ISCA Archive
http://www.isca-speech.org/archive

Auditory-Visual Speech Processing
2007 (AVSP2007)
Hilvarenbeek, The Netherlands
August 31 - September 3, 2007

Arabic Pharyngeals in Visual Speech

Slim Ouni[1], *Kais Ouni*[2]

[1] LORIA – BP 239 – 54506 Vandoeuvre-les-Nancy, France
[2] LSTS – ENIT, Tunisia

Slim.Ouni@loria.fr , Kais.Ouni@enit.rnu.tn

Abstract

Many perceptual experiments show that human talkers provide more intelligible visual speech than synthetic talkers. This inferiority of synthetic visual speech might be due to a lack of finer modeling of the parts of the face that are important to lipreading or that some parts of the face that are not generally considered as relevant to visual speech or as not visible in face-to-face communication, might actually provide some information, which humans are capable of decoding. This information might therefore not be modeled accurately in the synthetic speaker. In this paper, we provide evidence from Arabic that some sounds, which are not known as visible, might be recognized correctly visually. We performed a lipreading recognition experiment on Arabic, where a set of consonant-vowel stimuli were presented as visual-only speech and participants were asked to report what they recognized. The resulting consonant confusion matrix shows that some of these pharyngeals were, to some extent, well discriminated. Results are discussed based on the category of phonemes and the vowel context.

Index terms: visual speech, pharyngeal, Arabic language, viseme.

1. Introduction

In audiovisual speech, many studies showed that human talkers outperform synthetic talkers in perceptual experiments [1, 2, 3]. This is probably due to a lacking of finer modeling of the face or very likely because we still did not capture some important aspects of visual speech. For instance, some parts of the face that we do not consider relevant to visual speech or that they are not noticeably visible from outside, might actually provide some information, which human are capable of decoding. Studies as that presented in [4] might be helpful to better understand how information is communicated visually.

In this paper, we present a study of some aspects of visual speech specific to Arabic. In fact, some Arabic phonemes are produced at the pharynx and the articulation is followed by an important backing of the tongue. As these phonemes are mainly produced back in the vocal tract, we are interested to see if they have an effect on visible speech perception. Can human perceivers recognize these phonemes visually or they will be confused with their non-pharyngeal counterparts? Are pharyngeals visible from outside? Does their visibility depend on the vowel context?

Answering these questions will help to better understand the mechanism of visual speech for this category of sounds.

For this purpose, we performed a perceptual experiment on Arabic phonemes, and we report here the results for pharyngealization. In the following sections, we start by describing pharyngealization in Arabic, followed by a presentation of the experiment and we conclude by a discussion.

2. Pharyngealization in Arabic

An important articulatory feature of Arabic is the presence of pharyngealized and pharyngeal phonemes. There are two pharyngeal fricatives (/ħ/ and /ʕ/). These phonemes are characterized by the constriction formed between the tongue and the lower pharynx in addition to the rising of the larynx. There are two velars (/x/, /ʁ/) and one uvular (/q/) characterized by a constriction formed between the tongue and the upper pharynx for /x/ and /ʁ/ and a complete closure for /q/ at the same level. These five consonants are considered pharyngeal phonemes. In addition, there are four pharyngealized or emphatic phonemes: /sˤ/, /dˤ/, /tˤ/ and /ðˤ/. These phonemes are a pharyngealized version of the oral dental consonants /s/, /d/, /t/ and /ð/.

In Figure 1, we present tracings of X-Ray data of pharyngealized /tˤ/ and non-pharyngealized /t/ [6]. The phonetic place of articulation of both sounds is the same (both are dentals), however, they differ by the position of the back of the tongue, as /tˤ/ is far back than /t/. We may expect though that visually, they are indistinguishable. The same remark apply also to (/ð/, /ðˤ/), (/d/, /dˤ/) and (/s/, /sˤ/).

The main characteristic of the pharyngealization is the rearward movement of the back of the tongue. Thus, the vocal tract shape presents an increased oral cavity and a reduced pharyngeal cavity because of the retraction of the body and the root of the tongue toward the back wall of the pharynx [5, 6, 7]. The pharyngealized consonants also induce a considerable backing gesture in neighboring segments, which occurs primarily for the adjacent vowels (the pharyngealized consonants affect the neighboring vowels in such a way that they become pharyngealized. Thus, /æ/ becomes /ɑ/, /i/ becomes /ɨ/, /u/ becomes /ʉ/ (see Figure 2). The shape of the lips is different for /æ/ versus /ɑ/, which may help to better recognize the consonants in the context of these vowels. The vowels /ɨ/ and /ʉ/ are pharyngealized. Visually, we did not see any difference with their non-pharyngealized counterpart. We know, for instance, that for /ʉ/, the tongue moves toward the back wall of the pharynx (see for instance [10]), but visually the protrusion makes it practically impossible to see the tongue from outside. The tongue position for the phonemes /i/ and /ɨ/ are visually similar and characterized by stretching the lips, and thus a very little of the tongue is visible.

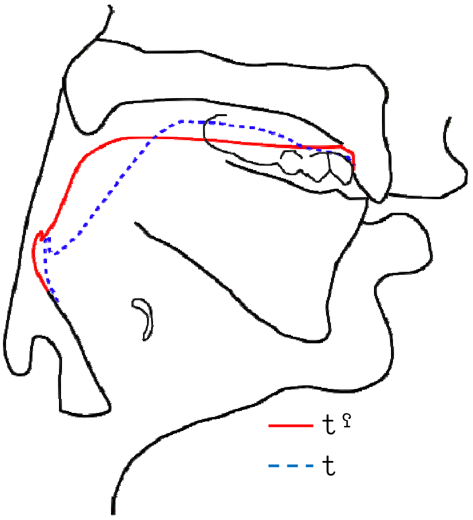

Figure 1 - ***Tracings of X-Ray data: pharyngealized /tˤ/ vs. non-pharyngealized /t/ (After [6]).***

The pharyngealization is observed in the pharyngealized consonants /sˤ/, /dˤ/, /tˤ/ and /ðˤ/ in almost all of the Arabic dialects. Nevertheless, in many regions of Tunisia, the phoneme /dˤ/ (sound of the letter ض) is pronounced as /ðˤ/. For the other pharyngeal phonemes, pharyngealization varies from one dialect to another. In many regions of Tunisia, the phonemes /ʕ/, /ʁ/, /r/, /ħ/, /q/ and /x/ are pharyngeals. In few other regions these same phonemes are non-pharyngeal (i.e., they are followed by non-pharyngeal vowels). The pharyngealization may also affect /l/ and /j/ in certain cases. As expected, researchers are not unanimous about the properties of these pharyngeal and pharyngealized phonemes in Arabic, its various dialects and the mechanism used for pharyngeal consonant production [5, 6, 8, 9].

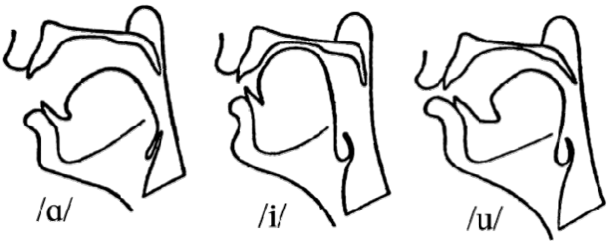

Figure 2 - ***Tracings of X-ray images of three vowels: /ɑ/, /ɨ/ and /ʉ/ in the context if the pharyngeal consonant /ħ/ (After [10]).***

3. Perceptual Experiment

A big part of the articulation of pharyngeal and pharyngealized phonemes takes place in the back of the vocal tract, which is not visible from outside. Our goal here is to verify whether some visible information is communicated to human perceivers: Are pharyngeal phonemes well recognized or mismatched with other phonemes? A lip-reading experiment on Arabic helps to find some answers. We report in the following results from the experiment described in [11], which is a more general study of visual speech in Arabic. We briefly describe the experiment and we focus here only on pharyngealization results.

3.1. Experiment Description

Ten native Arabic speakers, all Tunisian students, participated in this experiment. They were 23 to 29 years old in age, 4 females and 6 males. They all reported normal hearing and normal seeing abilities. They were living in the town of Tunis for several years.
The stimuli were 19 consonants : ***C*** = {/b/, /t/, /θ/, /ʃ/, /k/, /s/, /f/, /l/, /n/, /h/, /w/, /j/, /ħ/, /x/, /r/, /q/, /sˤ/, /tˤ/, /ðˤ/} and 3 vowels ***V*** = {/ɑ/, /ɨ/, /ʉ/} when the consonant is pharyngeal or pharyngealized and ***V*** = {/æ/, /i/, /u/} otherwise.
These sets of consonants and vowels form a total of 57 consonant-vowel syllables (CVs). These CVs were presented visual only without any audio. The consonant and vowel stimuli were chosen because they were representatives of distinct consonant viseme categories [11]. In addition, we kept all voiceless pharyngeal consonants and pharyngealized consonants, as they are the focus of this paper. These CVs were presented three times: each time, the set of 57 CVs was randomized. Therefore, the total number of trials was 171 presented in random order.

A Tunisian male talker was recorded uttering the 57 CVs. His presentations were video clips (size: 640 x 480, presented on 1600 x 1200, 21” monitor). We developed a presentation software which shows the video (without audio), and waits for a response from the participant. The participant chooses, among a panel showing the entire Arabic alphabet, the syllable that was pronounced. When processing these data, we took into account that some phonemes in the panel were not present in the stimuli. Thus, we considered a response to be correct, when the phoneme selected by a participant was visually indistinguishable from the one presented, based on our choice of visemes. For example, if a participant chooses the phoneme /z/, this will be converted to /s/, the glottal /ʔ/ is replaced by /h/, etc.

Table 1 - *Confusion matrix in the context of the vowel* /æ/ *(resp.* /ɑ/ *with pharyngeals). The main diagonal presents the proportion correct identification for each phoneme. For sake of clarity,* (.) *represents* 0*. Highlighted rows represent pharyngeals: dark grey cells show results for pharyngeal phonemes and light grey cells show results for pharyngealized phonemes.*

	ðˤ	θ	l	r	n	b	tˤ	ħ	h	q	t	f	s	sˤ	w	k	ʃ	j	x
ðˤ	**.95**	.	.	.	.	.	.05	.	.	.	.	.	.	.	.	.	.	.	.
θ	.	**.62**	.1	.	.	.	.	.	.	.	.2	.	.05	.	.	.	.	.05	.
l	.	.	**.89**	.	.11	.	.	.	.	.	.	.	.	.	.	.	.	.	.
r	.05	.	.	**.86**	.	.	.	.	.	.03	.	.	.	.	.	.	.	.	.1
n	.	.05	.29	.	**.19**	.14	.	.	.	.	.38	.	.	.	.	.	.	.1	.
b	.	.	.	.	.	**1.0**	.	.	.	.	.	.	.	.	.	.	.	.	.
tˤ	.1	.	.	.	.	.	**.52**	.	.	.	.05	.	.	.33	.	.	.	.	.
ħ	.	.	.	.	.	.	.	**.61**	.29	.	.	.	.	.05	.	.	.	.	.05
h	.	.	.	.	.	.	.	.06	**.89**	.	.06	.	.	.	.	.	.	.	.
q	.	.	.	.14	.	.	.05	.14	.1	**.38**	.	.	.	.	.	.	.	.	.19
t	.	.09	.09	.	.	.	.	.	.	.	**.76**	.	.	.	.	.	.	.05	.
f	.	.	.	.	.	.	.	.	.	.	.	**1.0**	.	.	.	.	.	.	.
s	.	.	.	.	.	.	.	.	.	.	.04	.	**.86**	.	.	.	.	.1	.
sˤ	.	.	.	.	.	.	.33	.	.	.	.05	.	.14	**.47**	.	.	.	.	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**1.0**	.	.	.	.
k	.	.	.05	.	.	.	.	.	.	.	.	.	.	.	.	**.28**	.	.66	.
ʃ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**1.0**	.	.
j	.	.	.05	.	.	.	.	.	.05	.	.05	.	.	.	.	.05	.	**.8**	.
x	.	.	.	.23	.	.	.	.14	.	.14	.	.	.	.	.	.	.	.	**.47**

3.2. Results

Tables 1, 2 and 3 show the results in three different vowel contexts: /ɑ/, /ɨ/ and /ʉ/ (resp. /æ/, /i/ and /u/ when the consonant is not pharyngeal). The results are presented by a confusion matrix based on the analysis of responses of all participants. We present the result for all the phonemes, but we highlighted pharyngeal and pharyngealized phonemes. In each context, the main diagonal presents the proportion correct for each phoneme.

The mean percentage of correct CVs recognized by participants was 45% in the three vowel contexts. Participants were able to identify the CVs with a reasonable accuracy in such conditions. We note that vowels were very accurately recognized (more than 95% correct). This result was expected as the three vowels were visually distinct and easily discriminated.

The pharyngealized phoneme /ðˤ/ was well recognized in the /ɑ/-context and highly mismatched with its non-pharyngealized counterpart /θ/ in the two other contexts. This implies that the presence of the vowel /ɑ/ or /æ/ provides perceivers with additional clues to identify the phoneme. In fact, based on the context, perceivers make a decision using the following kind of rules:

(a) (/ðˤ/ or /θ/) + /ɑ/ → /ðˤɑ/;

(b) (/ðˤ/ or /θ/) + /æ/ → /θæ/.

Table 2 - *Confusion matrix in the context of the vowel* /i/ *(resp.* /ɨ/ *with pharyngeals). The main diagonal presents the proportion correct identification for each phoneme. For sake of clarity,* (.) *represents* 0*. Highlighted rows represent pharyngeals: dark grey cells show results for pharyngeal phonemes and light grey cells show results for pharyngealized phonemes.*

	ðˤ	θ	l	r	n	b	tˤ	ħ	h	q	t	f	s	sˤ	w	k	ʃ	j	x
ðˤ	**.28**	.55	.	.	.	.	.05	.	.	.	.05	.	.05	.	.	.	.	.	.
θ	.06	**.62**	.19	.	.	.	.	.	.06	.	.06	.	.	.	.	.	.	.	.
l	.	.05	**.7**	.15	.05	.	.	.	.	.	.	.	.	.	.	.	.	.05	.
r	.	.	.1	**.6**	.1	.	.	.	.	.	.	.	.05	.	.	.	.	.	.15
n	.	.	.41	.06	**.06**	.	.	.	.	.	.35	.12	.	.	.	.	.	.	.
b	.	.	.	.	.	**1**	.	.	.	.	.	.	.	.	.	.	.	.	.
tˤ	.1	.	.	.	.	.	**.52**	.	.	.	.05	.	.	.33	.	.	.	.	.
ħ	.	.	.	.	.	.	.	**.62**	.28	.	.	.	.05	.	.	.	.	.	.05
h	.	.	.	.	.	.	.	.05	**.89**	.	.05	.	.	.	.	.	.	.	.
q	.	.	.	.14	.	.	.05	.14	.1	**.38**	.	.	.	.	.	.	.	.	.19
t	.	.1	.1	.	.	.	.	.	.	.	**.76**	.	.	.	.	.	.	.05	.
f	.	.	.	.	.	.	.	.	.	.	.	**1**	.	.	.	.	.	.	.
s	.	.	.	.	.	.	.	.	.	.	.05	.	**.85**	.	.	.	.	.09	.
sˤ	.	.	.	.	.	.	.33	.	.	.	.05	.	.14	**.47**	.	.	.	.	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**1**	.	.	.	.
k	.	.	.05	.	.	.	.	.	.	.	.	.	.	.	.	**.28**	.	.66	.
ʃ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	**1**	.	.
j	.	.	.05	.	.	.	.	.	.05	.	.05	.	.	.	.	.05	.	**.8**	.
x	.	.	.	.23	.	.	.	.14	.	.14	.	.	.	.	.	.	.	.	**.47**

We believe that this is very often used by perceivers to discriminate pharyngeal and pharyngealized phonemes from other phonemes that are visually equivalent, in this context. Nevertheless, in the /ʉ/-context the pharyngealized phoneme /ðˤ/ seems to have additional clues which make it sufficiently distinguishable from

/θ/. In the two contexts of /ɑ/ and /ɨ/, the pharyngealized phonemes /sˤ/ and /tˤ/ were correctly identified with a proportion of about 0.5, and mutually mismatched with a proportion of about 0.33 (as highest mismatch score). We also note that in these two contexts, (/sˤ/, /tˤ/) were not mismatched with their non-pharyngealized counterparts (/s/, /t/). This shows that there is an additional clue provided by the pharyngealization, which makes them sufficiently different from (/s/, /t/).

Table 3 - *Confusion matrix in the context of the vowel* /ɰ/ *(resp.* /ʉ/ *with pharyngeals). The main diagonal presents the proportion correct identification for each phoneme. For sake of clarity,* (.) *represents* 0*. Highlighted rows represent pharyngeals: dark grey cells show results for pharyngeal phonemes and light grey cells show results for pharyngealized phonemes.*

	ðˤ	θ	l	r	n	b	tˤ	ħ	h	q	t	f	s	sˤ	w	k	ʃ	j	x
ðˤ	**.67**	.33	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
θ	.6	**.4**	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
l	.	.05	**.**	.1	.1	.	.	.1	.42	.	.	.	.	.	.	.05	.	.1	.1
r	.05	.	.	**.15**	.	.	.	.05	.25	.	.05	.	.	.	.	.2	.	.05	.2
n	.	.19	.1	.	**.**	.	.	.05	.05	.	.38	.	.	.	.14	.	.	.1	.
b	.	.	.	.	.	**1.0**	.	.	.	.	.	.	.	.	.	.	.	.	.
tˤ	.5	.	.	.	.05	.	**.1**	.	.	.	.28	.	.1	.05	.05	.	.23	.1	.
ħ	.	.	.	.	.	.	.	**.05**	.76	.05	.	.	.	.	.	.	.	.	.14
h	.	.	.	.	.	.	.	.	**.71**	.	.	.	.	.	.	.	.	.	.14
q	.	.	.	.	.	.	.	.05	.66	**.1**	.	.	.	.	.14	.	.	.	.05
t	.	.05	.05	.	.14	.05	.05	.05	.23	.	**.19**	.	.	.	.	.	.	.	.
f	.05	.	.	.	.	.	.	.05	.	.	.	**.9**	.	.	.	.	.	.	.
s	.	.05	.	.	.	.	.	.	.	.	.19	.	**.1**	.	.05	.	.52	.1	.
sˤ	.	.	.	.	.	.	.	.	.05	.	.1	.	.05	**.14**	.	.	.52	.1	.05
w	.	.	.	.	.	.	.	.	.57	.05	.	.	.	.	**.38**	.	.	.	.
k	.	.	.05	.19	.	.	.	.1	.38	.14	.	.	.	.	.	**.05**	.	.	.1
ʃ	.	.	.	.	.	.	.	.	.05	.	.1	.	.05	.1	.05	.	**.52**	.1	.05
j	.05	.	.1	.	.05	.	.	.	.19	.	.05	.	.	.	.05	.1	.	**.38**	.05
x	.	.	.	.5	.	.	.	.05	.61	.1	.	.	.	.	.05	.	.05	.	**.1**

Pharyngeal phonemes (/ħ/, /x/, /r/ and /q/) presented better results in the two contexts of /ɑ/ and /ɨ/ than in the /ʉ/-context. The highest recognition score (86%) was that of /r/ in the /ɑ/-context. In the two contexts of /ɑ/ and /ɨ/, all the identification scores were higher than mismatch scores, even when the score is lower than 40%. This shows that these phonemes present some characteristics that help to make them perceptually distinguishable from the other phonemes. We believe that backing of the tongue helps to make this distinction, and thus, the recognition of these phonemes is very likely to be independent of the context (/ɑ/ and /ɨ/). The low recognition scores of pharyngeals in the /ʉ/-context is basically due to the importance of the protrusion which make it very difficult to obtain information from the tongue. None of the 4 pharyngeals presented above had higher identification score than a mismatch score. All of them were highly mismatched with /h/, which is probably a natural choice that participants tend to choose when they are in presence of similar phonemes (as for example, between a voiced and unvoiced phonemes, participants usually tend to choose unvoiced version of the phoneme). Note that /h/ is glottal which shows that participants perceived that they were back phonemes.

4. Discussion

The context of the vowel /ɑ/ helped in identifying pharyngeal and pharyngealized phonemes because the tongue is partially visible. The difference in the degree of the mouth opening between /ɑ/ and /æ/ helps to guide perceivers to distinguish between pharyngealized and their non-pharyngealized equivalent; and between pharyngeal and their closest non-pharyngeal phonemes, from an articulatory point of view.

The results in the context of the vowel /ɨ/, where the visibility of the tongue is very limited, inform that perceivers still recognize pharyngeal and pharyngealized phonemes, and more importantly they are not mismatched with the closest phonemes based on the place of articulation. It seems that pharyngealization provides extra clues due to the backing of the tongue. This is observable probably as (a) a little movement of the cheeks or (b) an acceleration during the release of a given consonant due to the imposing movement of the tongue, or (c) lowering the larynx which can be perceived through the skin of the neck.

The outcome of this experiment is that it is probably not sufficient to focus on modeling only the lips to build intelligible synthetic talker. The inferiority of synthetic visual speech is very likely due to a lack of finer modeling of the parts of the face that are important to lipreading. In addition, it could also be the case that some parts of the face that are not generally considered as relevant to visual speech or as not visible in face-to-face communication (as cheeks and neck), might actually provide some information, which humans are capable of decoding. The result of the experiment presented in this paper showed that for pharyngeals and pharyngealized phonemes, which do not seem to be produced by articulators that are easily visible in face-to-face communication, some visual information correlates with the production of pharyngeal phonemes that can be exploited by humans to recognize the pharyngeal place of articulation. This information might therefore not be modeled accurately in the synthetic speaker.

5. Conclusion

Does Pharyngealization in Arabic help to provide visible information? The answer to this question is very likely yes. Perceivers can get much information from the face, even though when the articulation is based on the back of the vocal tract. It is truer for pharyngealized phonemes than pharyngeals, and in some contexts than others. As pharyngealized and pharyngeal phonemes were to some extent recognized even in difficult vowel contexts, this shows that these phonemes provide additional information to be visually recognized. The described perceptual experiment in this paper provides evidences in this direction. However, we need probably more specific experiments to measure this information. We will address these issues in our future work. In fact, studying pharyngealization dynamics by tracking some markers on the face might provide valuable information. In addition, we will design an experiment using selective visual masking, similar to the one described in [4] and we will compare results of natural talker vs. synthetic talker to assess this information. This kind of experiment may provide a finer explanation of this phenomenon.

6. REFERENCES

[1] LeGoff, B., Guiard-Marigny, T., Cohen, M.M., Benoît, C. 1994. Real-Time Analysis-Synthesis and and Intelligibility of Talking Faces, *2nd Conf. On Speech Synthesis*, Newark.

[2] Massaro, D. W. 1998. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press: Cambridge, MA.

[3] Ouni, S., Cohen, M.M., Ishak, H., and Massaro, D.W., 2007. Visual Contribution to Speech Perception: Measuring the Intelligibility of Animated Talking Heads. *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007.

[4] Preminger, J. E., Lin, H.-B., Payen, M., & Levitt, H. 1998. Selective masking in speechreading. *Journal of Speech, Language, and Hearing Research*, vol. 41, 564–575.

[5] Al-Ani, S. 1970. *Arabic Phonology, an acoustical and physiological investigation.*, The Hague: Mouton.

[6] Ghazali, S. 1977. *Back Consonants and Backing Coarticulation in Arabic*. University of Texas at Austin, Austin.

[7] Jakobson, R. 1962. "Mofaxxama", the emphatic phonemes in Arabic. *Selected Writings* vol. 1, pp. 510-522. The Hague: Mouton.

[8] Ali, L., & Daniloff, R. 1972. A contrastive cinefluorographic investigation of the articulation of emphatic/non-emphatic cognate consonants. *Studia Linguistica*. vol. 26, n. 2, 81-105.

[9] Elgendy, A. M. 2001. *Aspects of Pharyngeal Coarticulation*. University of Amsterdam, Netherlands.

[10] Delattre, P. 1971. Pharyngeal features in the consonants of Arabic, German, Spanish, French and American English. *Phonetica.* Vol. 23, 129–155

[11] Ouni S., Ouni K. 2007. Aspects of Visual Speech in Arabic. *Interspeech 2007.* Antwerp, Belgium.