



# Aspects of Visual Speech in Arabic

*Slim Ouni<sup>1</sup>, Kais Ouni<sup>2</sup>*

<sup>1</sup> LORIA – BP 239 – 54506 Vandoeuvre-les-Nancy, France

<sup>2</sup> LSTS – ENIT, Tunisia

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

## Abstract

In this paper, we present a study of visual speech in Arabic. More specifically, 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 overall lipreading scores were consistent with other experiments in other languages. The resulting consonant confusion matrix shows that some of the phonemes were well discriminated, however, for others it depends on the context. Results are discussed based on the category of phonemes and the vowel context.

**Index Terms:** lipreading, visual speech, Arabic, visemes.

## 1. Introduction

Several researches showed that the face provides relevant visual information which helps to improve speech intelligibility, particularly when auditory signal is degraded or absent [1, 2, 3, 4]. Studying visible speech intelligibility is beneficial as it helps to better understand how this communication is performed and what elements of the language are more visible.

In this paper, we present a study on visible speech for Arabic. Such a study is lacking and there is very little research in visible speech for this language. We are expecting to obtain similar visible speech recognition results as what already was obtained for other languages. In fact, Arabic is no exception as for gaining intelligibility in presence of visual information (see [5], for instance). However, it is reassuring to show the benefit of the visual information for Arabic for lipreading and to be able to compare results with other languages. Furthermore, the results may provide helpful insights to improve quality of synthesis in talking head systems (as the one presented in [5]). In addition, a finer study at phonetic level will help to better understand which phonemes are more visible. The vowel context may also affect intelligibility of these phonemes, which was the case for other languages [2, 6].

In this paper, we present a lipreading experiment in Arabic. We presented several initial consonants in consonant-vowel contexts. In the following section, we

present the Arabic phonemes and visemes used in our study. Then, we present the experiment and the results.

## 2. Arabic Phonemes and Place of Articulation Visibility

Arabic language has 34 phonemes: 28 consonants and 6 vowels. Three vowels are short (/æ/, /i/ and /u/) and three are long (/ææ/, /ii/ and /uu/). These vowels are usually transformed in their pharyngealized version (/ɑ/, /ɨ/, /ɤ/) in presence of pharyngeal or pharyngealized phonemes.

The Arabic phonemes are presented in Table 1, and organized based on their place of articulation. We have 5 regions: bilabial, dental, postalveolar, palatal, and back consonants. Inter-dental, dental and alveolar are grouped together and are considered as dental. Back consonants (velar, uvular, pharyngeal and glottal) are usually considered as pharyngeal and they are characterized by the rearward movement of the back of the tongue: 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 [7, 8, 9]. The dentals /t/, /d/, /s/ and /ð/ have pharyngealized counterpart phonemes which are /t<sup>ɣ</sup>/, /d<sup>ɣ</sup>/, /s<sup>ɣ</sup>/ and /ð<sup>ɣ</sup>/. These phonemes are also dentals but they are followed by an important backing of the tongue, which is different compared to pharyngeal phonemes as they are actually produced at the pharynx. As we can see in Table 1, almost half of the Arabic consonants are fricatives.

Bilabial and dental regions are visible or partially visible and it is possible to lipread the phonemes whose places of articulation are in these regions. The visibility of some of these phonemes may depend on the vowel context. The lips are more open for some vowels and thus it is possible to see inside the vocal tract. In Arabic, there is only one vowel that is open (/æ/ and resp. /ɑ/ for pharyngeals). The back consonants are not visible, or very hard to see. It seems naturally difficult to lipread these phonemes as the place of articulation is completely or partially invisible from the outside unless there are other clues that help humans to perceive these sounds visually.

Table 1 – Places of articulation for Arabic consonants, organized based on their positions in the vocal tract from the glottis to the lips. Inter-dental, dental and alveolar are grouped together and considered as dental. Velar, uvular and pharyngeal are usually considered pharyngeal. Grayed cells are the places where it is hard to see the articulation from the face. This table stresses also the fricatives. The phonemes /tʰ/, /dʰ/, /sʰ/ and /θʰ/ are not presented in this table. They are dental in addition to a backing of the tongue.

fricatives

	w , b , m	Bilabial	Bilabial
f		Labiodental	
θ , ð		Interdental	Dental
s, z	t, d, n, r, l	Dental	
ʃ, ʒ			Postalveolar
	j		Palatal
x, ɣ	k	Velar	Back Consonants
	q	Uvular	
ħ, ʕ		Pharyngeal	
h	ʔ	Glottal	

### Arabic visemes

To perform our perceptual experiment, we defined a viseme set for Arabic. We grouped consonants according to their similarities visually and according to their place of articulation. In English, early viseme classification was based on the place of articulation [10]. This was modified later on by taking into account ideal and usual view conditions [11, 12]. In [13], a qualitative study of the difficulties of perceivers to lipread under conditions of reduced phonetic distinctiveness was presented. In our study we used some of these results, but we kept several phonemes separate as the articulation may not exactly be the same for Arabic as for English. For example, we kept all the pharyngealized phonemes, and the voiceless pharyngeals. We removed, however, the phoneme /dʰ/ (sound of the letter ض) as it is pronounced by most of Tunisians as /ðʰ/. In fact, in our experiment, the talker and participants were Tunisians, and thus the stimuli should be adapted to these two parties.

The phoneme /b/ represents the pair {b, m}. For non-pharyngeals, we kept the following voiceless consonants that we consider as sufficiently distinct visually: /t/, /θ/, /ʃ/, /k/, /s/, /f/, /l/, /n/, /h/, /w/ and /j/.

As a result, the set of the visemes considered is the following:  $C = \{ /b/, /t/, /θ/, /ʃ/, /k/, /s/, /f/, /l/, /n/, /h/, /w/, /j/, /ħ/, /x/, /r/, /q/, /sʰ/, /tʰ/, /ðʰ/ \}$ .

This was our choice of visemes in this preliminary study of visual speech in Arabic. In the future, we plan to perform a classification specific to Arabic phonemes, to group phonemes in perceptually visually equivalent groups.

## 3. A Lipreading Experiment

We carried out a lip-reading experiment on a selected set of consonants in three vowel contexts. The presentation was unimodal visual natural speaker. As no audio was presented, the set of stimuli was selected among the consonants that were sufficiently different visually, as described above.

### 3.1. Participants

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.

### 3.2. Test stimuli

The stimuli were 19 consonants:  $C = \{ /b/, /t/, /θ/, /ʃ/, /k/, /s/, /f/, /l/, /n/, /h/, /w/, /j/, /ħ/, /x/, /r/, /q/, /sʰ/, /tʰ/, /ðʰ/ \}$  and 3 vowels  $V = \{ /a/, /i/, /u/ \}$  when the consonant is pharyngeal or pharyngealized and  $V = \{ /æ/, /ɪ/, /ʊ/ \}$  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 task was lipreading. 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.

### 3.3. Results

The mean percentage of correct CVs recognized by participants was 45% in the three vowel contexts. The

percentages were in the same range as in other similar experiments for English [14, 16] and for French (see [15], if we consider the highest noise level, for instance). 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. We performed also a finer analysis at a phoneme level that is presented in Table 2, 3 and 4. Results were presented in three different vowel contexts by a confusion matrix based on the analysis of responses of all participants.

Table 2 - 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.

	ð <sup>ɛ</sup>	θ	l	r	n	b	t <sup>ɛ</sup>	h	h	q	t	f	s	s <sup>ɛ</sup>	w	k	ʃ	j	x
ð <sup>ɛ</sup>	.95	.	.	.	.	.05	.	.	.	.	.	.	.	.	.	.	.	.	.
θ	.62	.1	.	.	.	.	.	.	.2	.05	.	.	.	.	.	.	.05	.	.
l	.	.89	.11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
r	.05	.	.86	.	.	.	.	.03	.	.	.	.	.	.	.	.	.	.	.1
n	.05	.29	.19	.14	.	.	.	.38	.	.	.	.	.	.	.	.	.	.	.1
b	.	.	.	.	.10	.	.	.	.	.	.	.	.	.	.	.	.	.	.
t <sup>ɛ</sup>	.1	.	.	.	.	.52	.	.	.05	.	.33	.	.	.	.	.	.	.	.
h	.	.	.	.	.	.61	.29	.	.	.05	.	.	.	.	.	.	.	.05	.
h	.	.	.	.	.	.06	.89	.06	.	.	.	.	.	.	.	.	.	.	.
q	.	.	.14	.	.05	.14	.1	.38	.	.	.	.	.	.	.	.	.	.	.19
t	.09	.09	.	.	.	.	.	.76	.	.	.	.	.	.	.	.	.	.05	.
f	.	.	.	.	.	.	.	.10	.	.	.	.	.	.	.	.	.	.	.
s	.	.	.	.	.	.	.	.04	.86	.	.	.	.	.	.	.	.	.1	.
s <sup>ɛ</sup>	.	.	.	.	.	.33	.	.05	.14	.47	.	.	.	.	.	.	.	.	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.10	.	.	.	.	.
k	.	.05	.	.	.	.	.	.	.	.	.	.	.	.	.28	.66	.	.	.
ʃ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.10	.	.	.
j	.	.05	.	.	.	.05	.05	.	.	.	.	.	.	.	.05	.8	.	.	.
x	.	.	.23	.	.	.14	.14	.	.	.	.	.	.	.	.	.	.	.	.47

Across the three contexts, /æ/-context (and /ɑ/-context for pharyngeals) presented relatively higher recognition results than /i/ and /u/-contexts. This may be explained by the fact that the vowel /æ/ (and /ɑ/ for pharyngeals) are open and thus, we very likely have some additional information from seeing the tongue. Some phonemes presented remarkable identification results. Bilabial presented the best results across the different contexts. /b/ and /f/ (across the different contexts) and /w/ (in the /æ/-context and /i/-context) were correctly identified. The post-alveolar /ʃ/ was correctly identified in the /æ/-context and /i/-context. The articulation of these phonemes is mainly based on lips, which are totally visible. This explains to some

extent the above results. The pharyngealized phoneme /ð<sup>ɛ</sup>/ was well recognized in the /ɑ/-context and highly mismatched with /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) (/ð<sup>ɛ</sup>/, /θ/) + /ɑ/ → /ð<sup>ɛ</sup>ɑ/; (b) (/ð<sup>ɛ</sup>/, /θ/) + /æ/ → /θæ/. We believe that this is very often used by perceivers to identify pharyngeal and pharyngealized phonemes from other phonemes visually equivalent, in this context. Nevertheless, in the /u/-context the pharyngealized phoneme /ð<sup>ɛ</sup>/ seems to have additional clues which help

Table 3 - 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.

	ð <sup>ɛ</sup>	θ	l	r	n	b	t <sup>ɛ</sup>	h	h	q	t	f	s	s <sup>ɛ</sup>	w	k	ʃ	j	x
ð <sup>ɛ</sup>	.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	.	.	.	.	.	.	.	.	.	.	.	.	.	.
t <sup>ɛ</sup>	.1	.	.	.	.	.52	.	.	.05	.	.33	.	.	.	.	.	.	.	.
h	.	.	.	.	.	.62	.28	.	.	.05	.	.	.	.	.	.	.	.05	.
h	.	.	.	.	.	.05	.89	.05	.	.	.	.	.	.	.	.	.	.	.
q	.	.	.14	.	.05	.14	.1	.38	.	.	.	.	.	.	.	.	.	.	.19
t	.1	.1	.	.	.	.	.	.76	.	.	.	.	.	.	.	.	.	.05	.
f	.	.	.	.	.	.	.	.10	.	.	.	.	.	.	.	.	.	.	.
s	.	.	.	.	.	.	.	.05	.85	.	.	.	.	.	.	.	.	.09	.
s <sup>ɛ</sup>	.	.	.	.	.	.33	.	.05	.14	.47	.	.	.	.	.	.	.	.	.
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.10	.	.	.	.	.
k	.	.05	.	.	.	.	.	.	.	.	.	.	.	.	.28	.66	.	.	.
ʃ	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.10	.	.	.
j	.	.05	.	.	.	.05	.05	.	.	.05	.05	.	.	.	.05	.8	.	.	.
x	.	.	.23	.	.	.14	.14	.	.	.	.	.	.	.	.	.	.	.	.47

perceivers to less mismatch it with /θ/. This was not the case for the other pharyngealized phonemes (/s<sup>ɛ</sup>/, /t<sup>ɛ</sup>/), as the context helped a little but the pharyngealization does not seem to help for these two phonemes and they were very often mismatched with their non-pharyngealized counterparts. One interesting result was that of /l/ and /r/ in the two contexts of vowels /æ/ and /i/. The place of articulation of these two phonemes is in the same region, but they seem to have differences that make them sufficiently distinct to perceivers. The phoneme /n/ was confused with /l/ and /t/ in the three vowel-contexts. These three phonemes are dentals, which may explain this result. However, /l/ was not mismatched with /n/ nor with /t/ (89% correctly

identified and 11% mismatched with /n/). In English, these three phonemes are considered perceptually equivalent [13].

#### 4. Conclusion

The lipreading experiment presented in this paper provided some insights about some visual aspects of Arabic language. As in many other languages, visual speech in Arabic improved intelligibility compared to the case where there is no audio and no video (no information at all), as visual only CVs were reasonably identified in this difficult condition (lipreading) where there is no acoustic information. The vowel context has to some extent an influence on recognition. However, some other phonemes were easily identified independently of the context.

We should also note that these results are based on one speaker, and probably a comparison with other speakers should provide more accurate analysis, taking into account speaker variability that may exist. In our future work, evaluating the intelligibility of the speaker, in addition to comparing several speakers will be performed. Besides, we will consider other stimuli to be able to provide some results on the perception of coarticulation in Arabic.

As a final remark, we note that these results are specific to Arabic spoken in Tunisia and may be different from region to region in the Arab world. In fact, dialects may have influence on production and thus on perception of certain sounds (pharyngeals for instance).

Table 4 - Confusion matrix in the context of the vowel /a/ (resp. /a/ with pharyngeals). The main diagonal presents the proportion correct identification for each phoneme. For sake of clarity, (.) represents 0.

	ð	θ	l	r	n	b	t <sup>ʕ</sup>	ħ	h	q	t	f	s	s <sup>ʕ</sup>	w	k	ʃ	j	x
ð	.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	.	.	.	.	.	<b>1.0</b>	.	.	.	.	.	.	.	.	.	.	.	.	.
t <sup>ʕ</sup>	.5	.	.	.	.05	.	<b>.1</b>	.	.	.	.28	.	.1	.05	.05	.	.23	.1	.
ħ	.	.	.	.	.	.	.	<b>.05</b>	.76	.05	.	.	.	.	.	.	.	.	.14
h	.	.	.	.	.	.	.	.	<b>.71</b>	.	.	.	.	.	.	.	.	.	.14
q	.	.	.	.	.	.	.	.05	.66	<b>.1</b>	.	.	.	.	.14	.	.	.	.05
t	.	.05	.05	.	.14	.05	.05	.05	.23	.	<b>.19</b>	.	.	.	.	.	.	.	.
f	.05	.	.	.	.	.	.	.05	.	.	.	.	<b>.9</b>	.	.	.	.	.	.
s	.	.05	.	.	.	.	.	.	.	.	.19	.	<b>.1</b>	.05	.05	.	.52	.1	.
s <sup>ʕ</sup>	.	.	.	.	.	.	.	.05	.	.1	.	.05	<b>.14</b>	.	.	.52	.1	.05	.
w	.	.	.	.	.	.	.	.57	.05	.	.	.	.	.	<b>.38</b>	.	.	.	.
k	.	.	.05	.19	.	.	.	.1	.38	.14	.	.	.	.	.	<b>.05</b>	.	.	.1
ʃ	.	.	.	.	.	.	.	.05	.	.1	.	.05	.1	.05	.	<b>.52</b>	.1	.05	.
j	.05	.	.1	.	.05	.	.	.19	.05	.	.	.	.	.05	.1	.	<b>.38</b>	.05	.
x	.	.	.	.5	.	.	.	.05	.61	.1	.	.	.	.05	.	.05	.	.	<b>.1</b>

#### 5. References

- [1] Sumbly, W. H., & Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.*, 26, 212-215.
- [2] Benoit, C. Mohamadi, T., Kandel, S. 1994. Effects of Phonetic Context on Audio-Visual Intelligibility of French. *J Speech Hear Res.*37, 1195-1203
- [3] Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica*, 36, 314-331.
- [4] Massaro, D. W. 1998. *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. MIT Press: Cambridge, MA.
- [5] Ouni S., Cohen M.M., Massaro D.W. 2005. Training Baldi to be multilingual: a case study for an Arabic Badr. *Speech Communication*. vol. 45, no. 2, pp. 115–137.
- [6] Jiang, J., Alwan, A., Bernstein, L., Auer, E., Keating, P. 2002. Similarity structure in perceptual and physical measures for visual consonants across talkers. *ICASSP*. Orlando, pp. 441-444
- [7] Al-Ani, S. 1970. *Arabic Phonology, an acoustical and physiological investigation.*, The Hague: Mouton.
- [8] Ghazali, S. 1977. *Back Consonants and Backing Coarticulation in Arabic*. University of Texas at Austin, Austin.
- [9] Jakobson, R. 1962. "Mofaxxama", the emphatic phonemes in Arabic. *Selected Writings* vol. 1, pp. 510-522. The Hague: Mouton.
- [10] Nitchie, E.H. 1950. *New Lessons in Lipreading*. J.B. Lippincott Company, New York.
- [11] Fisher, C.G. 1968. Confusions among visually perceived consonants. *J. Speech Hearing Res.* vol. 11, pp. 796–803.
- [12] Jeffer, J., Barley, M. 1971. *Speechreading (Lipreading)*. Thomas, Springfield, Illinois.
- [13] Auer, E. T., Jr., Bernstein, L. E. 1997. Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *J. Acoust. Soc. Am.*. 102, 3704-3710.
- [14] 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.
- [15] LeGoff, B., Guiard-Marigny, T., Cohen, M.M., Benoit, C. 1994. Real-Time Analysis-Synthesis and Intelligibility of Talking Faces, *2nd Conf. On Speech Synthesis*, Newark.
- [16] Bernstein, L.E., Demorest, M.E., and Tucker, P.E., 2000. Speech Perception without hearing. *Perception & Psychophysics*, vol. 62, no. 2, pp. 233-252.