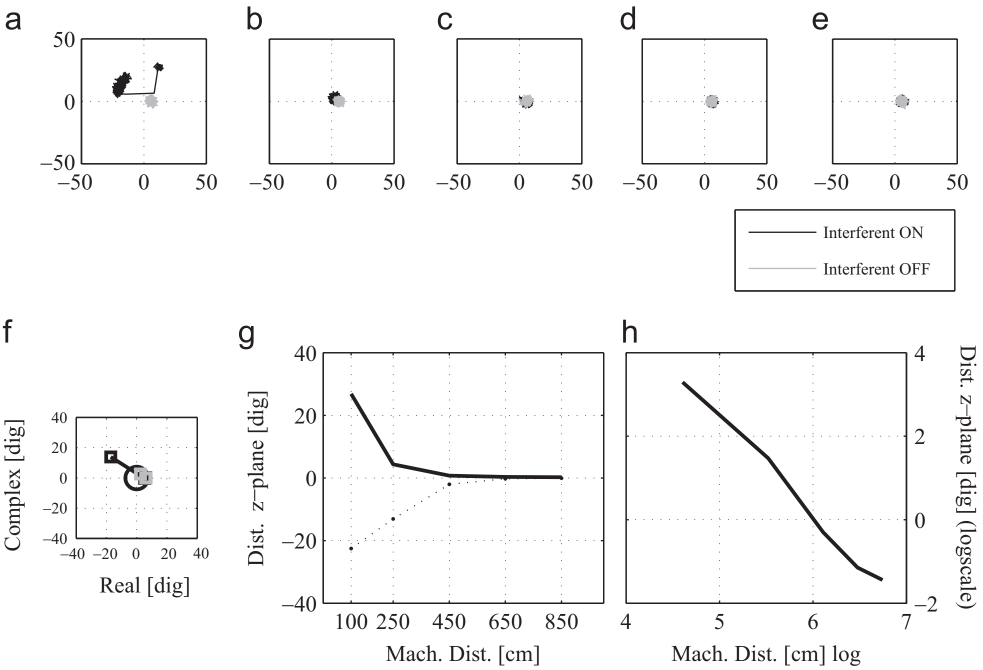

Fig. 6. Illustration of analysis of intermachine distances. Top panels (a)–(e): Displays of complex amplitudes as a function of distance between interfering and acquisition machines. The abscissa shows the real, the ordinate the imaginary part of the complex amplitude signal (both in dig). Bottom left (f): Averaged complex amplitudes in the *z*-plane; (g) decay of averaged complex distance as a function of intermachine distance (solid line) and first derivative; (h) same data as in (g), but linearized by taking the log of both intermachine distance and distance in the *z*-plane. (a) 100 cm, (b) 250 cm, (c) 450 cm, (d) 650 cm, (e) 850 cm, (g) Coil 01 Trans 1 and (h) Coil No. 01 Trans 1.

that of increased sensitivity: Interferences can not only be reflected in the signal amplitudes, but can also result in phase distortions that would not be captured otherwise.

4.2. Procedure

Fig. 6 illustrates the rationale of this analysis: The top five panels (a–e) give an example of raw complex amplitudes at different intermachine distances, ranging from 100 cm (a) to the maximal distance of 850 cm (e). The data acquired consist of static recordings acquired by placing 12 unused sensors in the manufacturer's calibration cartridges. Each panel shows two configurations, (i) with the interfering machine ON coded in black and (ii) with the interfering machine OFF coded in gray. The transmitters of the machine used for acquisition are always ON. With increasing distance, the bivariate distributions in both become increasingly similar. These patterns persist when the complete bivariate distributions are condensed to their mean value between interfering and acquisition machines in panel (e). The next step consists in transforming the complex amplitudes to the Euclidean distances between conditions in which the interfering machine was ON to the corresponding condition in which the interfering machine was OFF (g). The final transformation consists in a linearization of these patterns. By analogy to the distance voltage function of the old 2D system – see e.g. Hoole (1993) for details on the magnetic field functions – it makes sense to take the log of both measured amplitudes and distances between the two EMA machines to achieve a linear relationship, which in turn allows us to apply linear modeling techniques. Panel (h) gives an example for the type of linear relationship between predictors and criteria. Sometimes the patterns of decay did not conform to the expected exponential decay in Fig. 6(g,h). When this occurred, the whole set of five observations for that particular sensor/transmitter pairing was considered invalid and discarded from further

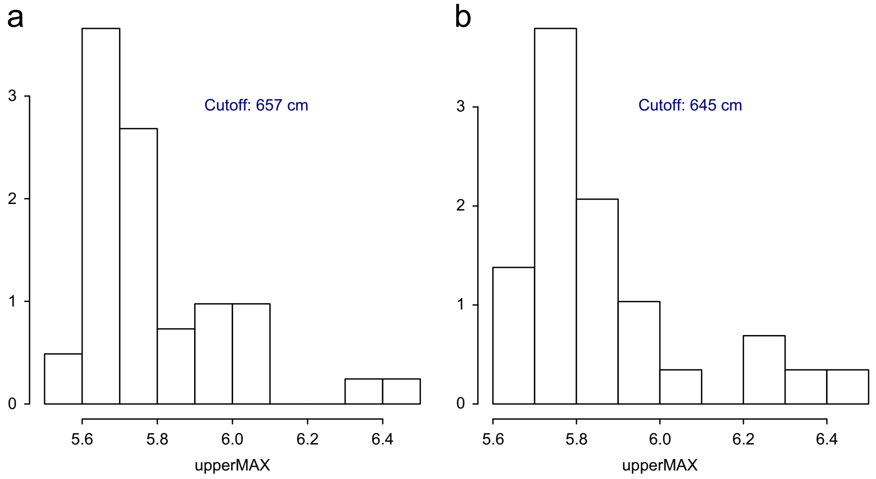

Fig. 7. The top left panel illustrates the rationale of the analysis: The cutoff reported is the distance at which the modeled data fall below the noise level. The bottom two panels display histograms of cutoff values modeled for each sensor–transmitter pairing, for each of the EMA devices separately. (a) EMA 1-Cutoff and (b) EMA 1-Cutoff.

analysis. In order to be able to determine a distance at which observed intermachine interference can be considered negligible, a baseline noise level is required. This noise level criterion was extracted from the data as follows: For each of the five distances, the mean distance over all samples per sensor–transmitter combination was calculated. In a second step, the standard deviation of these observations was calculated and subtracted from the data. This resulted in one noise level estimate for each of the five intermachine distances. From these, the minimum was selected as the final cutoff value. The determination of these noise floors was carried out independently for each of the two EMA devices; their numerical values were fairly similar amounting to 1.685 and 1.7531 digits.

4.3. Analysis and results

These data were analyzed by means of Linear Mixed Effects Models (e.g. Baayen, Davidson, & Bates, 2008). Unlike classical Generalized Linear Models, Linear Mixed Effects contain random effects in addition to the usual fixed effects in their linear predictor. All analyses described in this section were carried out using the programming language R (R Development Core Team, 2010), the Mixed Effect Modeling was carried out using the `lmer` function contained in the lme4 library (Bates & Maechler, 2010).

In addition to the Fixed effects, Linear Mixed Effects Models are capable of explicitly modeling random effects on slope, as well as on intercept. The design of the analysis was such that the log of distances of complex amplitudes in the *z*-plane functions as the dependent variable, and the log of the five levels of intermachine distances as the fixed factor. In addition to this fixed effect design, we calculated separate random adjustments of both intercepts and slopes for each sensor–transmitter pairing. Both parameter estimates of fixed and random effects were in a subsequent step used to calculate predicted values for each sensor–transmitter pairing. Thereby the contribution of the fixed effect stays constant, whereas this fixed effect is additively adjusted by the random contributions of the sensor–transmitter pairings modeled by an intercept and a slope each. This in turn allows one to calculate modeled interferences at arbitrary distances using model estimates. This was carried out at 1 cm intervals between 50 and 850 cm (log transformed) for each of the transmitter–sensor combinations. The final step consisted of determining for each of these values the distance at which this value fell below the log of the noise threshold defined above. These distances, transferred back into cm, present the final result of the analysis, and are summarized in the lower two panels of Fig. 7.

The cutoff points that are specific for each sensor–transmitter combinations are displayed as histograms, separately for the first AG500 (left panel) and the second device (right panel). In order to make reliable measurements, there must be no interference detectable. In other words, the maximum distance at which interference can occur – the worst case scenario – has to be considered the decisive criterion. These worst cases are also shown as text insets in each of the subplots, and amount to 657 and 645 cm for the two machines. The precise figures probably will depend on the exact AG500 devices, and also will in part vary with the physical properties of the rooms where they are set up. Still, we hope that this kind of information still mostly generalizes across machines, and therefore will be helpful for other laboratories setting up the same or similar hardware. Regardless of this issue, these results have repercussions for the setup of the Edinburgh facility. Necessary intermachine distances of approximately 650 cm allow us to satisfy the competing constraints that demands a fair distance from the rear studio wall – in our case a little more than 1 m.

5. Summary and discussion

In recent years, an increasing amount of work aiming at the validation of methods for speech motion research has been published in the speech production literature. For example, these have been dealing with algorithmic details of head correction (Kroos, 2012) and the process of position estimation and additional techniques to improve the accuracy of measured data (Hoole & Zierdt, 2010; Kroos, 2008). The position estimation issue was also extensively researched in the context of the current project. In particular, Korin Richmond developed an algorithm based on the unscented Kalman filter. In addition, the conversion of amplitudes into positions was also carried out by the method detailed in Hoole and Zierdt (2010), i.e. the TAPAD toolbox.[4] While the former has advantages over TAPAD in terms of computational efficiency, of greater importance for this project was the fact that it allowed us to compare two different solutions for the position estimation problem using heterogeneous formal approaches. Such an algorithm-independent perspective on raw articulatory data greatly to facilitates the interpretation of such data.

[4] Available at http://www.phonetik.uni-muenchen.de/~hoole/articmanual/index.html.

In contrast to these more general aspects tied to the particular acquisition technique used, the conceptual part of the present contribution identified specific problems associated with the setup of a facility designed specifically to acquire dialogue speech by means of two synchronized Carstens' 5D Electromagnetic Articulograph (EMA) systems and acoustic data. These were identified as (i) the synchronization of the devices and (ii) the distance between two identical EMA devices. Here, the co-registration and therefore the synchronization of different acquisition techniques are common throughout the psychological sciences, and its evaluation and the demonstration of the feasibility of the dual EMA approach were relatively straightforward. Also, the interference problem turned out to be influential on the design of the facility as a whole: It influenced basic design decisions of the facility, like the architecture, as the studios had to be built at a certain minimum size. It also had the consequence of making the design of a complex talkback system necessary, and had influences down through to the last detail like e.g. the design of custom piezoelectric headphones.

Concerning the timing of the EMA devices our results suggest that the amount of desynchronization of the devices is by no means linguistically relevant. With respect to the issue of electromagnetic interference between the devices, the optimal location of the machines is a mild compromise between intermachine and wall distance. However, results described in this paper as well as results from position estimation suggest that in comparison to single machine recordings, we only have to anticipate minor deterioration of data quality, if at all. Finally, data visualization, annotation and analysis are possible through the use of Articulate Instruments Advanced software, and data collected at the facility are stored in the data archive as detailed below.

5.1. Data archive

The project funded the development of custom-built data archive software for the facility. This software was created by Kevin White, and enables us to organize and access all relevant files and meta-data associated with any type of recording session made in this facility. This archive will be used to store all data collected from the facility. It enables files to be made accessible to appropriate groups, e.g. the experimenter, others associated with the facility, and/or the public according to the participants's and experimenters's wishes. In this way, it supports ethical aspects of data control. The project's dialogue sessions are called the DoubleTalk Corpus, and are available free of charge at the University of Edinburgh (for details see http://espf.ppls.ed.ac.uk/). The archive includes information about participants (e.g. dialect, age, scores on digit span and empathy psychometric tests, etc.). It also includes an indication of data quality, which relates to the success of the data post-processing algorithms, and to sensor detachment and/or malfunctioning.

5.2. Data visualization, annotation and analysis

The project also funded the purchase of advanced multichannel data capture, presentation and analysis software. This software was customized to the specific requirements of the project. The Articulate Assistant Advanced (AAA) application is commercially available and has been used successfully for the analysis of several pilot projects: The application is user-friendly and makes it possible for researchers with limited or no programming experience to display and analyze data from the facility. This includes synchronized recording and analysis of AG500 EMA data, EPG, audio and other analogue signals such as laryngograph. Although not part of the facility, the software is also capable of recording and analyzing ultrasound, video and 3D VICON camera tracking data.

Acknowledgment

This research was funded by EPSRC grants to Alice Turk (EPSRC Reference: EP/E01609X/1) and James M. Scobbie (EP/E016359/1).

Appendix A. Recording equipment

The inventory list for implementing the talkback system is given in Table A1.

Table A1
Inventory list implementing the talkback system as described in the main text.

Location	Produce
Studio booths (×2)	Neumann KM 100 (modular system) Neumann Capsule 31 (omnidirectional) Axia Microphone Audio Terminal RedBox RB HeadphonePreamp HD-2 Btech BT 928 Mic Preamp ArtCessories HeadAmp 4 – Mic Preamp K&M Round Base Mic Stand AKG SE300B Power Module AKG SA60 Mic Holder AKG CK98 Microphone AKG H30 shock absorber Vivanco 21472 wireless headphones, 2.4 GHz Custom-built Piezoelectric Headphones (Murata VSB50EWH0301B sounder)
Control system	Blade Server Dell Power Edge R300 Preamp Focusrite Octopre MK II RME ADI – 192 DD RB – DMA2 Soniflex Mic Preamp Axia 8x8 AES/EBU Audio Node BeyerDynamic DT 290 Headsets Axia Keypad Control Box

References

Anderson, A. H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Language and Speech*, *34*, 351–366.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. R Package Version 0.999375-34. URL: 〈http://CRAN.R-project.org/package=lme4〉.

Berry, J. (2011). Accuracy of the NDI wave speech research system. *Journal of Speech Language and Hearing Research*.

Cummins, F. (2009). Rhythm as an affordance for the entrainment of movement. *Phonetica*, *66*(1–2), 15–28.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, *103*, 386–412.

Gravano, A., Benus, S., Chávez, H., Hirschberg, J., & Wilcox, L. (2007). On the role of context and prosody in the interpretation of okay. In *45th Annual meeting of the association for computational linguistics* (*ACL*). The Association for Computer Linguistics, Prague, Czech Republic (pp. 800–807).

Honorof, D., McCullough, J., & Somerville, B., last retrieved June 24 (2013). *Comma gets a cure*. URL: 〈http://web.ku.edu/∼idea/readings/comma.htm〉.

Hoole, P. (1993). Methodological considerations in the use of electromagnetic articulography in phonetic research. *FIPKM*, *31*, 43–64.

Hoole, P., & Zierdt, A. (2010). Five-dimensional articulography. In: B. Maassen, & P. van Lieshout (Eds.), *Speech motor control* (pp. 331–349). Oxford, UK: Oxford University Press.

King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., & Wester, M. (2007). Speech production knowledge in automatic speech recognition. *Journal of the Acoustical Society of America*, *121*(February (2)), 723–742.

Kröger, B. J., Pouplier, M., & Tiede, M. K. (2008). An evaluation of the Aurora system as a flesh-point tracking tool for speech production research. *Journal of Speech Language and Hearing Research*, *51*(4), 914–921.

Kroos, C. (2008). Measurement accuracy in 3d electromagnetic articulography (Carstens AG500). In: R. Sock, S. Fuchs, & Y. Laprie (Eds.), *Proceedings of the eighth international seminar on speech production* (pp. 61–64). Strasbourg, France: INRIA.

Kroos, C. (2012). Evaluation of the measurement precision in three-dimensional electromagnetic articulography (Carstens AG500). *Journal of Phonetics 13*.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*, 522–523.

Pouplier, M., & Goldstein, L. (2010). Intention in articulation: *Articulatory timing in alternating consonant sequences and its implications for models of speech production*. *Language and Cognitive Processes*, *25*, 616–649.

Pouplier, M., & Hardcastle, W. (2005). A re-evaluation of the nature of speech errors in normal and disordered speakers. *Phonetica*, *62*, 227–243.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0. 〈http://www.R-project.org〉.

Richmond, K., Hoole, P., & King, S. (August 2011). Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Proceedings of Interspeech*, Florence, Italy (pp. 1505–1508).

Van Engen, K., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2007). Development of the wildcat corpus of native- and foreign-accented English. Poster presented at the mid-continental workshop on phonology, Ohio State University.

Van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., & Bradlow, A. R. (2010). The wildcat corpus of native-and foreign-accented english: *Communicative efficiency across conversational dyads with varying language alignment profiles*. *Language and Speech*, *53*(4), 510–540.

Wells, J. C. (1982). *Accents of English I: An introduction*. Cambridge University Press, Cambridge, New York.

Westbury, J. R. (1994). X-ray microbeam speech production database user's handbook, version 1.0. Waisman Center on Mental Retardation & Human Development, Madison, WI.

Wilson, M., & Wilson, T. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin and Review*, *12*(6), 957–968.

Wrench, A. A., & Hardcastle, W. J., (2000). A multichannel articulatory speech database and its application for automatic speech recognition. In *Proceedings of the fifth seminar on speech production: Models and data & CREST workshop on models of speech production: Motor planning and articulatory modelling*. Kloster Seeon, Bavaria, Germany (pp. 305–308).