Kuwait J. Sci. Eng. **33**(2) pp. 69-79, 2006

Articulatory models of Arabic vowels computed from magnetic resonance images

M. MOHAMMAD[1], T. SINAN[2], B. MOHAMMAD[3]

1 Public Authority for Applied Education and Training, College of Technological Studies, Electronics Engineering Department, P.O.Box 42325, Shuwaikh 70654, Kuwait.

[2] Kuwait University, Faculty of Medicine, Department of Radiology, Jabria, Kuwait.

[3] Public Authority for Applied Education and Training, College of Health Sciences, Department of Medical Records, Shuwaikh, Kuwait.

ABSTRACT

This paper presents the preliminary results of a long term project aimed at providing a database of articulatory models for the Arabic language. The articulation of 31 Arabic sounds was captured, and the three main Arabic vowels were studied using medical images of the vocal tract from four subjects. Magnetic Resonance Imaging (MRI) was used in this study to capture the shape and positions of speech articulators. The low temporal resolution of MRI limited this and most speech studies to sustained sounds. The images were processed, segmented, and the vocal tract models were computed. The resulting area functions are used for the calculation of formant frequencies. The results show expected natural intra-subject variabilities, but are generally closely matched and in line with the theory of vowel production. The computed models were used to generate the transfer functions and synthesize artificial Arabic vowel sounds. The quality of the synthesized sounds was confirmed to be intelligible.

Keywords: Arabic speech synthesis; MRI; vowels

INTRODUCTION

As speech is the principal means of human communication, the study of speech production is a topic of interest to many disciplines such as linguistics, neurology, psychology, physiology and engineering. The shape of the vocal tract changes with the position of the articulators, which consist of lips, tongue, velum, larynx and jaw. The aim of speech production research is to understand the relationship between articulatory configurations and acoustical output of both normal and pathological human speech. From an engineering point of view, the vocal tract can be thought of as a resonator that filters the sounds produced by different sources such as the periodic sounds produced by the vocal

folds and noisy sounds produced at the teeth or lips. The shape of the vocal tract and thus the characteristics of the acoustic filter are determined by the positions of the articulators.

Knowledge of the speech production mechanism is essential for development of mathematical models that can be used to produce artificial speech indistinguishable from human speech. The main obstacle to acquiring that knowledge is the inaccessibility of the vocal tract. In order to overcome this obstacle, a suitable measurement technique is needed to produce a database of quantitative measurements of articulator positions during speech. If the database was large enough to cover inter-and intra-subject variability, existing models would be improved greatly, leading to intelligible synthetic speech. The knowledge acquired from speech production modeling has direct clinical, linguistic, and industrial implications (Westall & Johnston 1996). Articulatory speech synthesis, which holds the lowest bit rate coding, can be used, for example, to replace humans in routine jobs, which is already done by non-articulatory methods of synthesis.

Various methods have been used to image and measure the vocal tract. X-rays, Ultrasound and Magnetic Resonance Imaging (MRI) (Narayanan 1996) are some of the most commonly used methods. All three have significant advantages and disadvantages: X-rays involve health risks and are poor at imaging soft tissues; ultrasound delivers poor resolution pictures, attenuates rapidly in bone and reflects at air (and thus cannot image much more than the tongue), but does capture soft tissue and movement well and MRI delivers high quality imaging of soft tissue, has the ability to select imaging planes at any angle with no health risks, but requires long acquisition time, and does not distinguish between bone and air.

The advantages of MRI over other imaging modalities have encouraged speech researchers to use it for accessing the vocal tract despite the low temporal resolution. Therefore, MRI was mostly used in speech studies for capturing static images of sustained speech sounds. Since the study by Baer's *et al.* (1991), many speech researchers reported using MRI to study the articulatory configurations for various sounds from different languages. Story *et al.* (1996) conducted a study on English sounds, where sets of axial and coronal images from one male subject were collected to show the vocal tract configurations for 12 vowels, 3 nasals, and 3 plosives. The collected images were considered averaged and could have been centralized due to the long scanning procedure and the resultant fatigue effects. However, it was shown that the simulated formant frequencies produced from segmented image data corresponded well with the formants from recorded natural speech. Kröger *et al.* (2000) reported another study, where six German vowels were examined using sagittal MR images from one male subject. A grid system was used to compute the cross-sectional areas by dividing the vocal tract into sections that are perpendicular to

airway midline. The resultant synthesized speech was compared to acoustical measurements and were shown to be closely matched. The literature is rich with comparable studies, however, the articulation of Arabic speech sounds was rarely analyzed. In this paper, we report using MRI modality to capture the vocal tract shapes during sustained production of 31 Arabic sounds. Vocal tract area functions were measured from the images and used to compute the spectrum to finally produce synthetic Arabic vowel sounds. We note that our focus was on the vowels since they play a major role in characterizing a language. It has long been accepted that there are six main Arabic vowels (Al-Ani 1970), including the three short vowels /a/, /i/, and /u/ that differ only in duration from their three long counterparts /ā/, /ī/, and /ū.

METHODS

Magnetic resonance images were collected using a GE 1.5 Tesla SIGNA machine. An anterior neck RF coil was placed around the neck-mouth area. The subject was prompted through the intercom channel to start producing the speech corpus and sustain the target phoneme on the second repetition. If the phoneme could not be sustained, such as plosive sounds (i.e. [p]), the subject was asked to hold the vocal tract position for the target sound on the second repetition. In this instance, the scan start key was pressed in the control room. A sagittal set of 10 adjacent images were collected for each sound. The image slice thickness was 6 mm without spacing. Scanning time was around 20 seconds for each set. Four native Kuwaiti Arabic speakers [2 males (RA, RK) and 2 females (WA, ZA)] volunteered as subjects for this study. Volunteers had no prior phonetic training and their ages ranged from 22 to 37 years old. Table 1 provides a list of the speech corpus consisting of 31 Arabic words, where the target sounds are underlined.

Table 1: The speech corpus with the target sounds underlined and phonetically translated.

1	[bab]	/ā/	باب	12	[nayi]	/y/	ناي	23	[sareer]	/r/	سرير
2	[karim]	/ī/	كريم	13	[elah]	/h/	إله	24	[hesab]	/b/	حساب
3	[sabur]	/ū/	صبور	14	[aziz]	/z/	عزيز	25	[wedad]	/d/	وداد
4	[jafaf]	/f/	جفاف	15	[rathath]	/ð/	رذاذ	26	[aswat]	/t/	أصوات
5	[athath]	/θ/	أثاث	16	[hafatha]	/ḍ/	حفظ	27	[bayad]	/ḍ/	بياض
6	[ras]	/s/	راس	17	[musaq]	/ġ/	مصاغ	28	[matat]	/ṭ/	مطاط
7	[rusas]	/ṣ/	رصاص	18	[shua'a]	/σ/	شعاع	29	[sharik]	/k/	شريك
8	[rushash]	/š/	رشاش	19	[zaman]	/n/	زمان	30	[funduq]	/q/	فندق
9	[jana'h]	/ḥ/	جناح	20	[kam]	/m/	كم	31	[qura'a]	?	قرأ
10	[tarikh]	/x/	تاريخ	21	[hajeej]	/j/	حجيج				
11	[law]	/w/	لو	22	[lail]	/L/	ليل				

RESULTS

Ten adjacent sagittal scans were collected for each sustained sound from all four subjects. Seven scans covered the vocal tract cavity for male subjects and six for females which left us with three or four out-of-range scans which were then discarded. A mid-sagittal scan was selected to give a preliminary view of the vocal tract configuration for the sustained sounds. Figure 1 shows seven sagittal images for subject RA while sustaining /$\overline{\text{u}}$/. Figure 2 shows 32 mid-sagittal images for the subject RA while at rest (no-speech) and while sustaining the 31 phonemes. The images give a general description of tongue positions, such as the low-front /$\overline{\text{a}}$/, the high-front /$\tilde{\text{I}}$/and the high-back /$\overline{\text{u}}$/. A three-dimensional view of the vocal tract cavity can be constructed from the rest of the sagittal images, which show more details such as tongue curvature.

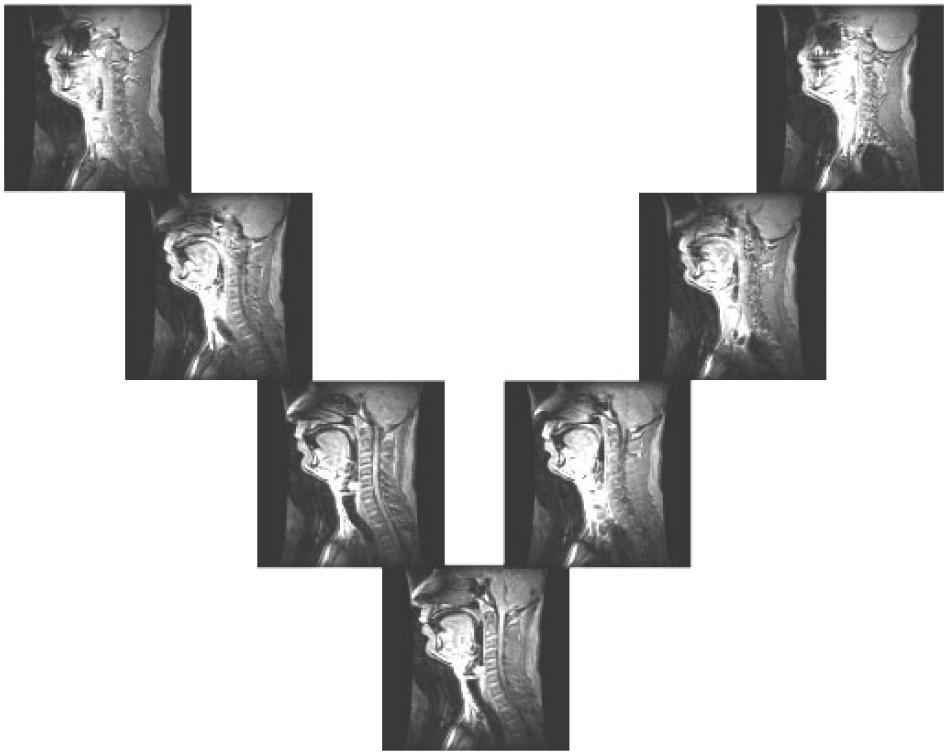


Fig. 1: Seven sequential sagittal views of subject RA while sustaining /$\overline{\text{u}}$/ with the mid sagittal view at the bottom center.

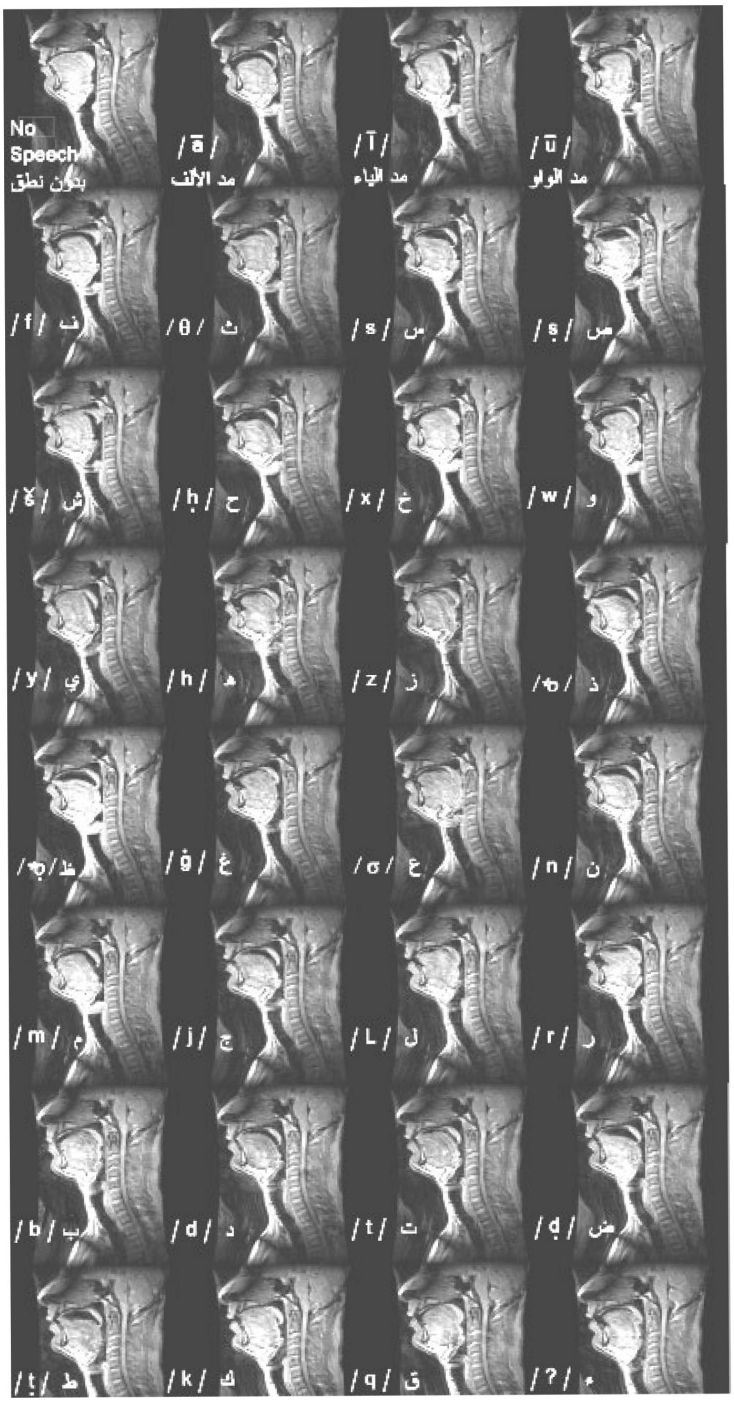


Fig. 2: Mid-sagittal views (32) of subject RA while sustaining 31 Arabic sounds with a reference no-speech image.

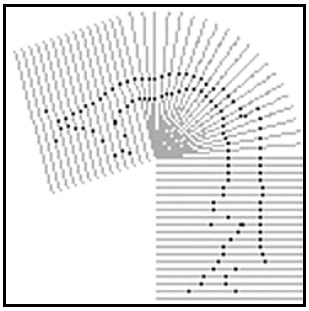
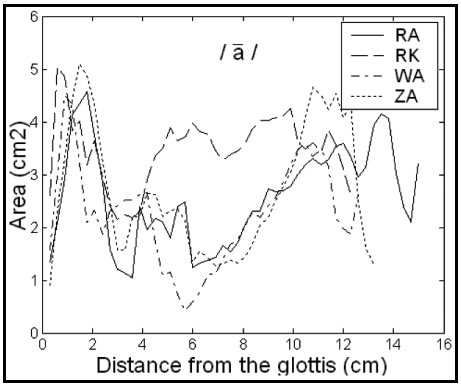
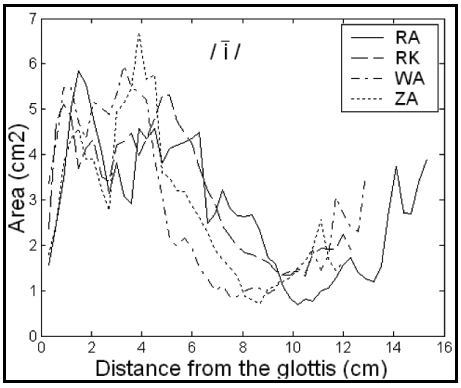


Fig. 3: The adopted grid system on top of the segmented mid-sagittal vocal tract for subject RA while sustaining /ū/.

The collected images were processed and the vocal tract was segmented by tracing the edges manually. A grid system was adopted to divide the vocal tract into sections perpendicular to the center of the vocal tract. The grid size was chosen to cover the whole vocal tract for all subjects; and the tongue center was measured, for each subject and sound, to determine the location of the grid. Figure 3 shows 55 sections aligned perpendicularly to the center of the vocal tract cavity. However, the actual grid size or the number of sections that are actually used for each model depends on the vocal tract length for each sound. The centered grid consists of a maximum of 20 horizontal sections in the larynx area separated by 3mm, 20 sections in the pharynx area separated by 5° from the center of the tongue and a maximum of 15 sections in the tongue front area (5° from vertical) separated by 3mm. When vocal tract side branches are excluded, two intersection points should exist between a grid line and the two edges of the main vocal tract. These points are extracted from all sagittal images for each grid line and used, along with image resolution parameters, to compute the cross-sectional area. The vocal tract is modeled as a set of concatenated tubes of variable cross-sectional area, which is commonly referred to as the area function (Fant 1960). Table 2 is a list of the area functions for all four subjects and all three vowels, excluding the side branches. Table 2 also lists the vocal tract lengths at the bottom. Figure 4 shows three diagrams for the vowels /ā/, /ī/, and /ū/ of the area functions from the four subjects, excluding side branches of the vocal tract.





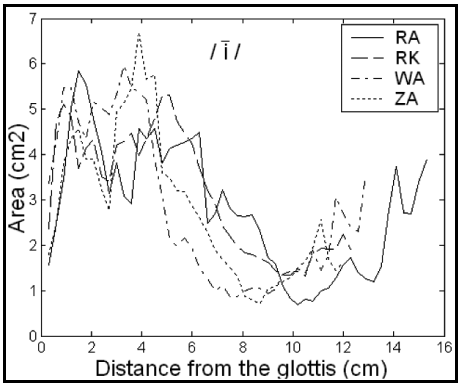


Fig. 4: Computed area functions (excluding side branches) from the images of four subjects for the vowels /ā/, /ī/, and /ū/.

Table 2: The area functions (cm^3) and vocal tract lengths (VT) for all four subjects and all three Arabic Vowels*.

Section	/ā/				/ī/				/ū/			
#	RA	RK	WA	ZA	RA	RK	WA	ZA	RA	RK	WA	ZA
1	1.35	2.614	1.575	0.9	1.575	2.359	3.375	1.8	1.875	3.442	3.825	0.225
2	2.1	5.036	3	2.325	2.475	4.718	4.275	2.475	1.5	5.61	5.1	2.4
3	2.925	4.845	4.5	3.3	3.45	5.1	5.475	3.825	2.25	5.61	4.875	2.85
4	4.125	3.825	4.35	4.425	5.025	4.845	5.475	4.35	2.775	5.228	5.7	3.975
5	4.35	4.016	3.225	5.1	5.85	3.698	4.725	4.575	4.575	4.144	5.325	2.85
6	4.575	3.187	2.1	4.875	5.55	4.144	4.35	3.9	4.875	4.59	4.725	4.875
7	3.75	3.634	2.325	4.35	4.8	4.335	5.175	3.9	5.325	5.1	4.125	5.1
8	2.85	2.933	1.875	3.225	4.05	3.506	5.025	3.225	4.275	3.825	3.45	4.125
9	1.575	2.386	2.325	2.475	3.15	3.399	4.875	2.782	4.875	3.698	3.225	4.125
10	1.2	2.148	2.421	1.575	3.825	4.206	5.325	4.95	5.7	3.397	2.25	4.875
11	1.125	2.24	2.524	1.575	3.075	4.294	5.96	5.129	4.875	3.154	2.325	3.6
12	1.05	2.18	2.523	2.191	2.925	4.478	5.481	5.446	4.65	2.974	2.55	3.225
13	2.4	2.295	2.617	2.223	4.575	3.984	5.382	6.707	4.5	2.992	2.567	3.531
14	1.95	2.895	2.652	2.64	4.35	4.351	5.208	5.658	4.65	2.939	2.545	3.106
15	2.175	3.366	1.694	2.615	4.575	4.794	3.88	5.757	5.55	2.693	2.4	3.1
16	2.1	3.5	1.075	2.195	3.825	5.294	3.079	3.594	4.95	3.045	2.404	3.19
17	1.8	3.892	1.154	2.319	4.125	5.315	2.161	3.484	4.95	3.308	2.185	3.451
18	2.4	3.648	0.72	2.347	4.2	4.745	1.981	3.201	4.425	3.322	1.595	3.508
19	2.48	3.707	0.411	2.125	4.275	4.474	2.159	3.191	3.675	2.928	1.058	2.887
20	1.239	3.974	0.584	1.337	4.356	4.252	1.959	2.874	3.314	2.899	1.108	2.489
21	1.329	3.813	0.78	1.545	4.486	3.707	1.505	2.601	3.194	2.682	1.131	1.439
22	1.369	3.771	1.12	1.419	2.483	3.193	1.254	2.291	3.262	2.358	0.84	1.439
23	1.424	3.443	1.119	1.265	2.692	2.924	0.993	1.902	2.099	2.203	0.901	1.404
24	1.657	3.275	1.412	1.308	3.217	2.429	1.121	1.688	1.593	1.859	1.293	1.298
25	1.543	3.382	1.669	1.374	2.815	2.191	0.88	1.499	1.467	1.633	0.995	1.119
26	1.712	3.446	1.758	1.33	2.644	1.969	0.88	1.311	1.511	1.557	1.596	1.122
27	2.056	3.601	2.039	1.449	2.64	1.828	0.984	0.88	1.734	1.406	1.778	0.906
28	2.311	3.863	2.284	1.688	2.673	1.782	1.055	0.793	1.953	1.547	2.239	0.94
29	2.304	4.019	2.176	2.039	2.316	1.699	1.055	0.716	1.695	1.926	2.255	0.774
30	2.73	4.035	2.421	2.18	1.724	1.632	0.94	1.039	1.612	2.579	2.461	0.984
31	2.679	4.086	2.574	2.52	1.578	1.438	1.029	1.089	1.656	3.088	2.63	0.984
32	2.687	4.107	2.847	2.758	1.085	1.331	1.357	1.2	1.282	3.952	2.657	1.372
33	2.781	4.248	3.186	3.045	0.831	1.344	1.42	1.282	1.799	4.803	2.977	1.376
34	3.006	3.629	3.471	3.453	0.683	1.483	1.416	1.502	2.002	4.95	3.156	1.792
35	3.187	3.478	3.584	4.076	0.805	1.381	1.325	1.639	2.135	5.209	3.984	2.179
36	3.305	3.599	3.342	4.656	0.774	1.784	1.887	1.93	2.25	8.01	4.3	2.489
37	3.19	3.424	3.478	4.502	0.984	1.938	1.439	2.578	2.461	5.976	5.396	2.932
38	3.281	3.853	3.142	4.233	1.055	1.918	1.771	1.682	2.397	4.162	5.332	3.559
39	3.537	3.544	2.184	4.529	1.274	1.905	3.05	1.441	2.352	2.984	4.863	3.849
40	3.583	3.107	1.971	4.082	1.571	2.25	2.691	1.634	2.6	2.461	4.679	6.066
41	3.339	2.611	1.867	4.322	1.729	1.89	2.399		2.533	1.811	2.963	4.592
42	2.962		2.625	2.509	1.412		2.306		2.493	1.287	1.86	3.186
43	3.138			1.616	1.254		3.545		5.06	1.229	1.189	2.734
44	3.941			1.301	1.201				7.97		1.333	2.024
45	4.147				1.548				6.853		1.322	1.176
46	4.061				2.777				4.235		3.022	1.346
47	3.032				3.726				2.401			
48	2.394				2.719				1.531			
49	2.101				2.692				1.514			
50	3.224				3.392				1.713			
51					3.903				1.685			
VT (cm)	15	12.3	12.6	13.2	15.3	12.3	12.9	12	15.3	12.9	13.8	13.8

* Side branches are excluded.

In this project, the computed area functions were transformed into formant frequencies using Maeda's vocal tract model VTCALCS program (Maeda 1982), which runs through MATLAB. We used the default values on the various speech parameters and the formants were generated with the default pitch (100Hz). Table 3 presents the predicted first three formants for the vowels. The predicted formants were then used to synthesize artificial vowel sounds using VTCALCS. The generated sounds were presented to a few listeners to identify and confirm that the quality of the sounds produced were intelligible.

Table 3: The predicted first three formant frequencies from MRI data.

	/ā/			/ī/			/ū/		
Subjects\Formants(Hz)	**F1**	**F2**	**F3**	**F1**	**F2**	**F3**	**F1**	**F2**	**F3**
RA	615	1453	2652	438	2002	2700	478	1496	2804
RK	695	1824	3109	533	2288	3291	559	1543	2826
WA	631	1676	3458	469	2347	3395	542	1389	3083
ZA	613	1587	2891	503	2548	3316	510	1645	2973

Many researchers chosen to compare synthetic formants with those produced from natural speech (such as Kröger *et al.* (2000) & Story *et al.* (1996)). The problem with this method is that the recordings are collected on a different day due to the noise encountered during MRI scanning. We choose to skip audio recording sessions because they are subject to intra-subject variabilities. A plain Electronic Medical Record (EMR) system was designed to form an Arabic speech database for the results from this study and future experiments.

CONCLUSIONS

Many speech production models exist and continue to be improved to provide an accurate relationship between the phonemes and different speech parameters such as cross-sectional areas, radiation effects, acoustic losses, and fluid viscosity. These models could improve greatly with a larger database of articulatory data covering a wider range of sounds and languages. Once the models are validated with a huge speech database, we would achieve a greater understanding of the speech production mechanism and provide an accurate empirical relationship between the phonemes and the different speech parameters. The main obstacle to building the articulatory configuration speech database is accessing the vocal tract.

This project utilized the increasingly popular MRI modality to access the vocal tract and study Arabic speech articulatory configurations, which have received very little attention in the past. The aim was to model Arabic vowels

and confirm the models by the production of synthetic sounds. The data show natural inter-subject variabilities, but are closely matched overall and in line with the theory of vowel production. The data from all subjects for all vowels shows an expansion of vocal tract cavity above the glottis. This can be attributed to the piriform sinuses joining with the vocal tract cavity. The data for the low-front vowel /$\overline{\text{o}}$/ shows a constricted pharynx for all subjects, except RK, and a widened oral cavity before a small mouth opening. The high-front vowel /$\bar{\text{I}}$/ is characterized by a wide pharynx and a restricted oral cavity leading into a larger mouth opening. The data for high-back vowel /$\overline{\text{u}}$/ shows a front (oral) and a back (pharynx) cavities separated by a tight constriction near the velum and a narrow mouth opening. Note that six sagittal images were used to show the vocal tract for females verses seven images for males. Generally, this did not seem to affect the results which were closely matched. The predicted formants also show closely matched results. However, it does appear that the synthesized formant frequencies from subject RA's results were almost always lower by varying degrees. However, most of the differences were small and probably insignificant, since they could be due to the fact subject RA was the tallest of all subjects.

This study is part of a long-term plan to provide and analyze articulatory data for Arabic speech, and future work should be targeted at further examination of Arabic vowels to confirm the three Arabic vowels' theory or identify more vowels. Also, the rest of the images collected in this study could be used to produce a three-dimensional articulatory model for the rest of the Arabic phonemes, and add results to our Arabic speech EMR database. Although static MRI limits speech studies to sustained sounds, it provides a powerful tool for accessing the vocal tract. However, recent techniques for dynamic MRI (Shadle *et al*. 1999, Takemoto & Honda 2003) could also be incorporated into future studies for dynamic models of Arabic sounds.

ACKNOWLEGMENTS

We would like to thank all the volunteers for the experiments of this project, and the MRI technicians at Mubarek Al-kabeer Hospital. We also thank Shinji Maeda for the VTCALCS program. This work has been supported by the Public Authority of Applied Education and Training (PAAET) of the State of Kuwait (grant TS-01-005).

REFERENCES

Al-Ani, S.H. 1970. Arabic Phonology: an Acoustical and Physiological Investigation. The Hague, Paris.

Baer, T. Gore, J.C. Gracco, L.C. & Nye, P.W. 1991. Analysis of vocal tract shape and dimensions

using magnetic resonance imaging. Journal of the Acoustical Society of America **90(2)**:799-828.

Fant, G. 1960. Acoustic Theory of Speech Production. Mouton, Graven-Hage.

Kröger, B.J. Winkler, R. Mooshammer, C. & Pompino-Marschall, B. 2000. Estimation of vocal tract area function from magnetic resonance imaging: preliminary results. Proceedings of the 5th Seminar on Speech Production, Kloster Seeon, 333-336.

Maeda, S. 1982. Digital simulation method of the vocal tract system. Speech Communication **1**:199-229.

Narayanan, S. 1996. Imaging applications in speech production research. International Society for Optical Engineering **2709**:120-131.

Shadle, C.H. Mohammad, M. Carter, J.N. & Jackson, P.J.B. 1999. Multi-planar dynamic magnetic resonance imaging: new tools for speech research. Proceedings of the 14th International Congress of Phonetics Sciences 623-626.

Story, B.H. Titze, I.R. & Hoffman, E.A. 1996. Vocal tract area functions from magnetic resonance imaging. Journal of the Acoustical Society of America **100(1)**:537-554.

Takemoto, H. & Honda, K. 2003. Measurement of temporal changes in vocal tract area function during continuous vowel sequence using a 3D cine-MRI technique. Proceedings of the 6th International Seminar on Speech Production, Sydney, 284-289.

Westall, F.A.R.D. & Johnston, L.A.V. 1996. Speech technology for telecommunications. British Telecom Technology Journal. **14(1)**:101-117.

Submitted : *25/8/2005*
Revised : *12/6/2006*
Accepted : *16/7/2006*

النماذج المفصلية للصوائت العربية باستخدام صور أشعة الرنين المغناطيسي

محمد عبد الجليل محمد[1] و طارق سنان[2] و باسمة عبد الجليل محمد[3]

[1]الهيئة العامة للتعليم التطبيقي والتدريب، كلية الدراسات التكنولوجية، قسم الهندسة الإلكترونية، ص.ب. 42325، الشويخ 70654، الكويت

[2]جامعة الكويت، كلية الطب، قسم الأشعة، الجابرية، الكويت

[3]الهيئة العامة للتعليم التطبيقي والتدريب، كلية العلوم الصحية، قسم السجلات الطبية، الشويخ، الكويت

خلاصة

يستعرض هذا البحث النتائج الأولية من مشروع طويل الأمد هدفه توفير قاعدة بيانات لنماذج النطق المفصلي للغة العربية. تم فى هذا البحث إلتقاط مواقع مفاصل النطق لإحدى وثلاثين صوتا عربيا بواسطة صور الرنين المغناطيسي و دراسة الصوائت العربية الثلاث الأساسية من أربعة متطوعين. السرعة البطيئة لإلتقاط صور أشعة الرنين المغناطيسي وضعت حدودا لهذة الدراسة و معظم الأبحاث المشابهه إلى دراسة الأصوات الثابته. الصور الملتقطه تم معالجتها و تجزيئها للتمكن من تطوير النماذج الحسابيه للمسالك الصوتيه. وقد تم إستخدام المعادلات المساحية الناتجة لحساب الذبذبات الأساسية للصوائت. تظهر النتائج فروقا طبيعية ومتوقعة بين المتطوعين، إلا إنها متقاربه بشكل عام وتتوافق مع نظرية نطق الصوائت. كما تم إستخدام النماذج الحسابية لتوليد المعادلة الصوتية وإصدار صوائت عربية مصطنعه. ومن ثم تم التأكد من أن نوعية الأصوات المصطنعة مفهومة.