Speech Communication 123 (2020) 83–97

Contents lists available at ScienceDirect

Speech Communication

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

Accuracy, recording interference, and articulatory quality of headsets for ultrasound recordings

Michael Pucher[a,*], Nicola Klingler[a], Jan Luttenberger[a], Lorenzo Spreafico[b]

[a] *Acoustics Research Institute (ARI), Austrian Academy of Sciences (ÖAW), Vienna, Austria*
[b] *Department of Foreign Languages, Literatures and Cultures (DLLCS), University of Bergamo, Italy*

ARTICLE INFO

Keywords:
Ultrasound tongue imaging
Articulatory phonetics

ABSTRACT

In this paper we evaluate the accuracy, recording interference, and articulatory quality of two different ultrasound probe stabilization headsets: a metallic Ultrasound Stabilisation Headset (USH) and UltraFit, a recently developed headset that is 3D printed in Nylon. To evaluate accuracy, we recorded three native speakers of German with different head sizes using an optical marker tracking system that provides sub-millimeter tracking accuracy (NaturalPoint OptiTrack Expression). The speakers had to read $C_1V_1C_2V_{1/2}$ non-words (to diminish lexical influences) in three conditions: wearing the USH headset, wearing the UltraFit headset, and without a headset. To estimate the relative headset movement, we measured the movement between tracked points on the probe, headset, and speaker's nose. By also tracking visual marker points on the speaker's lip and chin, we compared the movement of the outer articulators with and without a headset and, thereby, measured how the headsets interfere with the articulatory space of the speaker. Additionally, we computed the differences in tongue profiles at the acoustic midpoint of V_1 under the three conditions and evaluated the articulatory recording quality with a distance index and an area index. In the final evaluation, we also compared formant measurements of recordings with and without headsets. With this objective evaluation we provide a systematic analysis of different headsets for Ultrasound Tongue Imaging (UTI) and also contribute to the discussion of using UTI stabilization headsets for recording natural speech. We show that both headsets have a similar accuracy, with the USH performing slightly better overall but introducing the largest error for one speaker, and that the UltraFit headset shows more flexibility during recordings. Each headset influences the lip opening differently. Concerning the tongue movement, there are no significant differences between different sessions, showing the stability of both headsets during the recordings. Acoustic analysis of formant differences in vowels revealed that the USH headset has a larger influence on formant production than the UltraFit headset.

1. Introduction

Ultrasound Tongue Imaging (UTI) is a medical-derived technique developed within articulatory phonetics to study real-time and offline tongue movements during speech (Stone, 2005). In the last decade, the technique, which appeared on the scene in the early 1980s (Shawker and Sonies Phd, 1984), has made progress both on the technical side, with the introduction of systems that are increasingly performing well in terms of spatial and temporal resolution (de Jong et al., 2019); and on the methodological side, with the development of techniques for the analysis of static and dynamic data that are increasingly informative (Pini et al., 2019).

In addition to analyses of articulatory phonetics, the UTI technique is also well suited to technological (Hueber et al., 2010; Fabre et al., 2017), educational (Wilson and Gick, 2006; Nakai et al., 2016; Ribeiro et al., 2019) and clinical (Preston et al., 2016) applications. Among the technological applications, the most interesting ones are silent speech interfaces, which are systems that allow speech communication without audible vocalization (Bruce et al., 2010).

Among the educational and clinical applications, interfaces have been developed that allow for visualization - also in mixed reality environments - of the tongue profile, which can improve speech articulation thanks to the positive action of the visual feedback (Eleanor et al., 2019).

[*] Corresponding author.
E-mail addresses: michael.pucher@oeaw.ac.at (M. Pucher), nicola.klingler@oeaw.ac.at (N. Klingler), jan.luttenberger@oeaw.ac.at (J. Luttenberger), lorenzo.spreafico@unibg.it (L. Spreafico).
URL: http://www.sociolectix.org (M. Pucher)

https://doi.org/10.1016/j.specom.2020.07.001
Received 4 November 2019; Received in revised form 7 May 2020; Accepted 6 July 2020
Available online 11 July 2020

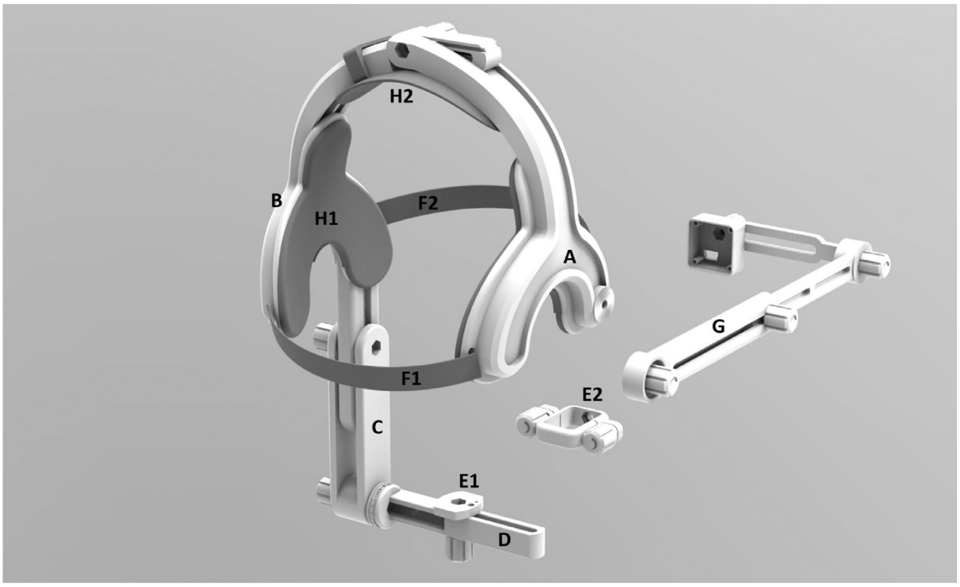

Fig. 1. Exploded view of the UltraFit system.

Whatever the field of application of UTI, one of the open issues is the stabilization of the ultrasound probe under the chin of the speaker to enable the definition of a fixed reference system for analysis or observations (Davidson and Decker, 2005). Although the resolution of this problem is felt differently by those who use the technique for research purposes or clinical practice, in recent years a number of ideas have been proposed to solve it; these include, for example the usage of mechanical systems (Stone and Davis, 1995; Scobbie et al., 2008; Davidson and Decker, 2005; Cai et al., 2011; Derrick et al., 2015; 2018); or software (Whalen et al., 2005); or simply holding the ultrasound probe by hand (Zharkova et al., 2015).

In this contribution we intend to deepen the evaluation of one of those solutions, the UltraFit headset (see Fig. 1) developed by Matosova (2016) and subsequently perfected and marketed by Articulate Instruments, and compare it with the Ultrasound Stabilisation Headset (USH) developed and marketed until 2018 by the same company (Scobbie et al., 2008; Articulate Instruments Ltd., 2008). The evaluation and comparison are relevant because the USH is among the most used stabilization devices in articulatory phonetics laboratories around the world.

The two headsets differ first in the material with which they are made: UltraFit is made of nylon, a synthetic polymer, while USH is made of aluminium, a non magnetic metal. A previous paper (Spreafico et al., 2017) described the process of developing the UltraFit headset and analyzed its usability.

The difference in the choice of materials has repercussions for many other aspects. First, it affects the shape of the UltraFit headset, because the polymer can be printed in 3D, enabling the headset to obtain a more organic shape, which is better with regard to both the fit and the maneuverability of the headset during setup, as well as with regard to the stability of the headset. Second, the choice of the polymer has positive repercussions for the weight, which is less than the metal headset. This is likely to be reflected in greater tolerability during prolonged sessions of use. Finally, the choice has an advantage in terms of integration with other techniques for investigating speech articulation. If necessary, the headset can be assembled without using metallic screws and bolts, thus, for example, ensuring compatibility in data collection sessions involving the use of Electromagnetic Articulography (EMA) or Magnetic Resonance Imaging (MRI).

Despite the many advantages, the accuracy of the measurements achievable using the UltraFit system remained to be tested. A preliminary assessment of the stability and accuracy of UltraFit was made by recording a speaker and showing that the overall error range of the headset movement for this speaker lay within 3 mm, with most errors lying in a 1–2 mm range (Spreafico et al., 2018).

Hence, in this paper we compare the accuracy of two different headsets “USH and UltraFit” by using data from three different speakers analyzed in reference to visual data about the movements of the headset. These movements were detectable externally using an optical tracking system. Additionally, we report acoustic data on the production of vowels and articulatory data on discrepancies detectable in the positioning of the tongue. Furthermore, we compare the results to acoustic and visual recordings of natural speech with and without wearing the headset.

This paper is organized as follows: In Section 2 we describe the visual, articulatory, and acoustic recordings. In Section 3 we present the analysis based on the visual data, which shows the accuracy of the headset and it’s influence on mouth opening. Section 4 contains the analysis of the headsets based on articulatory data and Section 5 those based on formants derived from acoustic data. Section 7 concludes the paper.

2. Data elicitation

For the evaluation of the accuracy of the two headsets shown in Fig. 2, we designed and ran a dedicated experiment. The experiment involved three informants. All informants were German native speakers of Standard Austrian German or Standard German German. The informants were characterized by having heads of different circumference, so as to highlight whether this parameter affects the stability of the helmet and therefore the accuracy of the measurement. The first speaker (*spk1*), female, had a small head size (53 cm in circumference); the second (*spk2*), male, had an average circumference (57 cm); the third (*spk3*), male, had a large circumference (60 cm).

Each speaker was seated in front of a computer in a semi-anechoic booth, and was instructed to read aloud the stimuli presented to him/her. The stimuli consisted of the following non-words of the type $C_1V_1C_2V_{1/2}$ repeated three times: /’paka ’paka ’paka/, /’taka ’taka ’taka/, /’tuki ’tuki ’tuki/, /’tipi ’tipi ’tipi/. Each non-word, pronounced with a trochaic stress in accordance with German phonotactics, was repeated three times during each recording session.

Each session began with a *silence* trial in which the speakers were instructed to keep the tongue in rest position and ended with a *swallow* trial. Each speaker attended three recording sessions: one wearing the metal helmet, one wearing the polymer helmet, one without helmets. During each session visual and synchronized articulatory and acoustic

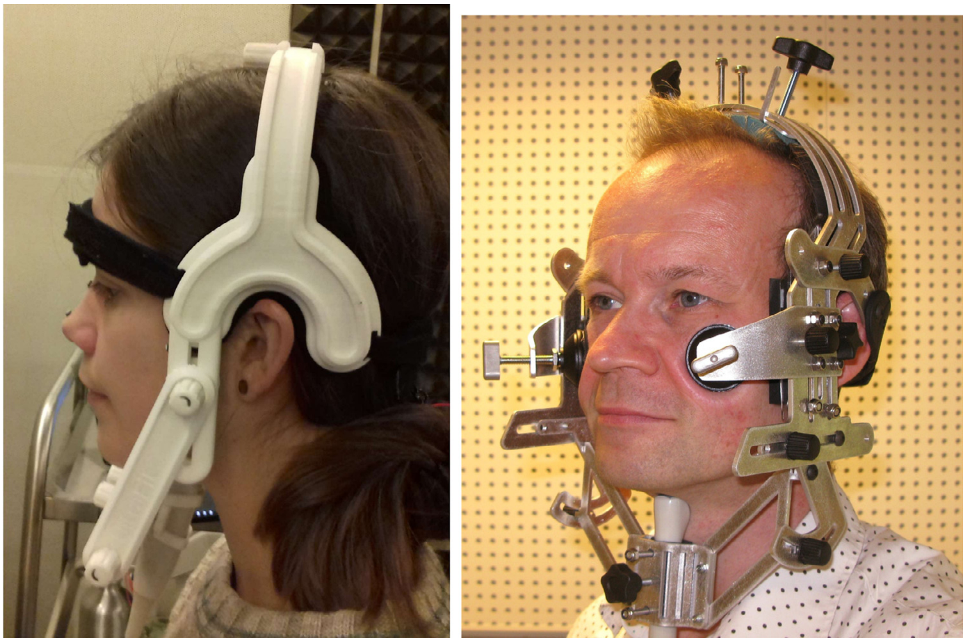

Fig. 2. UltraFit headset (left) and Ultrasound Stabilisation Headset (USH) (right).

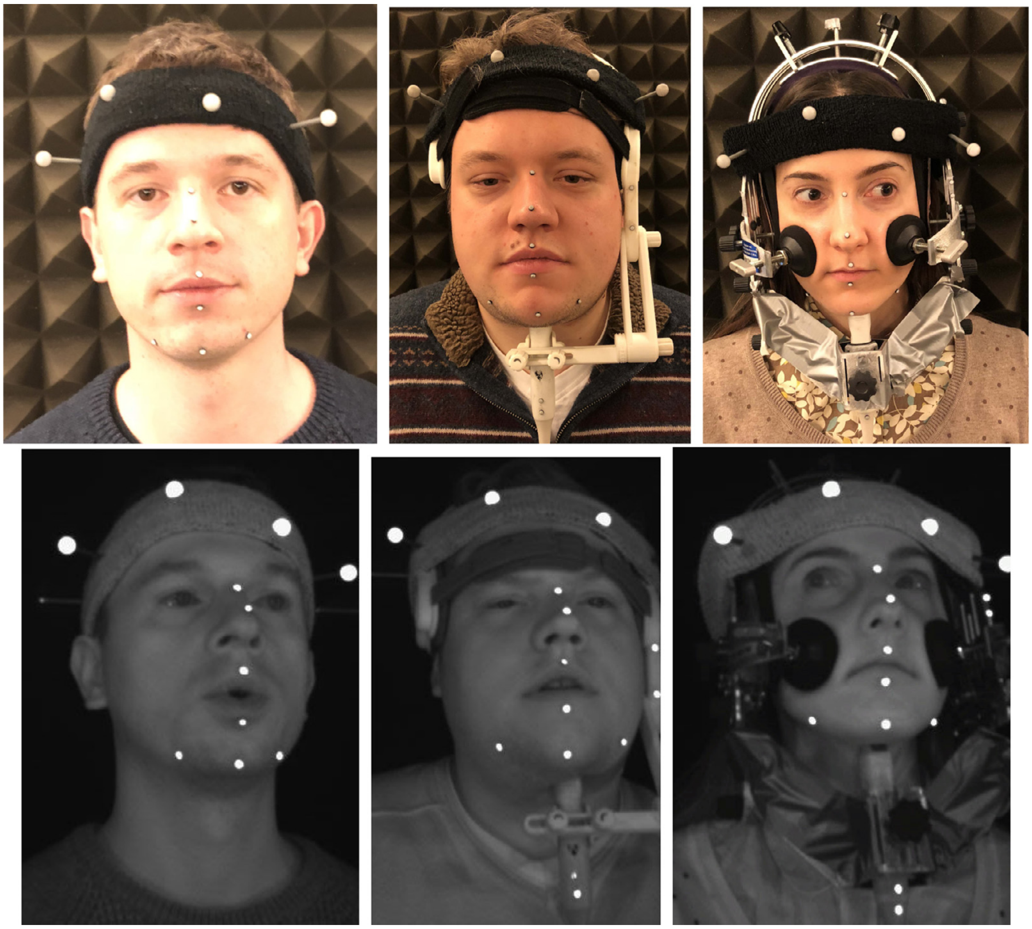

Fig. 3. Visual marker configuration (top). Video still from recordings (bottom). Natural - *spk2* (left column), UltraFit - *spk3* (middle column), and USH - *spk1* (right column) recording condition.

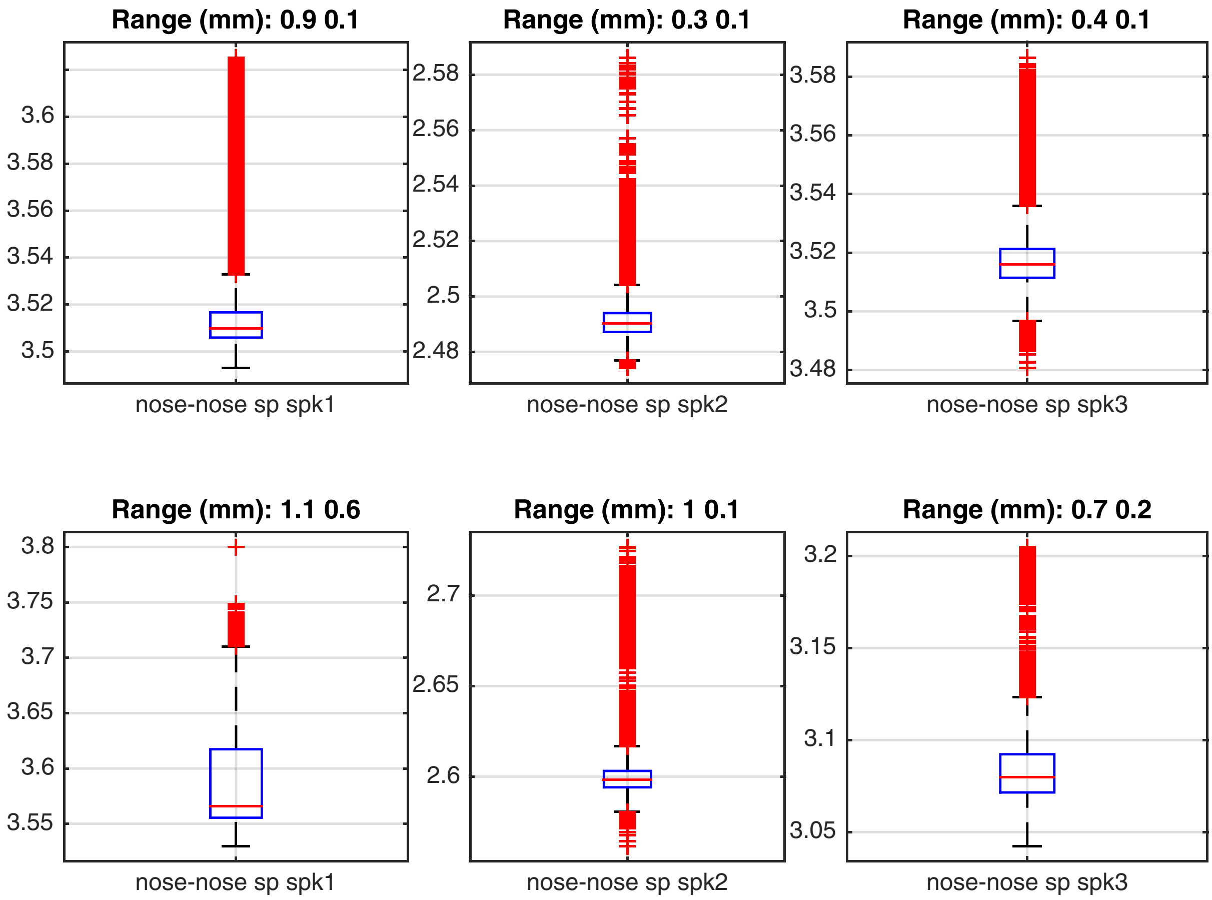

Fig. 4. Distribution of Euclidean distance between nose marker 1 and nose marker 2 for **sp**eech in UltraFit (top), and USH (bottom) condition..

data was collected. Altogether, the database contains 648 trials, namely 216 for each speaker.

2.1. Visual recordings

Facial movement was recorded using a NaturalPoint OptiTrack Expression system using seven FLEX:V100R2 infrared cameras. This system records the 3D position of reflective markers glued to the speakers face at 100 Hz.

We recorded the speakers without headset (*natural*), with the UltraFit headset (*UltraFit*) and with the USH (*USH*). The helmets were fixed by the same operator as firmly as possible to the head of the speakers. Regarding UltraFit, the auxiliary Velcro straps were not used to stabilise the probe arm laterally. The natural recordings were made to compare the lip opening with and without headsets. Depending on the recording condition we glued markers to the speaker's nose, the lips and jaw, the headset, and the ultrasound probe as shown in Fig. 3 for one speaker.[1] Here we only need a reduced set of markers, in previous work we used this system to record a full set of facial markers for facial animation (Schabus et al., 2014).

Additionally we also use the four headband markers that are used to remove head movement from the recordings. For the evaluation we use the output of the system directly without applying any manual corrections. Table 1 also shows the different marker configurations for the different conditions. Markers on the nose are also used to measure the inherent error of the system, distances between nose and probe markers are used for measuring the error of the recordings, and distances between lip markers are used to compare mouth opening.

Table 1
Number of markers per location / condition.

	Natural	UltraFit/USH
Headband	4	4
Nose	2	2
Lips	2	2
Jaw	3	3
Headset	0	2
Probe	0	2
	11	15

The different head sizes of the speakers are also indicated by the distances between nose and probe markers in Fig. 5. Fig. 3 shows the different marker configurations for *spk1*-*spk3*. *Spk3* is a Standard German German speaker, *spk1* and *spk2* are Standard Austrian German speakers.

2.2. Articulatory recordings

The analysis of articulatory data is of value because it can lead to the detection of differences between the two headsets that are not detectable by the analysis of visual recordings only. While the visual recordings referred to in Sections 2.1 and 3 are based on the observation of reflective markers in direct or indirect contact with the skin (which is independent from the position of the tongue), those referred to in this Section and Section 4 are based on the observation of ultrasound recordings from the probe, the fixation of which is why the headsets were developed.

[1] The adhesive tape on the USH headset was used to cover glossy parts and thereby improve the visual tracking robustness.

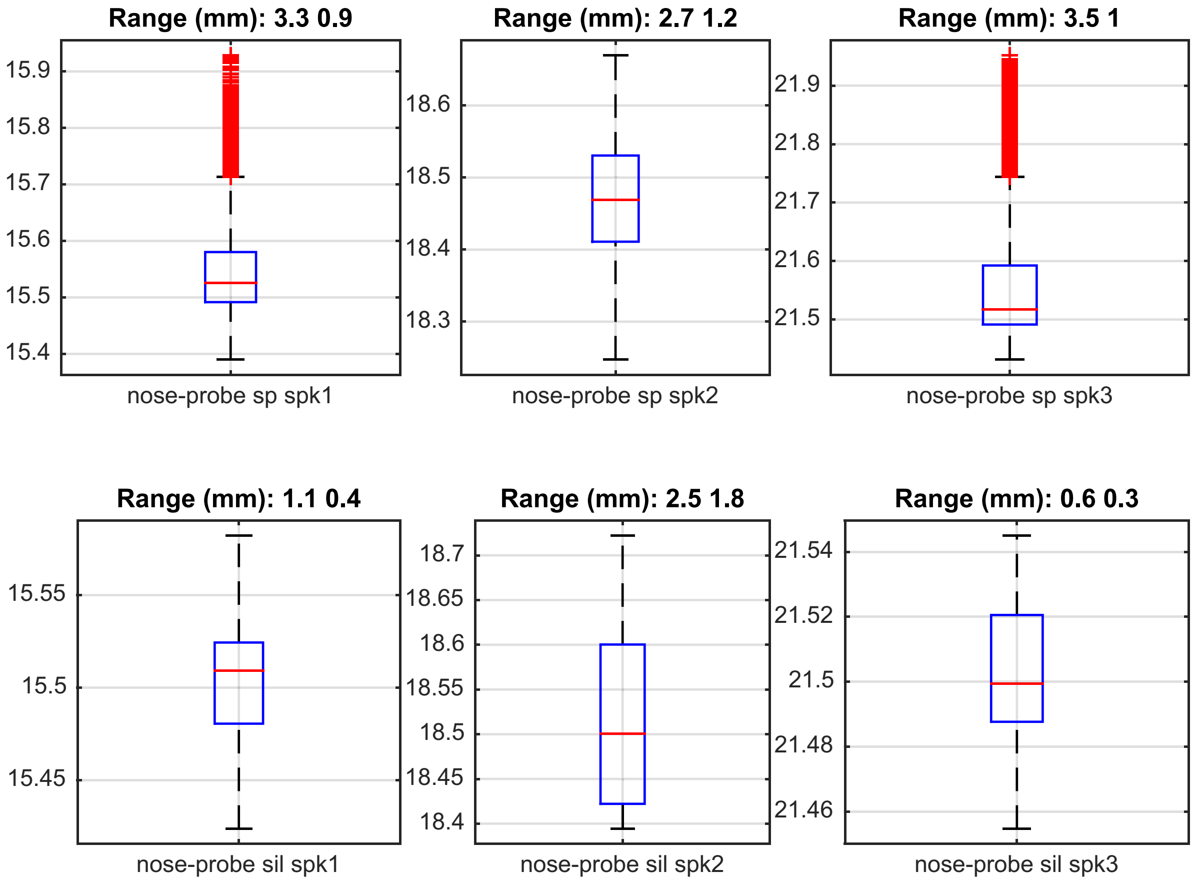

Fig. 5. Distribution of Euclidean distance between nose marker 1 and probe marker 1 for **sp**eech (top) and **sil**ence (bottom) in the UltraFit condition. *Spk1* (left column), *spk2* (middle column), *spk3* (right column).

The experimental data concern only the sets of recordings in which the speakers wore headsets. In fact, collecting ultrasound data by fixing the probe under the speaker's chin by hand would not guarantee reliable results for comparing the accuracy of the stabilization devices.

The articulatory recordings were made using the Micro Speech Research Ultrasound System (Articulate Instruments Ltd., 2017b) marketed by Articulate Assistant Advanced^{TM} coupled to the 5–8 MHz microconvex probe. The weight of the probe and two-thirds of the length of the hanging cable, excluding the weight of the connector, is 0.17 kg. This value should be taken into account because recent modelling work (Canella, 2019) has shown that the stability of UltraFit (and, therefore accuracy) is strongly influenced by the mass of the ultrasound probe. Ultrasound tongue imaging data was recorded with a fixed field of view of 150 degrees, at depths varying from 70 mm to 80 mm, at a sampling rate varying from 85 fps to 95 fps. The collected data was analyzed using the Articulate Assistant Advanced^{TM} software (AAA, v. 2.17.10; (Articulate Instruments Ltd., 2017a)).

With regard to the recording of articulatory UTI data, it is necessary to highlight some possible methodological criticalities. In fact, all the sessions took place on the same day and involved the same researchers, so as to try to partly mitigate the problems related to the reproducibility and repeatability requirements of data analysis involving biomarkers (Toeger et al., 2017).

Unfortunately, in an absolute sense this was impossible. In particular, the most difficult factors to control were operator variability, technical variability and image analysis variability for articulatory UTI data. With reference to the first factor, it was possible to exclude the inter-operator variability because the set-up of the articulatory instrumentation was entrusted to two researchers specialized in the practice. However, given the duration of the experiments, it was not possible to control or estimate a possible intra-operator variability.

With reference to the second factor, technical variability, the anatomical differences of the subjects concerning both the size of the head (intentional) and the shape of the chin and the mouth cavity (non-intentional) did not allow the repositioning of the ultrasound probe in anatomically identical positions for each of the three subjects. However, during the elicitation phase of the articulatory data - a subject that will be discussed in more detail in Section 4 - an attempt was made to find a functional correspondence of the images, orienting the ultrasound probe so as to include the points of contact between tongue and palate for the consonants /k/ and /t/. Since for our study we are mainly interested in the within speaker not the between speaker variability the technical variability is less critical.

Finally, with reference to the third factor, for the processing of the articulatory data and the extraction of the relative values it was necessary to adopt a semi-automatic analysis technique that also requires the initial manual definition of the language profile. Although the operation was entrusted to the same researcher, also in this case it is not possible to exclude a possible intra-operator variability.

2.3. Acoustic recordings

Acoustic recordings were conducted using a desk microphone and a USB audio interface (Focusrite Scarlett Solo) with a sampling rate of 44.1 kHz in a semi-anechoic booth.

3. Analysis of visual data

3.1. Accuracy

Measuring the distance between different visual markers allows for the measurement of the movement of the headset relative to the

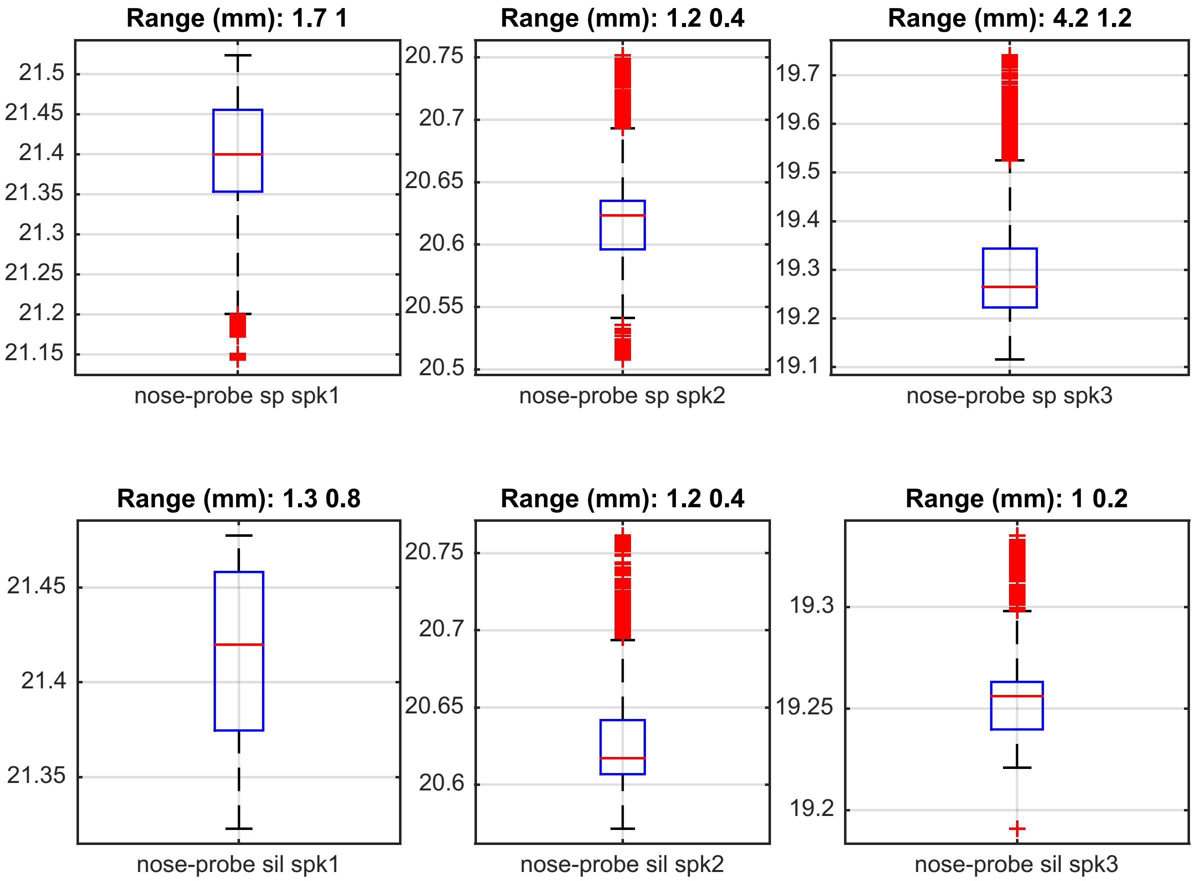

Fig. 6. Distribution of Euclidean distance between nose marker 1 and probe marker 1 for **sp**eech (top) and **sil**ence (bottom) in the USH condition. *Spk1* (left column), *spk2* (middle column), *spk3* (right column).

speaker's head. With no movement the distance of the markers should stay fixed with no variance. To measure the movement of the probe in relation to the speaker's head we measure the distance between a marker on the nose and a marker on the probe. To measure the internal error of the visual tracking system we measure the distance between the two nose markers. These two types of measures allow us to evaluate the accuracy of the headsets.

To measure the error of the recording setup we measured the distance between both nose markers, assuming that there is only little change in distance between the nose markers. Small changes are possible between the nose markers, when the speaker produces a facial movement that includes movement of the face.

So any changes in the nose-nose distance measurements can then be attributed to the visual tracking hardware and software, or small movements of the nose. The distribution of the nose-nose error is shown on Fig. 4. The range in millimeter (mm) in the title of each sub-figure is given for the 2.5th to the 97.5th percentile (first number) and for the 25th to 75th percentile of the data (second number).

We can see that the error is between 0.1 mm and 0.6 mm for 50% of the data for all speakers, and between 0.3 mm and 1.1 mm for 95% of the data for all speakers, such that we can conclude that the system performs with sub-millimeter accuracy almost all the time. For the USH condition (bottom) there is a slightly larger error of 1.1 mm (*spk1*), 1 mm (*spk2*), and 0.7 mm (*spk3*) for the 2.5th to 97.5th percentile.

Fig. 5 and 6 shows the Euclidean distances between the 3D points nose marker 1 and probe marker 1 for the whole recording session for the three speakers. This shows the error of the UTI headset during the recording session.

We can see that the maximum error in the 2.5th to 97.5th percentile is 3.5 mm for UltraFit and 4.2 mm for USH. The values for UltraFit in the 2.5th to to 97.5th percentile range from 0.6 - 3.5 mm, for USH from 1.0 - 4.2 mm. The head circumference can also be indirectly seen in the distances between nose and probe markers on the *y*-axis for the UltraFit condition, for the USH condition this relation does not hold due to different placements of the ultrasound probe for the three speakers.

We can see that the largest error occurs for the USH condition with 4.2 mm for *spk3*, although this speaker has a low intrinsic error of 0.6 mm as shown in Fig. 4. This shows that the nose-probe error is not dependent of the nose-nose error.

An F-test was performed to test if the samples are from a distribution with the same variance and significant differences ($p < .001$) were found for all speakers between the two conditions (UltraFit vs. USH). For *spk1* and *spk2* USH shows a higher accuracy, for *spk3* UltraFit shows a higher accuracy. Since the accuracy values of both conditions are in a similar range we may conclude that both headsets have a similar performance with the USH being slightly better.

Figs. 7 and 8 shows the distributions for the individual coordinates (x, y, z). This shows the error of the ultrasound headset in the different dimensions. To measure the movement in the different coordinates we have to remove the head movement first. This is done by using the four points of the headband, although we observed small movements of the head band during the recordings due to movements of the forehead. This is likely to have introduced errors in the numbers shown in Figs. 7 and 8. The fixing of the headband markers was easier with the USH than with the UltraFit condition. Furthermore one change of position of the headband markers introduces an error that is then present during the rest of the recordings.

The large errors especially for the UltraFit condition (Fig. 7) of 22.2 mm for *spk1* and 33.6 mm for *spk2* can be explained by the errors introduced through head movement removal. If we simply compute the

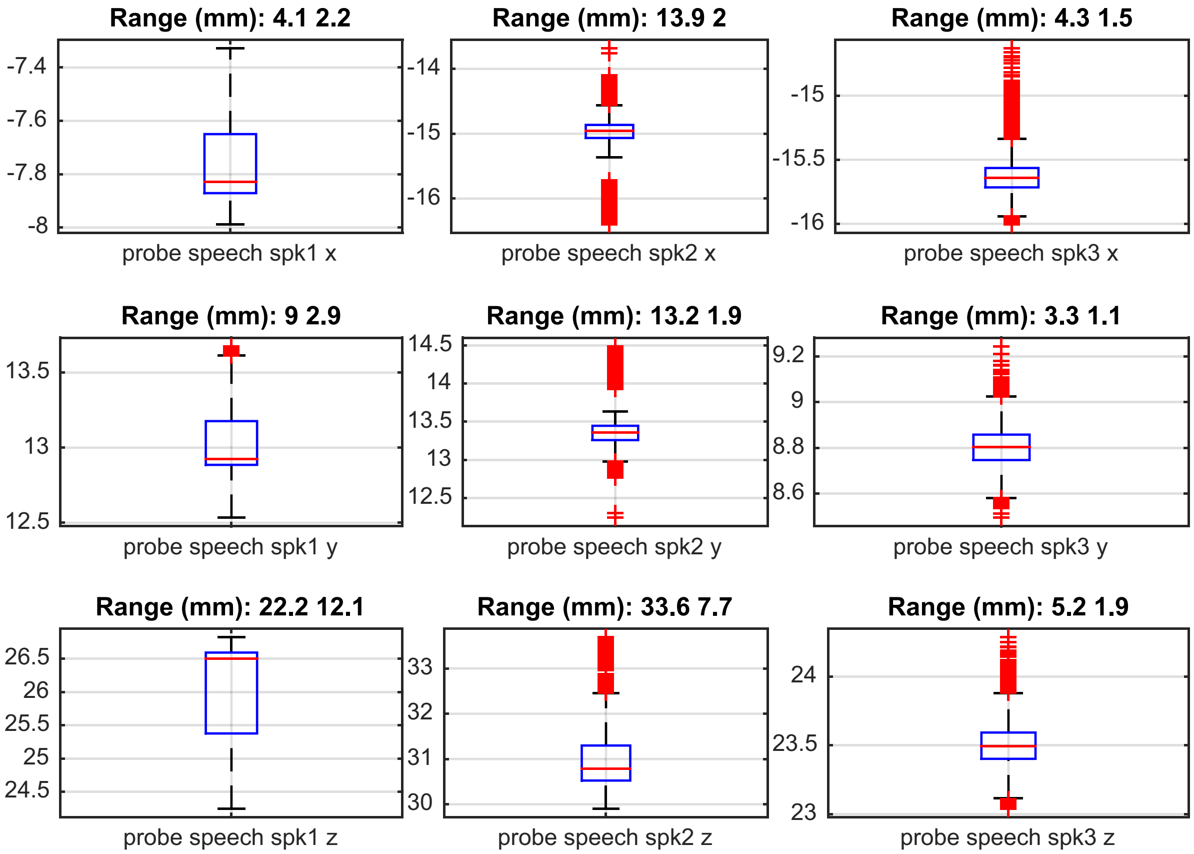

Fig. 7. Distribution of individual coordinates (*x, y, z*) for probe marker 1 for **sp**eech in the UltraFit condition after head movement removal.

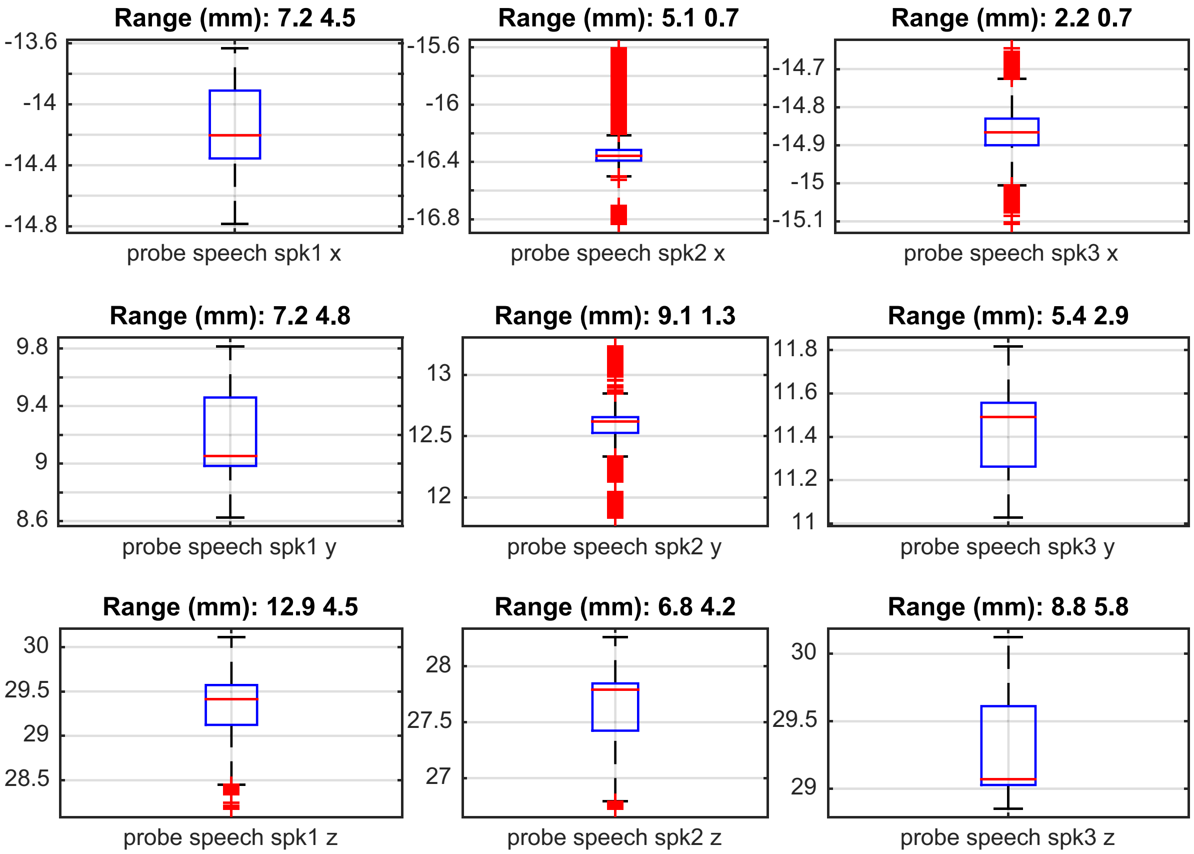

Fig. 8. Distribution of individual coordinates (*x, y, z*) for probe marker 1 for **sp**eech in the USH condition after head movement removal.

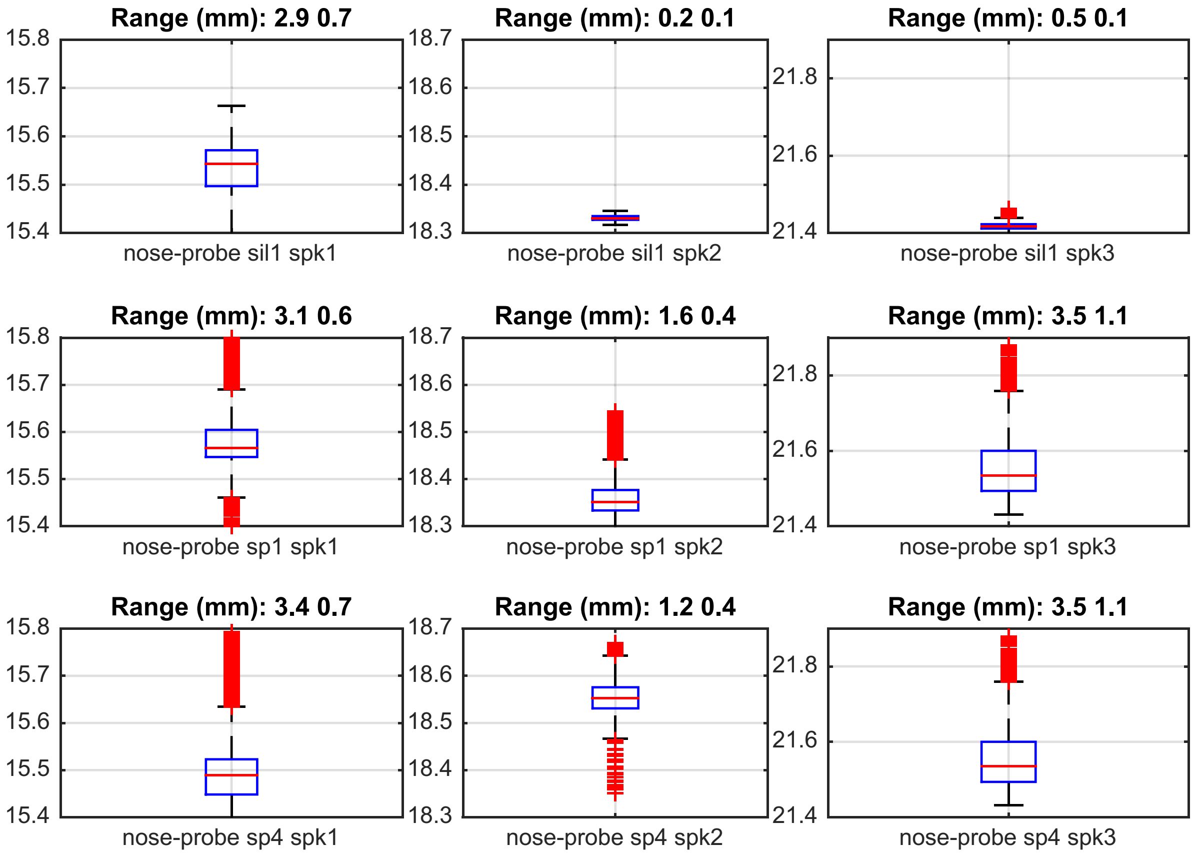

Fig. 9. Distribution of Euclidean distance between nose marker 1 and probe marker 1 for **sil**ence1, **sp**eech1, and **sp**eech4 in the UltraFit condition.

Table 2
Distribution of Euclidean distance between nose marker 1 and probe marker 1 for **sp**eech before and after head movement removal for the UltraFit headset.

	Spk1	Spk2	Spk3
Range before removal	3.3	2.7	3.5
Range after removal	9	13.5	2.6
Difference	-5.7	-10.8	0.9

distances between nose and probe marker from the data where head movement was removed and compare it with the distances in Fig. 5 we get 3.3 vs 9.0 mm (speech of *spk1*) and 2.7 vs 13.5 mm (speech of *spk2*) as shown in Table 2.

What we can still infer from Figs. 7 and 8 is that the largest error lies in the *z*-direction that is from the speakers head into the direction of the microphone.

To investigate the dynamics of recordings Figs. 9 and 10 show the distances for the first silence and speech, and last speech recordings for UltraFit and USH condition. In this way we can evaluate if there are changes during the recording session.

As can be seen in Fig. 9 the UltraFit is more flexible since it allows for expansion during the recording session from a smaller size in the first silence and then expanding during the recording. For the USH in comparison the median values do not change so much during the recordings.

A Wilcoxon rank sum test for equal medians shows significant ($p <$ .001) differences between the first silence recording (sil1) and the fourth speech recording (sp4) for the UltraFit condition for all three speakers and for the USH condition for *spk2* and *spk3*.

3.2. Recording interference

By recording visual markers at the outer articulators (lip, jaw) in the natural and two headset conditions we are able to measure if there is a difference between these recording conditions, which can be due to constraints that are set by the headsets leading to hypo- or hyperarticulation.

Fig. 11 shows the amount of lip opening during the recordings, which was measured by the distance between the upper and lower lip marker in cm. It can be seen that the largest difference between the three conditions appears at the rounded /u/ vowel (leftmost Figure), which indicates a larger amount of rounding of /u/ vowels in the USH condition (hyperarticulation), and a lower amount of rounding in the UltraFit condition (hypoarticulation).

The production of the /a/ vowel shows a very similar distribution for all three conditions. In /i/ vowels there are also small differences between the three conditions.

4. Analysis of articulatory data

During the experiment, the transition from the metal headset to the polymer headset forced the researchers to re-position the ultrasound probe. This re-positioning was done without any aid that would ensure that the probe was positioned in the identical location for both recording sessions. Because of this, each time the probe was re-positioned, a new spatial reference system was defined (Stone, 2005), differences in transducer angles relative to the head were introduced and different portions of the tongue and palate were visualized. This made it difficult to run a direct comparison of tongue and palate profiles in the two recording sessions of the same speaker and between the recording sessions of the different speakers (Pini et al., 2019). Therefore, in order to evaluate

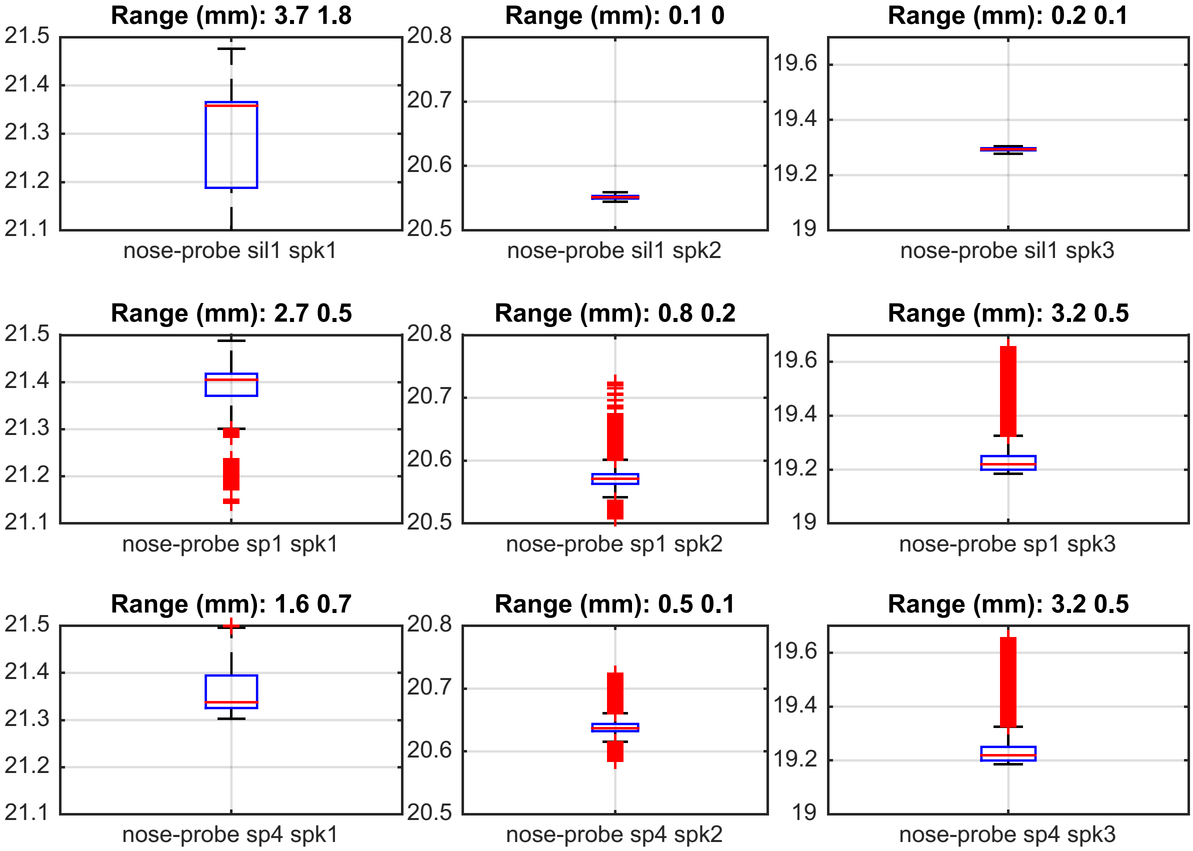

Fig. 10. Distribution of Euclidean distance between nose marker 1 and probe marker 1 for **sil**ence1, **sp**eech1, and **sp**eech4 in the USH condition.

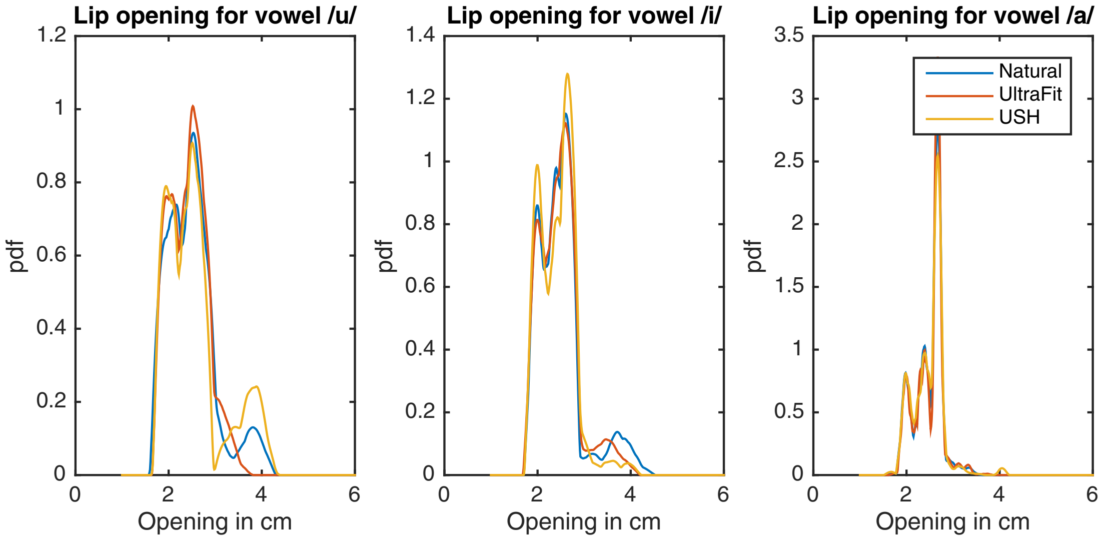

Fig. 11. Lip opening measured by distance of upper and lower lip marker for the vowels /u, i, a/..

the differences between the two headsets, we compared two indices a posteriori.

The first index considered is the distance between the ultrasound probe and the tongue profile. This distance is measured along the central radius of the fan superimposed on each ultrasound image by the software AAA (see Fig. 12). The choice of the radius is deliberate; since the depth setting of the ultrasound system is calibrated by taking this segment as a rough reference, a clear image of the tongue surface with a high spatial resolution is expected to always be on this radius, allowing for more accurate measurements to be obtained. This decision is similar to the one discussed by Vietti et al. (2015) which has also proven to be accurate enough to solve the task of word recognition from ultrasonic tongue images (Alessandro et al., 2015). Since this index is based on the measurement of the maximum tongue displacement for the radius in question, any low quality palate images are irrelevant.

The second index considered is the area of the geometric figure shown in Fig. 13. The figure has four sides defined by the intersection of the following: (a) the profile of the tongue of the speaker (AD line); (b) the line joining the place of articulation of /*t*/ and /*k*/ (BC line); (c) and (d) the radii joining the origin of the ultrasound

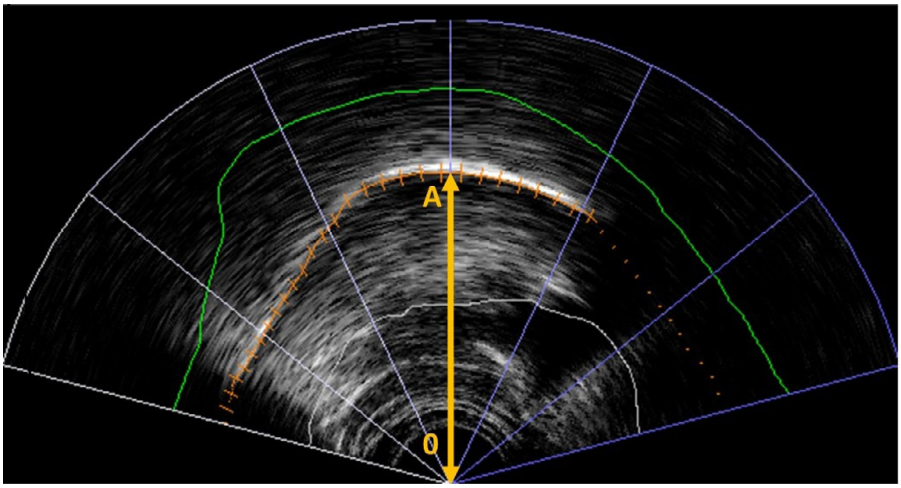

Fig. 12. Radius used for the computation of the distance between the probe and the surface of the tongue.

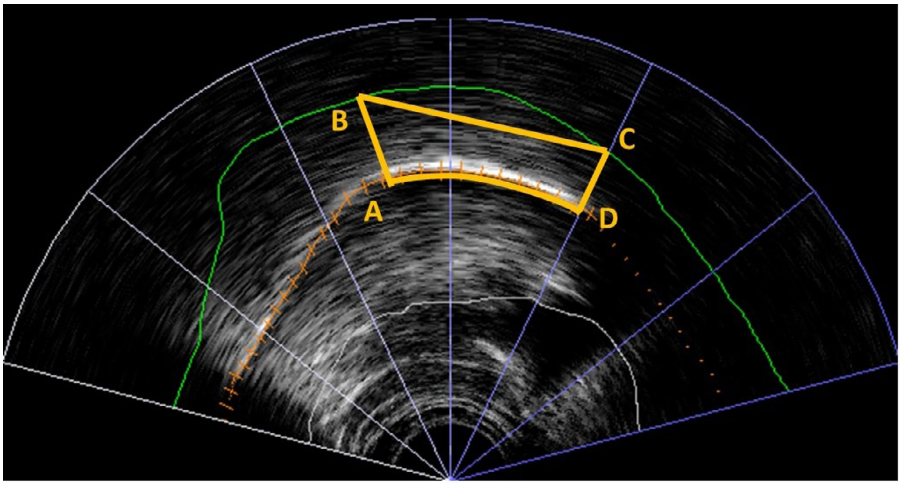

Fig. 13. Area taken into account for the computation of the area index.

probe with the place of articulation of /*t*/ (line AB) and /*k*/ (line CD), respectively.

The choice to use the BC segment as the upper side of the geometric figure instead of the more usual palate profile is due to the low quality of the ultrasound images during swallowing, namely those images that are used, typically, for the reconstruction of the palate profile itself (Epstein and Stone, 2005). Moreover, the decision to consider the places of articulation of /*t*/ and /*k*/ as points of reference because of the disappearance of the tongue profile from the ultrasound image due to the contact with the dental alveoli and the palate, respectively, was made because of the need to locate a region of the vocal tract that is significant for the production of linguistic sounds. This decision partly takes over the decision taken by Spreafico et al. (2015), Recasens and Rodríguez (2016), and Daniel and Clara (2018) to identify articulatory zones.

Since one of the aims of the research was to show whether the size of the speaker's head affected the accuracy of the headset in any way, the values of the second index have not been normalized so as to compensate for the different vocal tract sizes of the three informants, hence the following paragraphs show the results of the comparisons of the absolute values of the measurements made for the area index.

In this section we refer to the values measured for the two indices in reference to the production of the sequences /'taka 'taka 'taka/. The choice falls on this pseudo-word only because it contains the vowel /*a*/, which is supposed to determine the maximum displacement of the probe (and, thus, of the headset) because it involves the maximum jaw opening.

Of the four repetitions recorded by each speaker, only the second, third and fourth were considered. It was necessary to discard the first repetition because in two cases out of three (*spk1* and *spk2*) there were synchronization problems between the audio and the ultrasound signal that would affect the reliability of the data.

In addition, for each of the three repetitions, the values of the two indices were calculated based on the acoustics at the midpoint of the pronunciation of /*a*/. These values were extracted automatically after manually identifying the coordinates for the area index and defining the formula to calculate it using the “Analyse Value” function of the AAA software.

A first box plot representation of the absolute values for the area and distance indices shows that for each recording there is homogeneity in the variance of both. These differences are due to the fact that the comparison concerns absolute values (expressed in cm^2 and cm) while the size of each speaker's head and, therefore, the positioning of headsets and probe, as well as the field of view and depth settings in the ultrasound system, are different for each subject (see Fig. 14).

The values were compared with each other. The first comparison concerned the variations in area and distance between the second, third and fourth repetition of /'taka 'taka 'taka/ as pronounced by each of the three speakers. The hypothesis was that if the headset had been moved because of the cycles of pronunciation and swallowing, then the values of the two indices would have changed from repetition to repetition. However, since there are no significant differences between the three repetitions (ANOVA, ($p > 0.05$)), it can be deduced that the position of the ultrasound probe attached to the headset remains stable across all

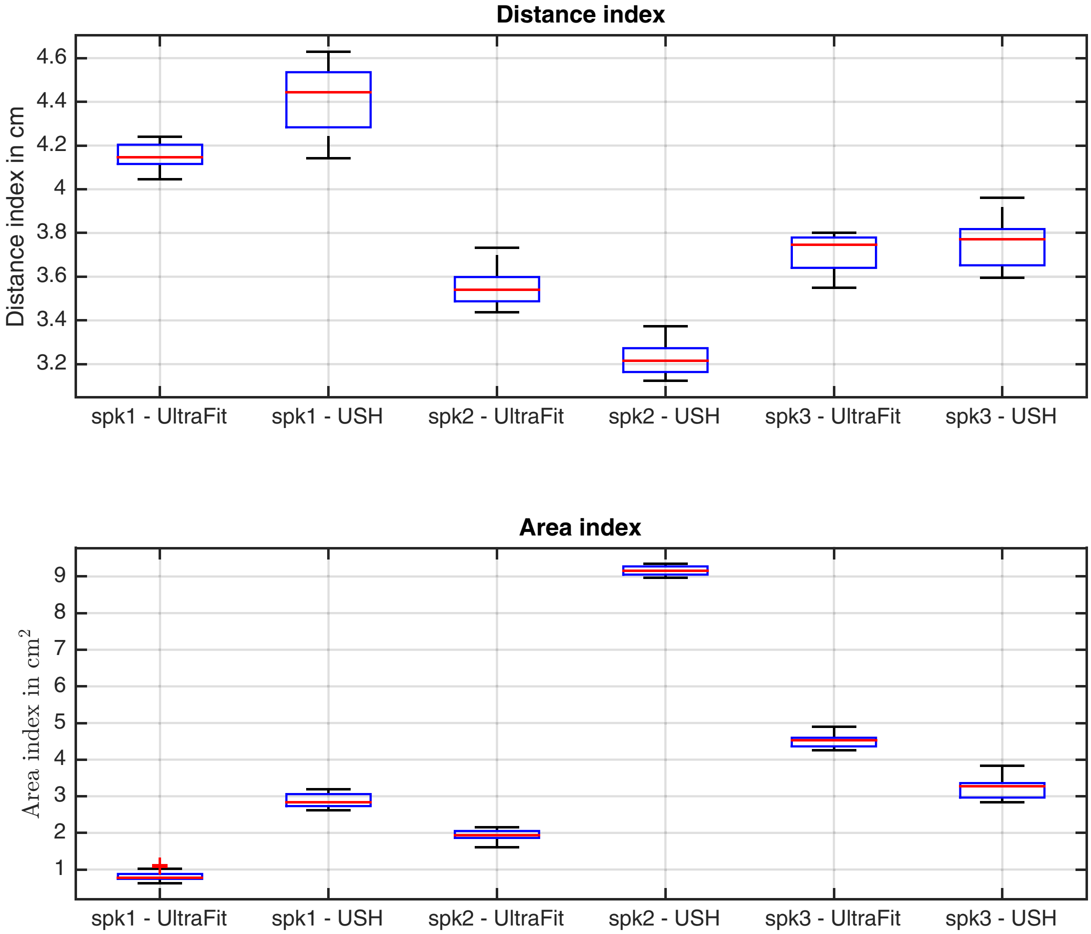

Fig. 14. Distribution of values for the distance index in cm (top) and area index in cm^2(bottom).

repetitions and recording sessions regardless of the type of headset or speaker.

The second comparison concerned the variation of the area and distance indices between two recordings made by the same speaker wearing the two types of headset. The objective in this case was not to check for significant differences between the two indices when wearing UltraFit or USH because, for the reasons set out above (the re-positioning of the probe), this was expected to be the case. Instead, the objective was to verify whether the difference in the indices between the recordings of the speakers with the polymer headset and the metal headset was significant. Indeed, if this were the case, it could be deduced that the difference is dependent on the type of headset used.

In fact, the results of the comparisons show that the differences are significant for all speakers, and for both indices (*t*-test, area index: *spk1* ($p < .001$), *spk2* ($p < .001$), *spk3* ($p < .001$); distance index: *spk1* ($p < .001$), *spk2* ($p < .001$), *spk3* ($p = .012$). This may be due to variances in absolute values related to the different sizes of the speakers' heads and to inconsistencies in the re-positioning of the headsets and the probe between one session and the other.

We also report the absolute values of the differences between the two indices because they are relevant for the purposes of our research. According to the data, the differences are, on average, 175 mm^2 (*spk1* = 204 mm^2; *spk2* = 191 mm^2; *spk3* = 128 mm^2) for the area index and 2.1 mm for the distance index (*spk1* = 2.6 mm; *spk2* = 3.2 mm; *spk3* = 0.4). While the data relating to the area is more difficult to interpret because it would also deserve a quantitative discussion of the differences in the geometric figure, the data relating to the distance is very informative.

First, the linear distances from the origin of the probe to the tongue surface detected with the polymer headset are always smaller than those detected with the metal headset, which could indicate a greater insertion of the probe between the metal protuberances. Second, the average difference between the maximum and minimum measurements is always lower for UltraFit (average: 2.3 mm) than for USH (average: 3.8 mm), perhaps testifying to a greater stability of the probe's positioning during the experiment. Moreover, these last values are relevant because they present orders of magnitude in line with those obtained from the analysis of the visual markers conducted with the NaturalPoint OptiTrack Expression system.

5. Analysis of acoustic data

The acoustic analysis includes measurements of the formants F1 to F3, and the duration of the stressed and unstressed vowels. The spectral information was extracted using a semi-automatic procedure in PRAAT (Boersma and Weenink, 2017).

The duration of the vowel was measured manually, relying on the periodicity and amplitude of the waveform. The script calculated the temporal midpoint of the interval (start and end of the vowel) and extracted the formants at these time points. Fourteen stimuli had to be excluded due to technical problems during the recording, resulting in 634 stimuli as a basis for the formant analyses.

The formant analyses included separate analyses of the F1, the F2, and combined measures of F1 and F2. In Fig. 15 the F1 (in Hz) values are given on the *y*-axis, the speakers and conditions (type of headset) are on the *x*-axis. As can be seen, the formant frequencies for F1 differ in regard of the speaker and condition. Furthermore, there is much higher variability in *spk1 - female* than in the other two speakers. Regarding F2 (Fig. 15 bottom), all speakers display a higher vari-

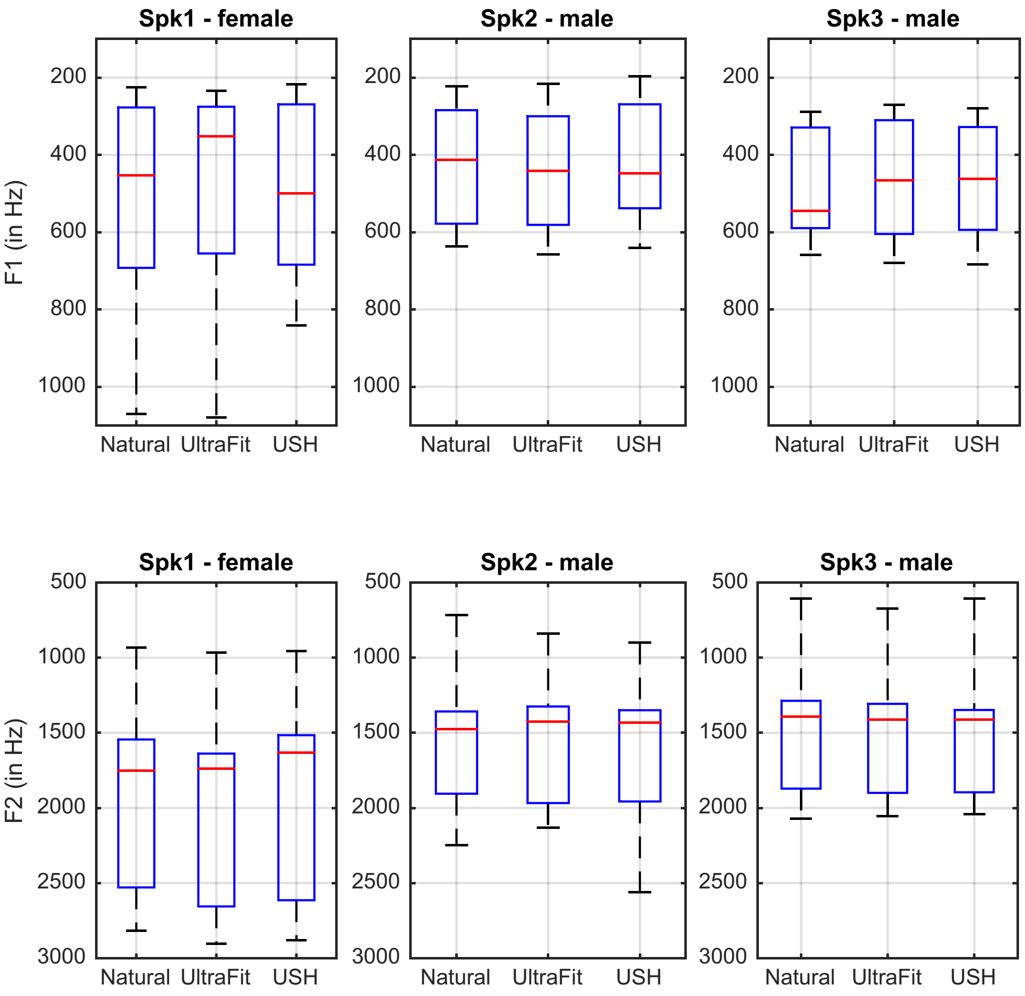

Fig. 15. Formant distributions for F1 (top) and F2 (bottom).

ability; however again *spk1 - female* seems to have the most variable production.

Fig. 16 displays the vowel space during the three conditions per speaker, with the F1 on the x-axis and the F2 on the y-axis. The colors differentiate the three conditions. Commensurate the three plots, the formant analysis shows that the production of vowels is influenced by condition and speaker. Furthermore, it seems as if the conditions natural and UltraFit produce a much more similar vowel space than the USH, indicating that the speakers were influenced to a higher degree by the metal head in their vowel production.

To estimate the influence of the three conditions (Natural, UltraFit, USH) and to account for the combination of the two dependent variables F1 and F2, we fitted a multivariate analysis of variance (MANOVA) in R (R Core Team, 2018). The model found significant main effects for speaker $(F(4,129), p < .001)$, triplet $(F(2,4.6), p = .01)$, condition $(F(4,3.1), p < .001)$, and target word $(F(6,1025), p < .001)$ on the formant values of the stressed vowel. The follow-up ANOVA proofed that F1 and F2 show significant influence from speaker (F1: $(F(2,417),\ p < .001)$ F2: $(F(2,417), p < .001)$) and target word (F1: $(F(3,417), p < .001)$, F2: $(F(3,417), p < .001)$) but only F1 is influenced by triplet $(F(1,417), p = .003)$ and condition $(F(2,417), p = .01)$.

For comparing the two headsets we are interested in the influence of the recording condition, the other variables (speaker, triplet, target word) are expected to have an influence on the formants as also revealed by the ANOVA. Only F1 not F2 is influenced by the recording condition, since the probe restricts the jaw opening in the USH and UltraFit condition, and the first formant is the most informative for restrictions in the jaw opening. F2 values are commonly associated with back-front vowels and less likely to be influenced by the different conditions.

Therefore we did a separate analysis (Linear Mixed Effect Model, lmer using (Bates et al., 2015)) for F1 of each vowel to account for the possibility of effects in different directions masking each other.

For the vowel /a/ the model revealed significant influences of the repetition ($t(210) = -2.927$, $p = .0038$), no significant difference between UltraFit and USH condition ($t(210) = 0.888$, $p < .37$), but a tendency for a difference between USH and natural condition ($t(210) = 1.697$, $p = .091$).

For /i/ the model did not reveal significant effects neither for repetition ($t(96) = -0.49$, $p = .62$) nor condition (UltraFit: $t(96) = 1.170$, $p = .245$; natural: $t(96) = 0.53$, $p = .568$).

For /u/ we found significant differences between USH and UltraFit condition ($t(102) = 2.424$, $p = .017$) and USH and natural condition ($t(102) = 2.536$, $p = .013$) as well as a tendency for a influence of repetition ($t(102) = -1.934$, $p = .056$).

As is documented in the statistics and can also be seen in Figs. 16 and 17, the speakers are influenced by the headset condition. The influence of condition on the vowel /a/ for F1 suggests that speakers were restricted in their jaw opening to a greater extent in USH condition thereby producing lower F1 and more centralized /a/ vowels (see Fig. 17). For the vowel /u/ in USH condition also lower F1 was produced, which may be attributed to general articulation restrictions since the lowering of the jaw plays a minor role in the production of /u/ vowels (see Fig. 17). Overall we can see that the USH condition had a larger influence on the first formant then the UltraFit condition.

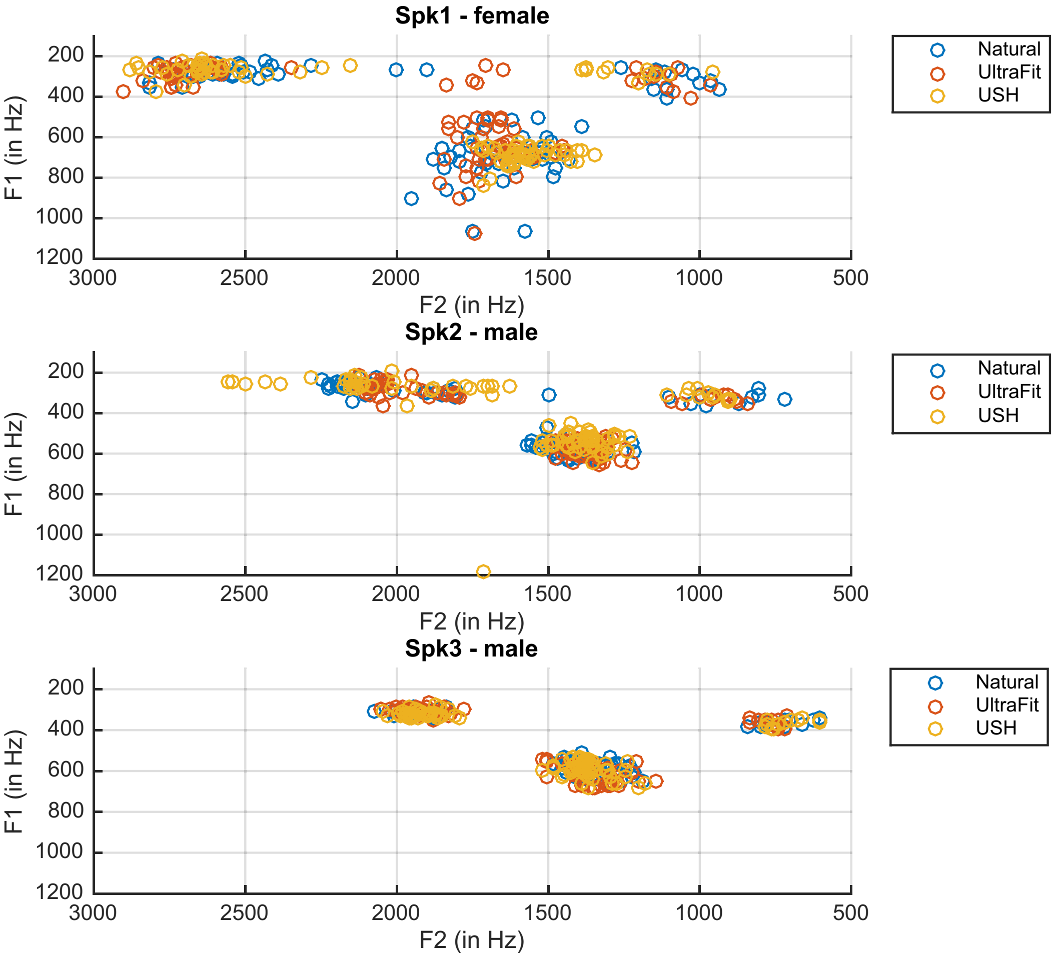

Fig. 16. Joint formant distributions of F1 (*y*-axis, from high to low values) and F2 (*x*-axis, from high to low values) for *spk1* (top), *spk2* (middle) and *spk3* (bottom).

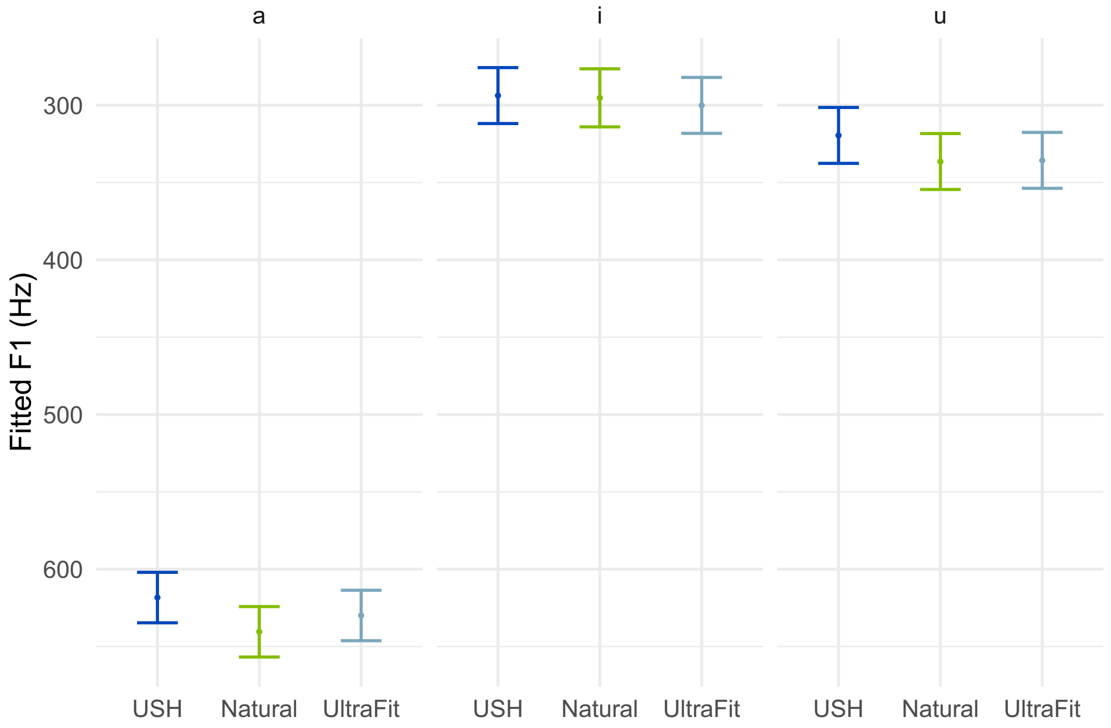

Fig. 17. Fitted F1 values taking speaker as a random factor.

Table 3
Overall evaluation of headsets.

	USH	UltraFit
Usability	★★★☆☆	★★★★★
Flexibility	★★★☆☆	★★★★★
Accuracy	★★★★☆	★★★★☆
Lip opening	★★★★☆	★★★★☆
Articulation	★★★★☆	★★★★☆
Formants	★★★★☆	★★★★★
Applications	★★★★☆	★★★★★

6. Discussion

The evaluation in this paper and previous work (Spreafico et al., 2018) allows us to compare the two headsets along different dimensions such as

- Usability, concerning the usability from the side of the speaker wearing the headset (comfort, easy to use) and the experimenter using the headset (fixing the headset). UltraFit has a much better usability since it is lighter and does not rest on parts of the head that can induce pain (Spreafico et al., 2018).
- Flexibility, concerning the possibilities to use the headset in different recording setups together with visual tracking software MRI, etc. Here also the UltraFit headset is more flexible, since it can be realized completely in plastic material (Spreafico et al., 2018).
- Accuracy, concerning the stability of the headset during recording, which was evaluated in Subsection 3.1 where we showed that the two headsets have similar accuracy.
- Lip opening, concerning the question if the headset influences the opening of the lips in some way, which was evaluated in Subsection 3.2 and showed that both headsets slightly influence the lip opening.
- Analysis of articulatory data in Section 4 showed that the position of the ultrasound probe remains satisfactorily steady across recording sessions regardless of the speaker or the type of headset.
- The formant analysis in Section 5 showed that the USH headset has a larger influence on the production of the first formant F1.
- Concerning application scenarios which are discussed in detail in Spreafico et al. (2018) the UltraFit headset has the advantage of being more easily usable with children for educational purposes for example.

Table 3 shows a scoring of the two headsets according to a five-star system that we derived from the overall evaluation.

7. Conclusion

We performed an objective evaluation of two headsets for Ultrasound Tongue Imaging (UTI), the USH, a metallic headset used in many laboratories today, and the UltraFit, a new headset made from polymer that was recently developed.

Using optical tracking hardware and software we showed that both headsets have a similar accuracy with the USH performing slightly better overall but introducing the largest error for one speaker, and that the UltraFit headset shows more flexibility during recordings. By measuring also the lip movement with visual tracking we showed that both headsets have a different influence to lip opening. Concerning the tongue movement there are no significant differences between different sessions showing the stability of both headsets during the recordings. Acoustic analysis of formant differences in vowels revealed that the USH headset has a larger influence on formant production than the UltraFit headset.

With these results we may conclude that both headsets are equally well suitable for recordings in speech science research, with the UltraFit being better in terms of usability, flexibility, and production of formants, which also makes it better suitable for technological, educational and clinical applications.

Declaration of Competing Interest

Dr. Lorenzo Spreafico has a financial interest in UltraFit, one of the evaluated headsets, since he was involved in the development of UltraFit, which is now commercialised by Articulate Instruments. For this paper he only performed the articulatory analysis, based on data that was collected at the Acoustics Research Institute (ARI) by the other authors. The analysis of visual and acoustic data was performed at ARI by the other authors. The other authors from the ARI have no financial interest/ personal relationships regarding UltraFit.

CRediT authorship contribution statement

Michael Pucher: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Formal analysis, Visualization, Supervision, Funding acquisition. **Nicola Klingler:** Data curation, Formal analysis, Visualization. **Jan Luttenberger:** Data curation, Formal analysis, Writing - review & editing. **Lorenzo Spreafico:** Data curation, Formal analysis, Visualization.

Acknowledgements

This work was supported by the Austrian Science Fund (FWF): F6002, I2539. This work was also supported with research funds from the project on Digital Humanities by the DLLCS Department of Excellence of the University of Bergamo.

References

Alessandro, V., Vittorio, A., Lorenzo, S., 2015. Verso un sistema di riconoscimento automatico del parlato tramite immagini ultrasoniche. In: Il farsi e disfarsi del linguaggio. Acquisizione, mutamento e destrutturazione della struttura sonora del linguaggio/Language acquisition and language loss. Acquisition, change and disorders of the language sound structure, pp. 477–489. doi:10.17469/O2101AISV000032.

Articulate Instruments Ltd., 2008. Ultrasound Stabilisation Headset—Users Manual, revision 1.5.. Articulate Instruments Ltd.. URL http://www.articulateinstruments.com/

Articulate Instruments Ltd., 2017a. Articulate assistant Advanced — Ultrasound Module User Guide, version 217.01. Articulate Instruments Ltd.URL http://www.articulateinstruments.com/.

Articulate Instruments Ltd., 2017. Installation manual for Micro Ultrasound system — Users Manual, revision 1.5.. Articulate Instruments Ltd.. URL http://www.articulateinstruments.com/

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1), 1–48. doi:10.18637/jss.v067.i01.

Boersma, P., Weenink, D., 2017. Praat: doing phonetics by computer. URL: http://www.praat.org.

Bruce, D., Tanja, S., Kiyoshi, H., Thomas, H., J.M., G., J.S., B., 2010. Silent speech interfaces. Speech Commun. 52 (4), 270–287.

Cai, J., Denby, B., Roussel-Ragot, P., Dreyfus, G., Crevier-Buchman, L., 2011. Recognition and Real Time Performance of a Lightweight Ultrasound Based Silent Speech Interface Employing a Language Model, pp. 1005–1008.

Canella, G., 2019. UltraFit: Modelling and Simulation of an Ultrasound Probe Stabilization Headset. Free University of Bozen, Italy Master's thesis. Unpublished BA thesis

Daniel, R., Clara, R., 2018. An ultrasound study of contextual and syllabic effects in consonant sequences produced under heavy articulatory constraint conditions. Speech Commun. 105, 34–52. doi:10.1016/j.specom.2018.10.007.

Davidson, L., Decker, P.D., 2005. Stabilization techniques for ultrasound imaging of speech articulations. J. Acoust. Soc. Am. 2544.

Derrick, D., Best, C., Fiasson, R., 2015. Non-metallic ultrasound probe holder for co-collection and co-registration with ema. In: Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015).

Derrick, D., Carignan, C., Chen, W.-r., Shujau, M., Best, C.T., 2018. Three-dimensional printable ultrasound transducer stabilization system. J. Acoust. Soc. Am. 144 (5), EL392–EL398.

Eleanor, S., Lloyd, S., Lam, J., Cleland, J., 2019. Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders. Int. J. Lang. Commun.Disord. 54 (5), 705–728.

Epstein, M.A., Stone, M., 2005. The tongue stops here: ultrasound imaging of the palate. J. Acoust. Soc. Am. 118 (4), 2128–2131. doi:10.1121/1.2031977.

Fabre, D., Hueber, T., Alameda-Pineda, X., Badin, P., 2017. Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. Speech Commun. 93 (4), 67–75.

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., reyfus, G., Stone, M., 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Commun. 52 (4), 288–300.

de Jong, K., Berkson, K., Lulich, S.M., Myers, S., Bohnert, A., 2019. The lingual topography of american english laterals in onsets and codas. J. Acoust. Soc. Am. 145 (3), 1928. doi:10.1121/1.5102009.

Matosova, A., 2016. UltraFit. Free University of Bozen, Italy Master's thesis. Unpublished BA thesis

Nakai, S., Beavan, D., Lawson, E., Leplatre, G., Scobbie, J., Stuart-Smith, J., 2016. Viewing speech in action: speech articulation videos in the public domain that demonstrate the sounds of the international phonetic alphabet (IPA). Innov. Lang. Learn. teach. 0 (0).

Pini, A., Spreafico, L., Vantini, S., Vietti, A., 2019. Multi-aspect local inference for functional data: Analysis of ultrasound tongue profiles. Journal of Multivariate Analysis 170, 162–185. Special Issue on Functional Data Analysis and Related Topics

Preston, J., Leece, M., Maas, E., 2016. Intensive treatment with ultrasound visual feedback for speech sound errors in childhood apraxia. Neuroscience 10 (240).

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Recasens, D., Rodríguez, C., 2016. A study on coarticulatory resistance and aggressiveness for front lingual consonants and vowels using ultrasound. J. Phonetics 59, 58–75.

Ribeiro, M.S., Eshky, A., Richmond, K., Renals, S., 2019. Speaker-independent classification of phonetic segments from raw ultrasound in child speech. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1328–1332. doi:10.1109/ICASSP.2019.8683564.

Schabus, D., Pucher, M., Hofer, G., 2014. Joint audiovisual hidden semi-Markov model-based speech synthesis. IEEE J. Sel. Top. Signal Process. 8 (2), 336–347.

Scobbie, J.M., Wrench, A.A., Linden, M.L.V.D., 2008. Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement. In: In Proceedings of the 8th Speech Production Workshop: Models and Data, pp. 373–376.

Shawker, T.H., Sonies Phd, B.C., 1984. Tongue movement during speech: a real-time ultrasound evaluation. J. Clin. Ultrasound 12 (3), 125–133.

Spreafico, L., Celata, C., Vietti, A., Bertini, C., Ricci, I., 2015. An epg + uti study of italian /r/. ICPhS.

Spreafico, L., Matosova, A., Vietti, A., Galata, V., 2017. Two head-probe stabilization devices for speech research and applications. Poster presentation. Ultrafest VIII. Potsdam, October 4–6, 2017.

Spreafico, L., Pucher, M., Matosova, A., 2018. Ultrafit: a speaker-friendly headset for ultrasound recordings in speech science. In: Proc. Interspeech 2018, pp. 1517–1520.

Stone, M., 2005. A guide to analyzing tongue motion from ultrasound images. Clin. Linguist. Phon. 19 (6–7), 455–502.

Stone, M., Davis, E., 1995. A head and transducer support system for making ultrasound images of tongue/jaw movement. J. Acoust. Soc. Am. 98 (6), 3107–3112.

Toeger, J., Sorensen, T., Somandepalli, K., Toutios, A., Lingala, S.G., Narayanan, S., Nayak, K., 2017. Test–retest repeatability of human speech biomarkers from static and real-time dynamic magnetic resonance imaging. J. Acoust. Soc. Am. 141 (5), 323–3336.

Vietti, A., Spreafico, L., Anselmi, V., Spreafico, L., 2015. Allophonic variation: an articulatory perspective. In: Presentation Ultrafest VII December 8th-10th, 2015, The University of Hong Kong.

Whalen, D.H., Iskarous, K., Tiede, M.K., Ostry, D.J., Lehnert-LeHouillier, H., Vatikiotis–Bateson, E., Hailey, D.S., 2005. The haskins optically corrected ultrasound system (hocus). J. Speech Lang. Hear. Res. 48 (3), 543–553.

Wilson, I., Gick, B., 2006. Ultrasound technology and second language acquisition research. In: Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006), pp. 148–152.

Zharkova, N., Gibbon, F., Hardcastle, W., 2015. Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. Clin. Linguist. Phon. 29, 1–17.