Whalen, D. H., Iskarous, K., Tiede, M. K., & Ostry, D. J. (2003). A combined ultrasound/Optotrak measurement system for speech kinematics. In S. Palethorpe & M. Tabain (Eds.), *Proceedings of the 6th International Seminar on Speech Production* (pp. 308-313). Sydney: Macquarie University.

1328

A COMBINED ULTRASOUND/OPTOTRAK MEASUREMENT SYSTEM FOR SPEECH KINEMATICS

D. H. Whalen[1], Khalil Iskarous[1], Mark K. Tiede[1,2] & David J. Ostry[1,3]

[1] Haskins Laboratories
[2] Research Laboratory of Electronics, M.I.T.
[3] Dept. of Psychology, McGill University

ABSTRACT: Real-time measurement of the tongue during speech is needed for a thorough study of the speech system, yet its inaccessibility has limited our ability to image it. This paper reports on a new combination of techniques, ultrasound and an optoelectric motion measurement system. With infra-red-emitting diodes placed on the ultrasound transducer and the speaker's head, it is possible to localize the image of the tongue within the hard structures of the vocal tract. The palate can be imaged with the ultrasound having the speaker push water as high up in the mouth as possible, allowing the ultrasound to penetrate to the palate. This palate trace can then be coregistered with the head and with a sagittal MR image of that same talker, which provides some information about the posterior structures (velum and posterior pharyngeal wall). The palate trace imaged by ultrasound and the posterior structures imaged by MRI are transferred to the ultrasound speech images, despite head movement, since such movement is tracked optically and compensated for. The result is a fairly complete view of the tongue within the vocal tract. A comparison with other systems for imaging the vocal tract is included.

INTRODUCTION

Real-time measurement of the tongue is essential to knowing the full extent of the relation between speech articulation and the resulting acoustic signal. The tongue is the most important articulator for speech sounds and is necessarily involved in virtually every sound that is produced in speech. However, because it is hidden from view, measuring it has proven a lasting problem in the field of speech research. Only recently have new systems, and modifications of old systems, enabled us to accurately quantify a large volume of tongue movement data.

The earliest approach to obtaining information about just what the tongue was doing was static palatography (see Scholz, 1966 for an historical account), but the earliest method of seeing the tongue in action was x-ray motion pictures (Russell, 1928). The high levels of radiation necessary in those early days limited the usefulness of this approach, however. More modern cases with carefully designed protocols have provided much useful data (Botherel, Simon, Wioland, & Zerling, 1986; J. Perkell, 1969), with continuing usefulness (e.g., Magen, Kang, Tiede, & Whalen, 2003). Even more recently, improvements in registration of the x-rays has allowed for a new set of protocols available for research use with unimpaired humans (Ericsdotter, Stark, & Lindblom, 1999; Fitch, 2000). The time available for collection, however, is still limited.

Electropalatrography allowed for a dynamic measure of tongue/palate contact, which has provided much useful information, but measuring the tongue surface when there is no contact required other measures. The most successful approaches have used electromagnetometry (J. S. Perkell et al., 1992; Schönle et al., 1987) and x-ray microbeams (Kiritani, Itoh, & Fujimura, 1975; Westbury, 1994) to track fleshpoints on the tongue and other articulators. These systems allow for accurate tracking of these points even though they are not visible since the source (electromagnetic waves or x-rays) can travel through the tissues of the vocal tract. These systems do, however, track only a limited number of points, and the receivers or pellets that must be glued to the tongue are both time-consuming to apply and interfere somewhat with articulation.

Ultrasound allows for a look at the tongue in greater detail. Since the pioneering work of Maureen Stone (M. L. Stone, Faber, Raphael, & Shawker, 1992; M. L. Stone, Morish, Sonies, & Shawker, 1987;

M. L. Stone, Sonies, Shawker, Weiss, & Nadel, 1983), the usefulness and ease of data acquisition with these devices have improved greatly. The image quality has steadily improved, so that it is now possible to see the tongue surface without an act of faith. Sampling rate is held back somewhat by the current dependence on videocassette technology; as digital video becomes standard, the sampling rate will be limited only by the ultrasound machine itself (currently up to 150 Hz), not that of the recording device.

One remaining difficulty with ultrasound, however, is that it is most useful when it can be related to the hard structures of the vocal tract. Since the probe typically "rides" along with the jaw, the image of the tongue shows the shape and location of the tongue relative to the jaw. While there are many useful insights that can be obtained with that kind of measurement, it is even more useful to have the tongue positioned relative to the palate and the rear pharyngeal wall. The present paper discusses a system under development at Haskins Laboratories which overcomes this difficulty by including the 3-D position of visible markers on the head and probe so that changes in head position and/or probe position can be compensated for.

HOCUS – THE HASKINS OPTICALLY CORRECTED ULTRASOUND SYSTEM

The approach we took at Haskins Laboratories to correlating the ultrasound image of the tongue with the rest of the vocal tract was to use a 3-D optical tracking system to place the ultrasound images within a head-centered coordinate space. Further, the system can be used to construct 3-D reconstructions of the tongue across a sustained utterance, or multiple repetitions of the same utterance.

The Ultrasound Component

The ultrasound machine does not have many unusual requirements, and so it can be any one of many machines available on the market. Our particular model is an Aloka SSD1000 with a 3.5 mHz probe. This probe allows easy imaging of the tongue in either the sagittal or coronal orientation, with only light pressure under the chin being needed to maintain contact between skin and probe. The probe can be hand-held or attached with elastic cords tied over the head.

The Optical Component

The optical registration is done with the Optotrak® system (Northern Digital, Waterloo, Ontario, Canada), which uses three coupled CCD arrays to locate infra-red-emitting diodes (IREDs) in space. These diodes are approximately 5 mm in diameter and can be easily attached to the skin with double-sided tape. In order to separately locate the head and probe unambiguously, three diodes are required for each. Given sufficient separation between these IREDs, rigid body transformations can provide the orientation of the imaging probe relative to the head at the moment of an imaged data frame. We place the IREDs for the head on goggles that the speaker wears, which allows for some savings in setup time, and avoids possible stretching of the skin that might cause difficulty if the IREDs were attached to the skin directly. The three IREDs for the ultra-sound probe are attached to its handle. It is also possible to have six attached, so that the orientation can be changed from sagittal to coronal without having to change the probe's configuration.

Our current alignment program is written in Matlab. It uses the Optotrak data to compute the position and orientation of the probe relative to the head. For those paradigms in which the midsagittal tongue position is of primary importance, the resulting orientation angle can be used to eliminate frames in which the probe is too far off the midline, either through shifting or through tilt. A more positive use is the combining of coronal or parasagittal sections from sustained phonation or multiple repetitions of an utterance to generate a pseudo-cine 3-D tongue surface. Our first test of this has shown the reconstructed tongue to be quite similar to a measured, static 3-D shape obtained from MRI.

IREDs can also be attached to the lips, chin, and other exterior points to allow for the collection of data on other articulators as well, as has been done for over 20 years. Since the Optotrak system is 3-D, the IREDs do not have to be in the same plane as the ultrasound probe. Thus if information about the movement of the corners of the mouth is desired, IREDs can be placed there even if the tongue is only imaged midsagittally. Information about jaw movement can be obtained from a chin IRED (though there is likely to be some sliding of the skin across the jaw) or as a result of the movement of the probe

itself as it "rides" on the jaw. Since the tissue under the chin is compressible, the probe height does not give a completely accurate measure of jaw height, but it can give an indication that could be useful.

The Hard Structures

The hard palate can be imaged with the ultrasound by having the speaker force a bolus of water as high in the mouth as possible. This allows the signal to pass through the water and image the hard palate. Since the relationship between the head (as measured with the diodes) and the palate does not change, this palate trace can be done once at the beginning of a session and applied to the remaining images of the tongue. (We obtain palate traces at the end of sessions as well to determine whether there was any slipping of the head IREDs during the experiment. We have yet to determine how great a discrepancy would be enough to invalidate our results.) If we have MRI of the speaker's vocal tract, we can also superimpose it onto the image for additional palatal verification. In the best case, this will give us the position of the rear pharyngeal wall as well. However, there may be changes in the angle between the hard palate and the rear pharyngeal wall due to posture (both the supine/upright difference and differences within upright) that will make this less feasible.

Time Alignment

Currently, time alignment is being accomplished via acoustic matching, aided by a sharp noise at the beginning of a session, made possible by the recording of the audio signal by both the optical and ultrasound systems. It may be that this will ultimately be unsatisfactory and a more complete and consistent means of synchronizing the systems will be constructed.

COMPARISON OF VARIOUS TONGUE IMAGING SYSTEMS

The various ways of imaging the vocal tract have different strengths and weaknesses. A sampling of these systems and these features is given in Table 1. This list is not exhaustive, and many of the systems change quite rapidly , so this comparison can only be considered temporary. Nonetheless, it gives a quick overview of the many ways now available to image the tongue and parts of the rest of the vocal tract. Those who want to include such measurements in their research can see some of the relevant points to note for deciding on a system from this chart. For example, if the target speaker population consists entirely of elderly speakers who are not in the best of health and may only tolerate being measured for a short amount of time, then the "setup time" column becomes very important. If research money is in short supply, then the costs per session may be the most important.

Note that there are three rows for ultrasound. The first row is simply labelled "ultrasound," and is intended to cover those systems in which the tongue is measured in jaw-relative space (Gick, 2002; M. L. Stone et al., 1987) or else in conjunction with a videotape-based correction system (Gick & Campbell, 2003). The second row is for those systems that use a combination of a holder for the ultrasound probe and a head restraint for the speaker (M. Stone & Davis, 1995). The third row applies to the system described here. Each has a different set of costs, constraints and uses.

CONCLUSION

The combination of video tracking and ultrasound imaging of the tongue is allowing us to take an unprecedented look at the entire vocal tract in real time. The combination promises a great deal of flexibility in selecting speakers and situations. The comparison across platforms has already proven useful (as with the comparison of 3D reconstructions from ultrasound and MRI), and will continue to be in the future. For example, being able to extract complete tongue shape information from MRI is useful but expensive; that kind of data can be combined with ultrasound to allow for inferences from the more inexpensive ultrasound dataset based on what we know from the MRI. All of these systems allow much more complete measurements of the vocal tract (especially used in conjunction) than has ever been possible in the past. We are certain to see an increase in the number and variety of descriptions of the world's languages with these technologies.

Table 1: Comparison of various speech measurement systems that image the tongue.

Feature / System	Tongue Imaging	Sampling Rate	Machine Cost/ Availability	Session Cost	Portability
2D Magnetometry	Flesh Points (usually 4)	200-500 Hz	~$70,000/ available	~$200[3]	Lab only
3D Magnetometry	Flesh Points (usually 4)	Usually 200 Hz	~$90,000/ pending[2]	~$200[3]	Lab only
Ultrasound	Full-length[1]	30 Hz[5]	$30-70,000/ available	~$5[4]	Can use portable system
Ultrasound with head holder	Full-length[1]	30 Hz[5]	$30-70,000/ available	~$5[4]	Lab only
Ultrasound/ Optotrak (HOCUS)	Full-length[1]	30 Hz[5]	~$80,000/ available	~$5[4]	Lab only
Static MRI	Full-length	---	~$1,500,000	~$500	Lab only
Cine MRI	Full-length	8-24 Hz	~$1,500,000	~$500	Lab only
X-Ray Microbeam	Flesh Points (usually 4-5)	40-160 Hz	Only available at U. Wisconsin	Free	Lab only

Notes:

[1]The ultrasound image extends from just above the hyoid bone to near the tip of the tongue. Whenever the signal hits air, it disappears, so any portion of the tongue tip that is over the sublingual cavity will not be imaged. Similarly, retroflex tongue shapes are not well imaged. Note, however, that the coverage of the tip may be similar to that of magnetometry and microbeam, since the receivers for the tip are placed 1 cm posterior to the actual tip; an experiment is under way to see how often this point is imaged by ultrasound.

[2] This has not yet been delivered in an operable form.

[3] This is the approximate cost of two receivers, which is a typical number that need to be replaced after each experiment.

[4] This is for supplies, such as latex gloves, ultrasound gel, etc.

[5] The internal combinations of settings result in a machine internal sampling rate of anywhere from 30 to 100 Hz. Currently, the image must be recorded on videotape, which runs at 30 Hz for North American videotape. It is possible, using machine internal storage, to obtain the higher sampling rates for limited periods of time.

Table 1 (continued)

Feature / System	Setup Time	Imaging Tongue Root	Imaging Velum	Head Move-ment	Special Features of Speech	Length of Data Collection	Special Popula-tions
2D Magne-tometry	~60 min.	No	No	Restr.	Receivers affect artic.	Up to 2 hours[6]	Limited
3D Magne-tometry	~60 min.	No	No	Free	Receivers affect artic.	Up to 2 hours[6]	Limited
Ultrasound	2-20 min.	Yes	No	Free	Probe slightly impinges on jaw.	Limited only by speaker	Broad
Ultrasound with head holder	~20 min.	Yes	No	Restr.	Some effect on jaw.	Limited only by speaker	Fairly broad
Ultrasound/ Optotrak (HOCUS)	~20 min.	Yes	No	Free	Probe slightly impinges on jaw.	Limited only by speaker	Fairly broad
Static MRI	~10 min.	Yes	Yes	Restr.	Supine position.	Usually 1 1/2 hour	Limited
Cine MRI	~10 min.	Yes	Yes	Restr.	Supine position.	Usually 1 1/2 hour	Limited
X-Ray Microbeam	~60 min.	No[7]	No[8]	Free	Some effects on speech[9]	12 min. of speech	Fairly broad

[6] The time limit is usually the length of time that most of the receivers stay glued on; for 2-D systems, there are also limits on how long the speaker can keep the head immobilized .

[7] Some early speakers had a pellet that was fairly low on the tongue root; this has not been the case recently.

[8] It is possible to suture a pellet to the underside of the velum, but this is very seldom done.

[9] Weismer and Bunton (1999) found that only a few individuals had noticeable effects of the pellets on their production of one sample sentence. Some other subgroups had tendencies toward perceptible effects.

Note that we have not described some of the full-body x-ray systems now in use. There are a variety of them, with different parameters, making it difficult to include them in the table.

ACKNOWLEDGEMENTS

This research was supported by NIH grant DC-02717 to Haskins Laboratories. We thank Bryan Gick for helpful comments.

REFERENCES

Botherel, A., Simon, P., Wioland, F., & Zerling, J.-P. (1986). *Cinéradiographie des voyelles et consonnes du français.* Strasbourg: Institut de Phonétique de Strasbourg.

Ericsdotter, C., Stark, J., & Lindblom, B. (1999). Articulatory coordination in coronal stops: Implications for theories of coarticulation. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (Vol. 3, pp. 1885-1888). San Francisco: University of California, Berkeley.

Fitch, W. T. (2000). The phonetic potential of nonhuman vocal tracts: Comparative cineradiographic observations of vocalizing animals. *Phonetica, 57*, 205-218.

Gick, B. (2002). The use of ultrasound for linguistic phonetic fieldwork. *Journal of the International Phonetic Association, 32*, 113-121.

Gick, B., & Campbell, F. (2003). Intergestural timing in English /r/. In M. J. Solé, D. Recasens & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1911-1914). Barcelona: Causal Productions Pty Ltd.

Kiritani, S., Itoh, K., & Fujimura, O. (1975). Tongue-pellet tracking by a computer-controlled x-ray microbeam system. *Journal of the Acoustical Society of America, 57*, 1516-1520.

Magen, H. S., Kang, A. M., Tiede, M. K., & Whalen, D. H. (2003). Posterior pharyngeal wall position in the production of speech. *Journal of Speech, Language, and Hearing Research, 46*, 241-251.

Perkell, J. (1969). *Physiology of speech production: Results and implications of a quantitative cineradiographic study.* Cambridge, MA: MIT Press.

Perkell, J. S., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992). Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America, 92*, 3078-3096.

Russell, G. O. (1928). *The vowel: Its physiological mechanism as shown by x-ray.* Columbus, OH: Ohio State University Press.

Scholz, H. J. (1966). Zur Geschichte der Palatographie. *Phonetica, 15*, 110-121.

Schönle, P., Grabe, K., Wenig, P., Hohne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language, 31*, 26-35.

Stone, M., & Davis, E. P. (1995). A head and transducer support system for making ultrasound images of tongue/jaw movement. *Journal of the Acoustical Society of America, 98*, 3107-3112.

Stone, M. L., Faber, A., Raphael, L. J., & Shawker, T. H. (1992). Cross-sectional tongue shape and linguopalatal contact patterns in [s], [ʃ], and [l]. *Journal of Phonetics, 20*, 253-270.

Stone, M. L., Morish, K., Sonies, B., & Shawker, T. (1987). Tongue curvature: A model of shape during vowel production. *Folia Phoniatrica, 39*, 302-315.

Stone, M. L., Sonies, B. C., Shawker, T. H., Weiss, G., & Nadel, L. (1983). Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system. *Journal of Phonetics, 11*, 207-218.

Weismer, G., & Bunton, K. (1999). Influences of pellet markers on speech production behavior: Acoustical and perceptual measures. *Journal of the Acoustical Society of America, 105*, 2882-2894.

Westbury, J. R. (1994). *X-ray microbeam speech production database user's handbook* (Manuscript): Madison, WI: Waisman Center, University of Wisconsin.