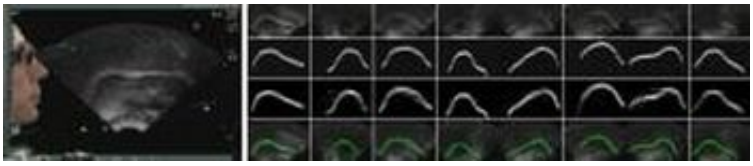
A Test in Producing a Visual Capture of Speech

June 14 2010, By La Monica Everett-Haynes



The top row in the image to the right indicates ultrasound inputs. The second row represents tongue contours drawn by a human. The third row represents the automatic system's raw output, with the bottom row showing the contour after they had been processed. (Image courtesy of: Diana Archangeli and Ian Fasel)

(PhysOrg.com) -- Diana Archangeli, a UA linguistics professor, is heading up a team using ultrasound and other devices to create a technology that would enable the detection of words without auditory cues.

Simply studying how people speak will not lead to the best understanding of why some individuals have difficulty pronouncing certain words or learning a second language.

To address this challenge, a University of Arizona research team is using ultrasound and other equipment to produce technology that would map the mouth's interior to aid in analyzing exactly how sound is produced.

Diana Archangeli, a UA linguistics professor, is heading up the project,

"Arizona Articulatory, Acoustic, and Visual Speech Database." It is meant to improve what is known about how words are formed in the mouth and, thereby, advancing what is known about speech.

"Learning a language involves mastering the exquisitely-timed coordination of multiple articulators - lips, tongue, velum and glottis - yet all but the lips are invisible to the language learner, hidden within the mouth," Archangeli said.

"When someone is listening to speech, there is often a lot of other noise going on," she said. "One question we are trying to figure out is what we do when the audio signal isn't enough?"

The project has implications for different types of language research, people learning to play certain wind instruments and may eventually aid individuals who have had their larynx removed.

The team recently earned a $30,000 Arts, Humanities & Social Sciences Grants for Faculty grant, a UA funding program established to aid University researchers in transitioning promising projects from conception to application. Other members include Ian Fasel, an assistant research professor of computer science, and Jeff Berry and Jae Hyun Sung, both graduate students in the linguistics department.

For now, Archangeli's team is using equipment to reocrd lip and tongue movements by capturing video recordings of the mouth, jaw and tongue while simultaneously recording the audio signal and measuring both vocal fold vibration and nasal airflow.

"If we want to understand how people make sound, you want to measure everything," Fasel said.

While other technologies exist to capture such images, including X-ray

and magnetic resonance imaging, or MRI, the team opted to use ultrasound - positioned between the chin when recording data - because it is non-toxic, and portable, both for fieldwork and classroom purposes.

The team's research will be fed into a new database to be called TIMIT-UA, which will expand upon TIMIT, a widely used speech recognition database that has been supported by the work of researchers at Texas Instruments and the Massachusetts Institute of Technology. The UA's contribution is the addition of ultrasound and video data.

Though the project only recently earned grant funding, the team's work already has been accepted for presentation.

Fasel co-authored a paper that been accepted for presentation during the Computational Neuroscience Meeting to be held in July in San Antonio.

Next month, Archangeli, Fasel and Berry's collaborative research will be presented at Laboratory Phonology in Albuquerque. Also, another paper co-authored by Fasel and Berry is slated to be presented at the International Conference on Pattern Recognition, to be held in Istanbul, Turkey in August.

Paul Cohen, the UA computer science department head, said he expects the team members and their research - particularly with the database - will garner increased attention.

"I am quite pleased that this work is concerned equally with advancing our scientific understanding of human production and perception of language as it is with potential technological applications," said Cohen, who also directs the UA School of Information Sciences, Technology and Arts.

He said the project is a clear example of researchers utilizing advanced

machine learning technologies with broad-based uses and to improve knowledge in the social sciences.

For instance, it takes a skilled researcher about 20 minutes to manually draw a four-second video of tongue contours - a taxing and time-consuming effort to improve speech recognition, said Berry, a doctoral degree student.

But producing a computer program that would be able to collect and interpret such data more quickly and without human intervention - as the team intends - would be a boon, Berry said. The team intends for its technology, when complete, to be able to capture and trace 30 frames per second in real time.

Fasel and Archangeli said this is the type of application that would be especially helpful in classrooms where educators are teaching language or music.

Archangeli offered for an example the different ways to articulate the sound of the sound "R" in speaking in English, one in which the tongue slopes upward; the other in which the tongue bunches up, forming a dome just behind the teeth.

This, Archangeli said, may contribute to reasons why “R” is a challenging sound for some learners of English to master.

But if learners are able to see their tongues in motion, as the team's technology allows, this would help improve language acquisition, Fasel said.

"Now, an expert is required," Fasel added. "But if you had a computer system to do this, you can imagine this being in classrooms and helping students learn."

Provided by University of Arizona