SYNCHRONIZING VIDEO, ULTRASOUND, AND AUDIO WITH A WATER BALLOON

Thomas Magnuson & Chris Coey

Department of Linguistics, University of Victoria, Victoria B.C., Canada
thomasm@uvic.ca; coey@uvic.ca

ABSTRACT

Synchronization complicates the use of ultrasound in phonetics research, particularly with methods that incorporate video data of external reference points (e.g., [3, 4]). The challenge of a third stream is compounded in longer spans of data at higher than NTSC standard frame rates. As part of work toward developing high-speed ultrasound techniques for long spans of running speech, this paper proposes using a water balloon to create a common analogue, non-speech reference event across tri-modal data. 'Snapping' the tied end of the balloon on camera while holding its base against an ultrasound transducer results in a synchronous optical and acoustic event. This allows for the validation of otherwise synchronized data as well as the alignment of otherwise unsalvageable, misaligned data. Using a water balloon clacker-board immediately before a span of interest potentially allows that immediate span to be synchronized to the point of analytical viability.

Keywords: ultrasound, synchronization

1. INTRODUCTION

A problem with using ultrasound in phonetic research is that it involves at least two streams of data that must be aligned before any meaningful analysis can be attempted: the ultrasound video and the audio signal. Techniques such as Palatoglossatron [3] and CHAUSA [4] that make use of video of external reference points to track the position of the palate involve a third stream: optical (non-ultrasound) video of the head. At the standard NTSC frame rate of 29.97 frames/second, synchronization of the ultrasound and audio is possible through external mixing hardware such as Canopus cards. At higher frame rates, synchronization is less straight-forward as a range of software, hardware, and general computer processing limitations can act to throw any of these streams temporally out of synch with one another. For this reason, ultrasound data collection architectures require digital or analogue 'clacker-boards' to validate alignment and recalibrate between subjects or takes. Clacker-boards are also useful with data which would otherwise be unsalvageable due to messy alignment. That is, inserting a reference event immediately prior to an analysis target allows for data streams to be satisfactorily aligned at least in the immediate vicinity of the object of analysis.

The precise form (e.g., digital versus analogue) of an appropriate mechanism ultimately depends of the needs, resources, and research direction of each institution. Hueber, et al. [2] video recorded a mallet striking the pump mechanism of a lotion dispenser, which subsequently deposited a fluid droplet onto an ultrasound transducer. Miller, et al. [4] combined an analogue bell and video clacker-board with a tri-modal pulse generator that electronically imprinted a landmark onto each data stream. Wrench & Scobbie [7] similarly involved imprinting a digital signal across signals, but without an analogue component.

The water balloon clacker-board proposed here was developed to satisfy three criteria for ongoing research at the University of Victoria's Speech Research Lab: 1) It had to be deployable with ease at any time during a collection session to compensate for progressive asynchrony in longer 60 f.p.s. recordings, 2) It had to involve a simultaneous analogue signal detectable across the three modalities, and 3) It had to be easily accessible to both fieldwork and pedagogy. A water balloon was the lowest-tech match for these criteria, and as a volume-preserving hydrostat it featured the added benefit of being analogous in size and shape to the tongue when viewed in ultrasound.

2. THE BALLOON ACROSS MODALITIES

The action of the water balloon in video and ultrasound is shown in Fig. 1, and the waveform and spectrogram of the resultant sound is shown in

Fig. 2. Holding the base of the balloon against the transducer then pulling upwards on the tied end 'arms' the mechanism. Once the tied end is released, the balloon rapidly contracts from an elongated pear-shape to the more egg-like shape of its resting state. While the balloon's tied knot is not visible in the ultrasound image, we do see a rapid contraction of the balloon along with the excitation of the air trapped within it.

Figure 1: 'Snapping' the knotted portion of a water balloon as a tri-modal clacker board. The sequence below shows consecutive frames extracted from video and ultrasound data recorded at 60 f.p.s. each.

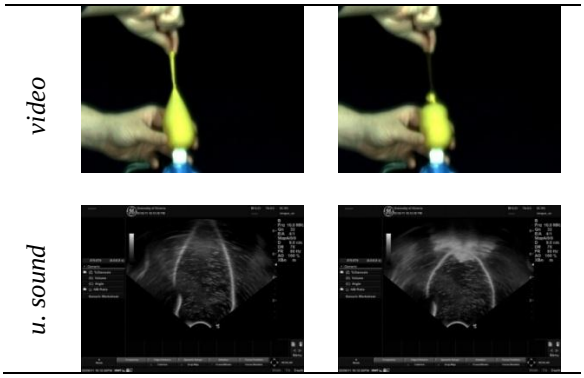
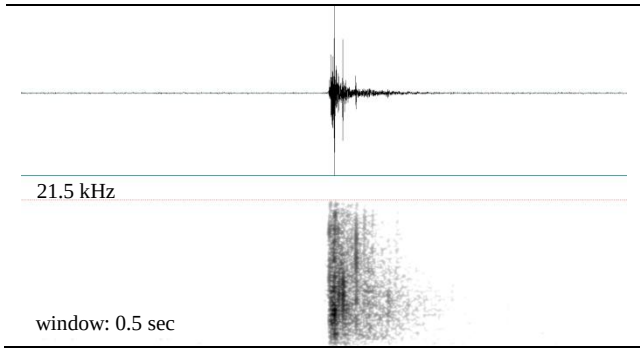


Figure 2: Waveform and spectrogram of a balloon snap.



The sound that results is a brief transient across a wide range of frequencies, to roughly 21.5 kHz. The waveform typically features one prominent as well as multiple less-prominent peaks in amplitude (four in the example in Fig. 2, the second being the most prominent).

2.1. Deconstructing the acoustic signal

As seen in Fig. 2, the acoustic transient caused by snapping a water balloon is complex with multiple peaks in amplitude. In order to ascertain what part of the transient corresponds to what sub-component of the action of the balloon (and thus what to base alignment on), we need to deconstruct how the balloon makes the sound that it does. To do this, and based on evaluative recordings of balloon snaps discussed later in this paper, 50 balloon snaps were video recorded at 60 f.p.s. via a AVP Stingray CCD Firewire 800 camera. With the aim of teasing out the acoustic contribution of the portion of the balloon above the knot, 5 conditions (with 10 repetitions each) were evaluated: 1) Releasing (in staggered succession) either side of the latex ring that forms the opening of the balloon while pinching the tied knot; 2) Again pinching the knot (to removing the acoustic contribution of the filled lower part of the balloon), simultaneously releasing both sides of the latex ring; 3) Pinching then releasing the only the knot, instead of the latex ring; 4) Releasing the latex ring in staggered succession without pinching the knot, and 5) Simultaneously releasing the latex ring without pinching the knot. Fig. 3 below illustrates a staggered release of the latex ring while the balloon's knot is pinched (i.e., condition 1). Fig. 4 shows a representative waveform and spectrogram for one repetition from each condition.

Figure 3: A staggered release of the latex ring at the open end of the water balloon, while holding the knotted portion. a) pre-release; b) initial release of one side of the latex ring; c) completed release.

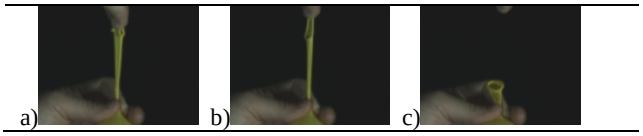
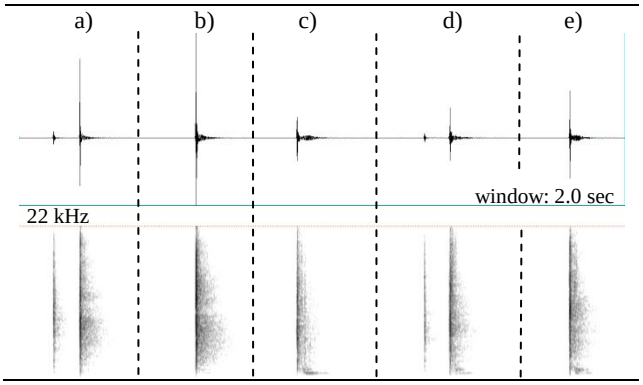


Figure 4: Waveforms and spectrograms for 5 test conditions: a) Staggered release of latex ring, pinching knot; b) Simultaneous release of latex ring, pinching knot; c) Release at knot only; d) Staggered release of latex ring w/o pinching knot; e) Simultaneous release of latex ring w/o pinch.



Snapping the balloon causes the stretched latex ring to recoil, resulting in a brief high-amplitude snapping noise. If the ring is released one side at a time, there are two amplitude peaks, as in Fig. 4(a, d). For the each of the test repetitions involving a

staggered release, the first of the two amplitude peaks was lower than the second. This is likely due to the increased potential energy transferred to the remaining held portion of the latex ring when the first portion is let go. Once the ring is released in either manner, it recoils at high velocity toward and into the lower part of the balloon. Where the lower part was pinched off at the knot (Fig. 4a, b; see also Fig. 3c), the researcher's fingers absorbed the impact of the rapidly descending ring.

The relatively higher amplitude peaks for the pinched knot condition as compared to the non-pinched knot conditions (Fig. 4d, e) suggest that the highest amplitude peak is associated with the ring's impact (as opposed to the release itself). The rationale for this is that, compared to a boney finger, a more massive and less rigid water-filled elastic bladder vibrates at a lower frequency than does a finger. Average intensities of the highest peaks in the pinched and non-pinched conditions support this intuition: 59.98 and 61.11 dB respectively for the staggered and non-staggered pinched conditions versus 54.25 and 55.37 dB respectively for the non-pinched conditions. Average intensity for the ten knot-only releases was lowest at 49.63 dB.

While a much higher frame rate video camera than the 60p device available to this study is necessary to determine the precise timing relationships between the release of the latex ring and its impact into either fingers or the lower part of the balloon, it seems reasonable to assume a close relationship between the highest peak in the waveform and the compaction/impact of the released part of the balloon. In any event, the two are temporally related within one frame at 60 f.p.s,, or 16.7 m.s.

3. APPLICATION TO DATA

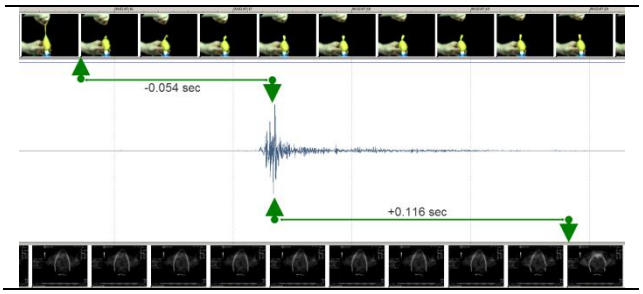
3.1. Alignment and measuring asynchrony

Fig. 5 shows the general alignment process using Sony Vegas 9 [5]; however, any audio/video editing software that allows multiple audio and video tracks to be decoupled and edited separately would be equally effective.

The data shown in Fig. 5 were 10-minutes in length, and captured by two computers, both running the capture software UltraCap [1]. One machine captured 60 f.p.s. video with the AVP Stingray camera while a GE Logiq-e set at 60 f.p.s. sent the ultrasound stream over VGA video output to a another machine equipped with a with an EMS RGB XTreme PCI-e video capture (frame-grabber) card. Audio (48 kHz, 16-bit) was recorded with a Sennheiser ME-60 shotgun microphone connected to the external video capture computer through a Mackie 1202VLZ mixer to an M-Audio Delta 1010LT PCI audio capture card. Through a coaxial SPDIF connection, an exact duplicate of the audio signal was mirrored to the ultrasound capture computer's m-audio Firewire 410 SPDIF interface. 'Locking' the ultrasound capture PC's audio clock to this signal allowed the audio stream to act as a common 'ruler' associated with the external video and the ultrasound recordings (see [4] for a description of a 29.97 f.p.s. Canopus system used as a ruler for higher frame rate ultrasound data).

Cloned or hardware-aligned audio tracks allow the data streams to be initially aligned with reasonable approximation. Importantly, due to dropped frames in both visual streams over the 10 minutes in the data here, this initial alignment was not enough to adequately synchronize all three signals. The video and ultrasound streams were typically out of sync to different degrees with the audio, as judged by the relative locations of the acoustic transient and key frames in the visual signals. Key frames were taken to be those corresponding most closely to the point of maximal compaction of the water balloon following the release of the tied end. This point was identified by toggling between single frames.

Figure 5: Measuring asynchrony between the highest amplitude peak in a waveform and corresponding key frames in 60 f.p.s. video and ultrasound data.



The asynchrony in the example in Fig. 5 was quantified by measuring the time difference between the highest amplitude peak in the waveform and the key frames' anchor points (denoted by small arrow marks in Sony Vegas 9, exaggerated in Fig. 5). A negative value thus represents a time before the acoustic peak, and a positive value represents one following it. In this case, the video preceded the audio by 0.054 seconds while the ultrasound followed the audio by

0.116 seconds. While Sony Vegas' editing features allow for the visual signals to be manually moved into synchronization with the audio, this is not strictly necessary for quantification purposes alone. Using this technique to evaluate our 10-minute 60 f.p.s. set-up (as described above), we found that the video was on average (based on 10 recordings) 0.016 seconds ahead of the audio at the 1 min. mark and 0.221 seconds ahead at the 10th minute. The ultrasound meanwhile on average trailed the audio by 0.073 seconds at 1 min. but preceded the audio by 0.293 seconds at the 10th minute. These discrepancies are not ideal, and point to the need for continued efforts at improving the system's performance. In contrast, we also evaluated an external Canopus 29.97 f.p.s. hardware mixer for one hour (with balloon events recorded every 5 minutes). While the ultrasound signal trailed the audio by roughly 0.08 seconds from the outset, this was relatively constant throughout. An initial adjustment at the first balloon event would therefore bring subsequent alignment to within a single frame, or 0.034 seconds at 30 f.p.s.

4. CONCLUDING REMARKS

This paper has demonstrated that a water balloon can be used as a clacker board to create a reference event in concurrent video, ultrasound, and audio data. Key frames in the visual data which correspond to the compaction of the balloon's shape following the release of its tied end were associated with the highest amplitude peak in the acoustic event's waveform. Based on the timing relationship between the key frames and the waveform we were able to demonstrate how asynchrony in collected data can be quantified using audio/video editing software. While results suggested that more work is needed to develop a higher-than-NTSC frame rate capture system that can record longer spans of data, a reliable tri-modal clacker-board represents one step towards that goal. That said, it is important to keep in mind that there is no such thing as 'perfect' alignment: frames are invariably dropped due to computers' processing bottlenecks. The frame rate of any visual data, too, is itself a limit: any single frame is one image representing a continuous length of time (34 m.s. at 30 f.p.s., 16.7 m.s. at 60). Ultrasound data present a further complication: the image projected as the ultrasound frame is itself actually asynchronous – it is compiled from continuous and rapid anterior-posterior scanning through imaging crystals within the transducer head [6]. While this process is extremely rapid, it nonetheless means that the images captured are not so much snapshots of the articulators at work, but rather panoramas – with one end photographed slightly before or after the other. Taken together, this means that one cannot achieve such a thing as perfect, absolute alignment in any ultrasound or video data. Rather, we must be contented being able to ascertain the degree to which our data are not aligned, and strive to mitigate that asynchrony as best we can.

5. ACKNOWLEDGEMENTS

This research was supported by the Social Sciences and Humanities Research Council of Canada, #767-2010-1146. Any and all errors are entirely the authors' own.

6. REFERENCES

[1] Coey, C. 2009. *UltraCap (Version 1.4)*. [Computer program]. University of Victoria Linguistics.

[2] Hueber, T., Chollet, G., Denby, B., Stone, M. 2008. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. 8th International Seminar on Speech Production*, 365-368.

[3] Mielke, J. Baker, A., Archangeli, D., Racy, S. 2005. Palatron: A technique for aligning ultrasound images of the tongue and palate. In Siddiqi, D., Tucker, B.V. (eds.), *Coyote Papers* 14, 97-108.

[4] Miller, A., Finch, K. 2011. Corrected high-frame rate anchored ultrasound with software alignment. *Journal of Speech, and Hearing* 54, 471-486.

[5] Sony Creative Software Inc. 2008. *Vegas Movie Studio Platinum (Version 9.0b (Build 92))*. [Computer program].

[6] Stone, M. 2005. A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics* 19(6-7), 455-501.

[7] Wrench, A., Scobbie, J. 2008. High-speed cineloop ultrasound vs. video ultrasound tongue imaging: Comparison of front and back lingual gesture location and relative timing. In: Sock, R., Fuchs, S., Laprie, Y. (eds.), *Proc. of the 8th International Seminar on Speech Production* Strasbourg, France: INRIA, 57-60.