様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２６ 年　 ８ 月 １５ 日現在

機関番号：２１６０２
研究種目：基盤研究(C)
研究期間：2011～2013
課題番号：２３５２０４６７
研究課題名（和文）舌の超音波画像及び顔・唇のビデオ画像による視覚音声認識

研究課題名（英文）Visual speech recognition using ultrasound tongue and video lip/face images

研究代表者
Ｗｉｌｓｏｎ　Ｉａｎ（Wilson, Ian）

会津大学・コンピュータ理工学部・教授

研究者番号：５０４４４９３０

交付決定額（研究期間全体）：（直接経費）　 3,700,000円、（間接経費）　 1,110,000円

研究成果の概要（和文）：（１）顎の動きを集めたビデオのデータに関して、CVC音節の中にある母音での下顎皮膚の伸張を計測したところ、頭子音だけが大きな影響力を持つ事が分かった。（２）英語を話す際の舌の位置を集めた超音波のデータに関して、ネイティブスピーカーは英語を母国語としない日本人より、より効率のいい位置で舌を休ませる。（３）MUTISと呼ぶ、特徴空間を写像し解釈する最適な方法に着目したことに関して、MUTISのより高次元な場所は最も人々を識別することに有効であること、VSSの低次元な場所のデータは主に音素を識別することに最も有効であることを示した。VSSのデータの軌道はＬ１とＬ２との違いをはっきりと示している。

研究成果の概要（英文）：There are three main results of our research: (1) Related to video data collection of jaw movement, when measuring the amount of skin stretching over the mandible for the vowel in a CVC syllable, the onset consonant (but not the coda consonant) has a significant effect. (2) Related to ultrasound data collection of tongue position when speaking English, native (L1) speakers rest their tongue in a more efficient location (closer to the median position for English speech sounds) than Japanese (L2) speakers do. (3) Related to our focus on how best to construct and interpret a feature space we call MUTIS (midsagittal ultrasound tongue image space), results indicated that higher dimensions of MUTIS are most effective for identifying people, and that primarily the lower dimensions of VSS (vocal sound space) data are most effective for identifying phonemes. Trajectories within the VSS data indicate clear differences between L1 and L2 speakers, but not within the MUTIS data alone.

研究分野：人文学

科研費の分科・細目：言語学・音声学

キーワード：ultrasound　video　tongue　articulation　jaw

１．研究開始当初の背景

Speech recognition, the ability of a computer to analyze the acoustic signal and determine what words have been spoken, has reached a high degree of accuracy – about 95% or more, according to Nuance Communications, developer of products such as Dragon NaturallySpeaking and Dragon Dictate. However, speech recognition based on moving images of the vocal tract (e.g., videos of lip/face movement) is still at a low level of accuracy – about 75% for some tasks (Hilder et al., 2009). Given the fact that the acoustic signal is a direct consequence of the motion of the articulators, and most of what is produced inside the vocal tract is also visible in the face (Yehia et al., 1998), why is the accuracy so much lower for visual speech recognition? Is it possible to recognize speech from the midsagittal movements of the tongue, instead of the movements of the face? If these two types of visual data (face/lip and tongue data) are combined, does visual speech recognition accuracy reach the same level as that of audio speech recognition? These are the questions that motivated our research.

In the history of phonetics and language-learning literature, most textbooks show midsagittal figures of the vocal tract. Chomsky and Halle's (1968) famous phonological feature set was largely based on the position of the tongue from a midsagittal plane perspective. However, this midsagittal bias may simply be due to the fact that our first imaging methods (e.g., x-ray) worked best in this plane. We do not have data that shows how well midsagittal movies of the tongue predict the acoustic signal. This is partly due to the fact that an imaging method has not been available to safely view the tongue's movements with good clarity. With the advance of ultrasound imaging of the tongue (allowing whole-tongue images, not simply point-tracking), such a method now exists, and we can more accurately map the relationship between midsagittal tongue shape/position and the acoustic signal.

If there is a strong correlation between the tongue's midsagittal shape/movements and the acoustic signal, then the bias toward the midsagittal plane is justified and future research, including articulatory speech synthesis would be simplified. However, if no such correlation exists, then this would have strong implications for the focus of future work in phonetics and phonology – namely, that we should not simply focus on the midsagittal plane, but consider the whole tongue/airspace, and textbooks would have to change to reflect this.

Computer lip-reading, more accurately called speech-reading, the ability of a computer to recognize speech using only the visual signal, has attracted many researchers. Computers that could do voice recognition in noisy environments, or voice recognition from only the video signal, would be valuable in a number of applications: military defense applications where silent speech is necessary, video surveillance / monitoring for anti-terrorism and law-enforcement applications, enhanced pronunciation evaluation systems, etc.

２．研究の目的

Newman and Cox (2009) recently showed that computers can distinguish between languages just based on speech-reading (i.e., analyzing images of the face), and this has generated much excitement in the speech research world. If this is possible, is it also possible just based on "tongue-reading" (i.e., processing of midsagittal tongue images during speech)? If so, we can generalize about the differences in tongue shape/movement between languages, and this has implications for the way foreign languages are taught and acquired.

One of the goals of the proposed research was to make a detailed description of the factors (other than differences in phonetic inventory between languages) that differ between languages - e.g., speed of the tongue, general tongue location – high/low/front/back, midsagittal tongue area, which part of the tongue is most active, etc. This work followed directly from previous JSPS kakenhi research that developed ways of measuring articulatory setting of the tongue in different languages.

３．研究の方法

1) Train and work with research assistants to develop the method of head motion tracking/correction so that ultrasound tongue data is maximally reliable.

2) Train and work with research assistants to analyze ultrasound video data of the tongue's movements.

3) Test various image-processing algorithms to find one that is best able to track the 2-D midsagittal tongue motion.

4) Train and work with research assistants to develop the image-processing method of tracking the lip and jaw movement during speech.

5) Collect both ultrasound data of the tongue and video data of the lips/jaw, and analyze this data to discover differences between native and non-native speech patterns.

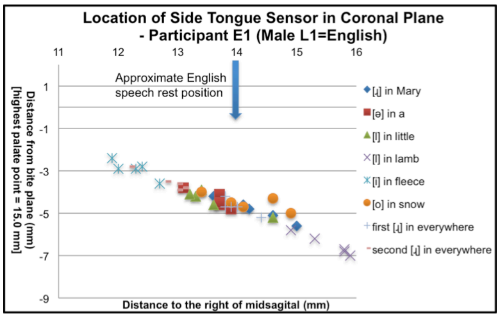
6) Write up results for presentation at conferences and publication in journals.

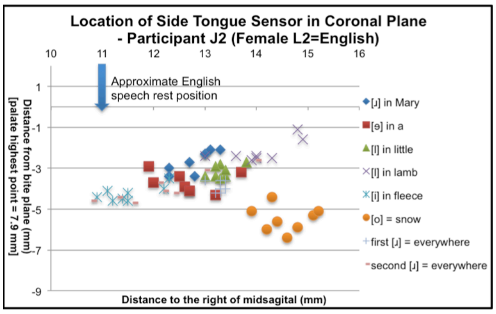
4．研究成果

A simple, inexpensive method of inferring movements of the mandible is to use video tracking of a chin marker during speech. One potential problem with video tracking of a chin marker, however, is that it records skin movement, not necessarily mandible movement. Since the skin stretches over the mandible during production of some speech sounds, especially labial consonants, one must exercise caution when inferring mandible movement from the position of a marker on the chin. In an experiment to measure the degree of skin stretching, we found that the onset consonant affected the degree of stretching, but not the coda consonant (see table below).

	onset [p]vs[t]vs[k]	vowel [æ]vs[ɪ]	coda [p]vs[t]vs[k]
male speaker	p = 0.046	p ＜ 0.001	p = 0.202
female speaker	p = 0.002	p = 0.040	p = 0.095

In another experiment, we looked at the rest position of the tongue during pre-speech posture, and found that it was more efficient for native speakers of a language than for second-language learners. A native speaker rests his/her tongue in the center of where it is required for speech sounds, but non-native speakers were found to have a more narrow tongue than was required for most speech sounds (see figures below).





The upper figure is for a native English speaker. The blue arrow indicates the rest pre-speech position of a marker on the side of the tongue (in the coronal view). The lower figure is for a Japanese speaker of English. Note that the side marker is much closer to the midline, meaning that the tongue is narrower when at rest. In each figure, the colored marks indicate the position of the tongue for various English speech sounds.

5．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計９件）

Wilson, I. & D. Erickson. 2013. Effect of syllable onset, coda, and nucleus on degree of skin stretching over the mandible. Proceedings of Meetings on Acoustics, vol.19. doi:10.1121/1.4799467

Moriya, S., Y. Yaguchi, N. Terunuma, T. Sato, & I. Wilson. 2013. Normalization and matching routine for comparing first and second language tongue trajectories. Journal of the Acoustical Society of America, vol.134, No.5, Pt.2, p.4244. doi:10.1121/1.4831607

Kanada, S., I. Wilson, B. Gick, & D.

Erickson. 2013. Coarticulatory effects of lateral tongue bracing in first and second language English speakers. Journal of the Acoustical Society of America, vol.134, No.5, Pt.2, p.4244. doi:10.1121/1.4831608

Moriya, S., Y. Yaguchi, N. Terunuma, T. Sato, & I. Wilson. 2013. 舌特徴空間における言語学習者の違いを比較するための正規化とマッチング手法. IEICE Technical Report, vol.113, No.308, SP2013-80, pp.53-57.

Wilson, I., D. Erickson, & N. Horiguchi. 2012. Articulating rhythm in L1 and L2 English: Focus on jaw and F0. Proceedings of the 2012 Autumn Meeting of the Acoustical Society of Japan (ASJ), pp.319-322.

Yaguchi, Y., N. Horiguchi, & I. Wilson. 2012. Finding phoneme trajectories in a feature space of sound and midsagittal ultrasound tongue images. In IEEE Proceedings of the 4th International Conference on Awareness Science and Technology (iCAST 2012), pp.156-162. doi:10.1109/iCAwST.2012.6469606

Abe, Y., I. L. Wilson, & D. Erickson. 2012. Video recordings of L1 and L2 jaw movement: Effect of syllable onset on jaw opening during syllable nucleus. Journal of the Acoustical Society of America, vol.132, No.3, Pt.2, p.2005. doi:10.1121/1.4755428

Okada, J., I. L. Wilson, & M. Yoshizawa. 2012. Pitch and intensity in the speech of Japanese speakers of English: Comparison with L1 speakers. Journal of the Acoustical Society of America, vol.132, No.3, Pt.2, p.2004. doi:10.1121/1.4755421

Sano, K., Y. Yaguchi, & I. Wilson. 2012. Comparing L1 and L2 phoneme trajectories in a feature space of sound and midsagittal ultrasound tongue images. Journal of the Acoustical Society of America, vol.132, No.3, Pt.2, p.1934. doi:10.1121/1.4755107

〔学会発表〕(計４件)

Wilson, I., J. Villegas, & T. Doi. Lateral tongue bracing in Japanese and English. Paper presented at Ultrafest VI, Edinburgh, Scotland. (2013.11.08)

Erickson, D. & I. Wilson. Articulatory and laryngeal contributions to rhythm in English. Poster presented at the Joint Research Meeting of the Dept. of Linguistic Theory and Structure, NINJAL, Tokyo, Japan. (2013.03.02)

Wilson, I. & N. Horiguchi. How accurately people follow articulation instructions. Paper presented at the 4th Pronunciation in Second Language Learning and Teaching conference (PSLLT 2012), Vancouver, Canada. (2012.08.24)

Yaguchi, Y., N. Horiguchi, & I. Wilson. 発音習得のための超音波舌画像に対する音素片マッピング [Mapping phonemes to midsagittal tongue images for pronunciation learning]. Paper presented at the joint meeting of the Technical Committees for Pattern Recognition and Media Understanding (PRMU) and Signal Processing (SP) of the Institute of Electronics, Information and Communication Engineers (IEICE), Sendai, Japan. (2012.02.10)

〔産業財産権〕
○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://clrlab1.u-aizu.ac.jp/index_j.html

６．研究組織
(1)研究代表者
ウィルソン　イアン　（Wilson　Ian）

研究者番号：50444930