Investigating the effects of posture and noise on speech production

Ingmar Steiner[1], Peter Knopp[2], Sebastian Musche[2],
Astrid Schmiedel[2], Angelika Braun[2], Slim Ouni[3]

[1]*Multimodal Speech Processing, Cluster of Excellence MMCI*
[1]*Language Technology Lab, DFKI GmbH*
[1]*Computational Linguistics & Phonetics, Saarland University, Germany*
[2]*Institute of Phonetics, University of Trier, Germany*
[3]*Université de Lorraine, LORIA, UMR 7503,Vandoeuvre-lès-Nancy, F-54500, France*

steiner@coli.uni-saarland.de,
{knopp|musche|schmiedel|brauna}@uni-trier.de,
slim.ouni@loria.fr

Abstract

In recent years, speech production research has benefitted greatly from magnetic resonance imaging (MRI). Two problem areas can be identified in conjunction with MRI, however: (a) subjects are typically recorded in supine posture and (b) they have to produce speech in noise. This paper investigates both of these issues by comparing articulatory behavior in upright and supine posture, with and without noise. The production data are recorded using electromagnetic articulography (EMA) and ultrasound tongue imaging (UTI) simultaneously. Preliminary analysis of the EMA data suggests that speakers are affected by posture, noise, and the combination of both in different ways, and use different strategies in compensating for these effects.

Keywords: *speech production, posture effect, speech in noise, EMA, ultrasound*

1. Introduction

Speech production research, like many other fields, is benefitting in ever-increasing amounts from advances in the area of magnetic resonance imaging (MRI). This non-hazardous, non-invasive imaging omodality offers unprecedented insight into the human vocal tract. By acquiring a rapid succession of 2D slices, it is possible to scan the static vocal tract in 3D, or a 2D slice of articulatory movements at video frame rate, with synchronized acoustic recordings (25 to 30 Hz, cf. Narayanan et al. 2013, Niebergall et al. 2013). Within certain constraints, both approaches can even be combined for 4D vocal tract imaging (Zhu et al. 2013).

Two aspects of MRI scanning are however potentially problematic for speech production: (a) the speaker is required to *lie* in the scanner, and (b) the MRI scanner emits a very loud noise during acquisitioni. The effects of posture and gravity have been explored in several previous studies using cineradiography (Whalen, 1990), electromagnetic articulography (EMA) (Tiede, Masaki, Wakumoto, et al., 1997), optical tracking (Shiller, Ostry, and Gribble, 1999), X-ray microbeam (Tiede, Masaki, and Vatikiotis-Bateson, 2000), ultrasound tongue imaging (UTI) (Stone, Crouse, and Sutton, 2002; Stone, Stock, et al., 2007; Wrench, Cleland, and Scobbie, 2011), and MRI (Kitamura et al., 2005; Engwall, 2006; Traser et al., 2013). The number of subjects studied is generally very small, and results vary. While Wrench, Cleland, and Scobbie (2011) observe a "slight superior and posterior displacement of the tongue root" in all of their four subjects, Tiede, Masaki, and Vatikiotis-Bateson (2000) note consistent, but differing behavior patterns in the two speakers studied. Stone, Stock, et al. (2007) as well as Kitamura et al. (2005) point to significant between-subject differences. The reason for this seems to be that different subjects choose different strategies to cope with the unusual posture. Considerable caution should therefore be exercised when interpreting the data, and it seems best to start by considering each speaker individually before making generalizations.

The effects of noise on speech production have also been widely studied (Van Summers et al., 1988; Junqua, 1993; Liénard and Di Benedetto, 1999; Lu and Cooke, 2008). Analyses were carried out in the acoustic domain, with inferences made with regard to the underlying articulatory gestures. The previous studies are in good agreement as far as word and vowel duration, F0, overall intensity, and spectral tilt are concerned: most speakers will speak with increased duration, average F0 and intensity, as well as a reduced spectral tilt in a noisy environment (cf. e.g., Van Summers et al., 1988; Junqua, 1993). This is precisely what is generally described as the Lombard reflex (Lombard, 1911). The effect of ambient noise on formants is less clear. Results are fairly consistent with respect to F1: generally, an increase of F1 is observed. Changes in F2, on the other hand, seem to vary with individual speakers and possibly also gender. Junqua (1993) found an increase in female speakers only. Furthermore, effects on formant bandwidth and formant separation have been reported (e.g., Junqua, 1993).

In order to systematically investigate the interactions between posture, noise, and the production of sustained, reiterant, and running speech, we studied these conditions in a factorial design. Preliminary results for the edge vowels of German (/i/, /a/, and /u/) are presented in this paper.

2. Data and methods

Speakers were recorded in supine and upright posture, with and without masking noise, using 3D EMA and synchronized UTI.

A total of 7 speakers were recorded; 3 female, 4 male. All are native speakers of German and thus recorded stimuli in German; one (bilingual) speaker also recorded stimuli in English.

The decision about the materials was not an easy one taking into account that everything had to be recorded in four different conditions: Upright and supine, each in combination with and

without masking noise. Since the EMA coils tend to become detached after some time and may need to be reattached (which makes it more difficult to interpret the data), we decided to limit the material rather than risk having to reattach several coils with the possibility of missing their original placement.

In designing the material, we attempted to emulate established MRI speech production experiments by including a set of sustained speech sounds as well as simple nonsense utterances, and a small number of benchmark utterances. We thus recorded the following production tasks:

1. A set of sustained vowels and diphthongs. The vowels correspond to the "long" vowel phonemes of German and therefore represent extreme vowel qualities as well as the roundedness dimension: /i, e, ɛ, a, o, u, y, ø, a͡ʊ, a͡ɪ, ɔ͡ʏ/.
2. The consonant phonemes of German in an [aCa] context, where C is each of /p, t, k, b, d, g, m, n, ŋ, l, f, v, s, z, ç, j, ʃ, x, ʁ, h/
3. CV repetitions of the consonants /f, s, ʃ, ç, x, ʁ, m, n, ŋ, l/ in vocalic context /i, a, u/, e.g., [fifififi], to study coarticulatory effects. Since it would have been too time-consuming to include all German consonant phonemes, the plosives were dropped altogether, and only voiceless fricatives were recorded.
4. A repetition of the sustained vowels and diphthongs (see above), in order to study potential compensatory effects.
5. *Nordwind und Sonne*, the German translation of "The North Wind and the Sun" passage (cf. *Handbook of the International Phonetic Association* 1999, pp. 89).
6. 10 sentences taken from a project corpus designed to study German vowels (kindly provided by Phil Hoole).

2.1. Acquisition setup

Each speaker was recorded in upright (sitting) posture, and in supine posture, lying on a non-ferromagnetic gurney constructed for an earlier study (Steiner and Ouni, 2011). The sequence of conditions was the following:

1. upright without noise;
2. upright with noise;
3. supine with noise;
4. upright without noise.

This allowed the speakers to become accustomed to the EMA coils first, before being presented with the noise condition; conversely, presenting the noise first in the supine condition allowed us to isolate (more or less) the posture effect during noise, before allowing the speakers auditory feedback in the final, supine condition. This rationale reflects the situation of speakers in an MRI speech production experiment, where they have little, if any, opportunity to compensate for posture before the noisy scanning procedure.

The recordings were made simultaneously using a Carstens AG501 articulograph with 16 channels and an Ultrasonix Mindray DP-6600 ultrasound imaging system. The audio was recorded with a Sennheiser MKH816 P48 directional microphone mounted approximately 2 m from the subject. In addition, the entire procedure, which lasted 90 to 120 min per speaker, was documented using a digital video camera.

In order to later synchronize the modalities, a "clicker" was used to record an audible pulse before and after each production task.

2.1.1. Ultrasound tongue imaging

The tongue contour was tracked using an electronic convex array transducer (Mindray 35C20EA).

The video signal from the UTI system was recorded twice: one output was fed to the Articulate Assistant Advanced (AAA) software package, which recorded individual production tasks; the second output was recorded directly in a digital video recorder, which multiplexed the UTI video with the microphone signal and the audio prompts (see section 2.1.3) into a continuous, uninterrupted MPEG-2 stream.

The probe stabilization headset normally used to maintain a constant probe position could not be utilized in our experiment, as its metal parts would have interfered with the magnetic field of the EMA device. Moreover, it would have physically hindered the speakers from resting their heads comfortably in the supine condition. The speakers therefore held the probe by hand during the recordings; its position was monitored and adjusted whenever it deviated from its optimum, while two EMA coils on the probe tracked its position relative to the speaker's head (see section 2.1.2).

2.1.2. Electromagnetic articulography and acoustic recordings

The positions and orientations of the EMA coils was recorded at 250 Hz. Of the 16 available coils, 13 were attached as follows:

- three reference coils, behind each ear and on the bridge of the nose, in order to correct for head movements;
- two coils on the upper and lower lip, respectively;
- five coils on the tongue, three in the mid-sagittal plane (on the tongue tip, blade and dorsum), as well as one on either side of the tongue blade;
- one coil near the lower incisors to track jaw motion;
- two coils mounted on the UTI probe, one near the top, the second roughly 5 cm further down on the handle. These coils enable tracking the position of the probe throughout the experiment and registration of the UTI and EMA modalities. This technique is similar to the one described by Aron et al. (2006).

The three remaining coils were held back as spares. They were also used to capture 3D palate traces in both postures, and the speaker's bite plane at the end of each recording session.

The acoustic signal from the microphone was recorded in synchronization with the EMA sweeps at 48 kHz, 16 bit.

2.1.3. Prompt presentation and noise

The stimuli for all production tasks were presented to the speakers via in-ear headphones. All stimuli except the last set had been recorded by a male speaker of Standard German without any regional markers; the 10 sentences were synthesized using a text-to-speech system. Each prompt was followed by two beeps spaced 2 to 5 s apart (depending on the task); the speakers were instructed to speak between these beeps. In the upright posture condition, the stimuli were additionally presented via a computer screen facing the speakers; this allowed them to familiarize themselves with the prompt list during the two upright repetitions with visual, as well as aural, input, before having to rely only on the audio prompts in the supine condition.

The simulated MRI noise was likewise presented via earphones, between the beeps. We selected a recording of gradient echo noise,[1] chosen for its roughly uniform structure. The noise level was set according to subjects' individual tolerance. Across the speakers, the sound pressure level varied from 75 to 90 dB.

[1] recorded in 2010 at the University of Iowa Hospital Radiology Lab, `http://www.cornwarning.com/xfer/MRI-Sounds/`

3. Results

3.1. Acoustic analysis

Even though acoustic analysis is not the focus of this paper, analyses were attempted regarding fundamental frequency, as well as the first two formants, in order to establish agreement with previous studies and thus the validity of the data. However, the quality of the audio recordings was far from perfect: while the directional microphone picked up a good-quality signal in upright position, the recording quality in supine position is seriously degraded. Furthermore, it is difficult to find a dedicated software package which will work with noisy recordings, let alone automatically. Therefore, due to the significant noise in the acoustic data, we cannot report reliable formant measurements in this paper.

3.2. EMA analysis

For a preliminary analysis of our data, we selected the sustained vowel prompts for four of the speakers (two male, two female). The vowels were manually labeled based on the recorded audio, and these annotations were used to automatically extract synchronized time segments of the recorded EMA data.

For each of the eight measurement coils, we applied a principal component analysis to the position data for the vowel segments to compare the differences between the four experimental conditions for each vowel and coil separately. Figure 1 displays the first principal component (PC1) for female speakers VP05 and VP06 and male speakers VP07 and VP08. Each plotted box represents quartiles 1 to 3, with whiskers extending to ± 1.5 interquartile range.

4. Discussion and conclusion

While more thorough analysis is of course planned, and only portions of the data have been annotated so far, the preliminary analysis shown here allow us to make several observations:

- Jaw movement is strongly affected by supine posture and noise for VP07 and VP08.
- All speakers show a clear effect of posture (and to a lesser extent, for noise, except for VP05) regarding lip motion, and characteristic effects of rounding. For the female speakers this is restricted to the lower lip, although this may also be influenced by the individual attachment of the coils.
- Tongue mid, left, and right coils are strongly correlated, with little lateral motion (as expected with the vowels).
- Noise seems to affect the tongue tip motion of VP08 in both positions, while for VP06, it is mainly posture.
- A few coils seem to have become detached or faulty during the recordings, notably those on the lower incisors and tongue back in the supine conditions of VP06. The nature of these erratic measurements is yet to be investigated.

We have yet to analyze the ultrasound recordings; the large amount of manual effort involved makes this a formidable task, but we expect to benefit from the registered EMA tongue coil data to improve the ultrasound processing.

Overall, we can confirm the influence of posture and noise on articulation, and that speakers are affected, or compensate, in different ways. Pending more detailed analysis, we should indeed be wary of these effects when interpreting speech production data acquired in a noisy environment, and in supine posture, such as during speech production MRI studies.

5. References

Aron, M. et al. (2006). "Coupling electromagnetic sensors and ultrasound images for tongue tracking: acquisition setup and preliminary results". In: *7th International Seminar on Speech Production*. Ubatuba, Brazil.

Engwall, O. (2006). "Assessing Magnetic Resonance Imaging Measurements: Effects of Sustenation, Gravitation, and Coarticulation". In: *Speech Production: Models, Phonetic Processes, and Techniques*. Ed. by J. Harrington and M. Tabain. New York, NY: Psychology Press, pp. 301–313.

Handbook of the International Phonetic Association (1999). Cambridge University Press.

Junqua, J.-C. (1993). "The Lombard reflex and its role on human listeners and automatic speech recognizers". In: *Journal of the Acoustical Society of America* 93.1, pp. 510–524. DOI: 10.1121/1.405631.

Kitamura, T. et al. (2005). "Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner". In: *Acoustical Science and Technology* 26.5, pp. 465–468. DOI: 10.1250/ast.26.465.

Liénard, J.-S. and M.-G. Di Benedetto (1999). "Effect of vocal effort on spectral properties of vowels". In: *Journal of the Acoustical Society of America* 106.1, pp. 411–422. DOI: 10.1121/1.428140.

Lombard, É. (1911). "Le signe de l'elevation de la voix". In: *Ann. Maladies Oreille, Larynx, Nez, Pharynx* 37, pp. 101–119.

Lu, Y. and M. Cooke (2008). "Speech production modifications produced by competing talkers, babble, and stationary noise". In: *Journal of the Acoustical Society of America* 124.5, pp. 3261–3275. DOI: 10.1121/1.2990705.

Shiller, D. M., D. J. Ostry, and P. L. Gribble (1999). "Effects of Gravitational Load on Jaw Movements in Speech". In: *Journal of Neuroscience* 19.20, pp. 9073–9080.

Steiner, I. and S. Ouni (2011). "Investigating articulatory differences between upright and supine posture using 3D EMA". In: *9th International Seminar on Speech Production*. Montreal, Canada.

Stone, M., U. Crouse, and M. Sutton (2002). "Exploring the effects of gravity on tongue motion using ultrasound image sequences". In: *Journal of the Acoustical Society of America* 111.5, pp. 2476–2477.

Stone, M., G. Stock, et al. (2007). "Comparison of speech production in upright and supine position". In: *Journal of the Acoustical Society of America* 122.1, pp. 532–541. DOI: 10.1121/1.2715659.

Tiede, M. K., S. Masaki, and E. Vatikiotis-Bateson (2000). "Contrasts in speech articulation observed in sitting and supine conditions". In: *5th Seminar on Speech Production*. Kloster Seeon, Germany, pp. 25–28.

Tiede, M. K., S. Masaki, M. Wakumoto, et al. (1997). "Magnetometer observation of articulation in sitting and supine conditions". In: *Journal of the Acoustical Society of America* 102.5, p. 3166. DOI: 10.1121/1.420773.

Traser, L. et al. (2013). "The Effect of Supine and Upright Position on Vocal Tract Configurations During Singing: A Comparative Study in Professional Tenors". In: *Journal of Voice* 27.2, pp. 141–148. DOI: 10.1016/j.jvoice.2012.11.002.

Van Summers, W. et al. (1988). "Effects of noise on speech production: Acoustic and perceptual analyses". In: *Journal of the Acoustical Society of America* 84.3, pp. 917–928. DOI: 10.1121/1.396660.

Whalen, D. H. (1990). "Intrinsic velar height in supine vowels". In: *Journal of the Acoustical Society of America* 88.S1, S54. DOI: 10.1121/1.2029052.

Wrench, A., J. Cleland, and J. M. Scobbie (2011). "An ultrasound protocol for comparing tongue contours: upright vs. supine". In: *17th International Congress of Phonetic Sciences*. Hong Kong, China, pp. 2161–2164.

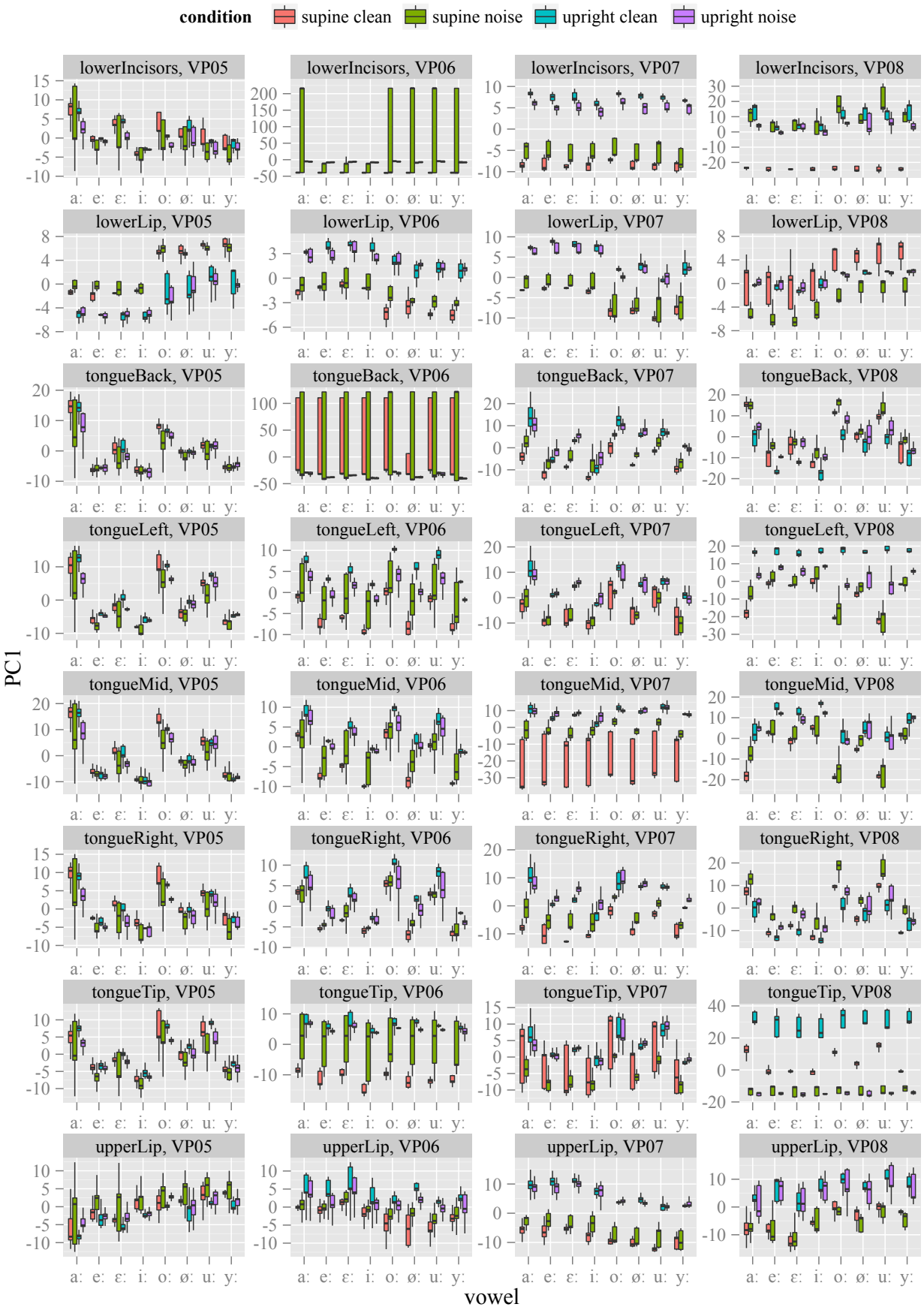

Figure 1: EMA results for sustained vowels from selected speakers