ACTA OTORHINOLARYNGOLOGICA ITALICA 2009;29:203-208

OTOLOGY

The McGurk phenomenon in Italian listeners

L'effetto McGurk nell'ascoltatore italiano

R. BOVO, A. CIORBA, S. PROSSER, A. MARTINI
Audiology and Phoniatrics Department, University Hospital of Ferrara, Ferrara, Italy

SUMMARY

In the classic example of the McGurk effect, when subjects see a speaker say /ga/ and hear a simultaneous /ba/, they typically perceive /da/, a syllable that was not presented either acoustically, or visually. This phenomenon, although non-natural and recreated in laboratory investigations, has been studied in order to better understand how, where and when the central nervous system processes and integrates visual and auditory signals. Till now, it has been demonstrated for English, Spanish and German languages, while in Japanese and Chinese it seems weaker. Aim of this study was to evaluate the entity of the McGurk effect for the Italian language. Results obtained demonstrate a robust McGurk effect for the Italian language, which has never been described before. The phenomenon is highly significant when an auditory bilabial Consonant-Vowel is dubbed with a visual apico-dental or velar Consonant-Vowel. Results are discussed on the basis of the recent hypothesis regarding the bimodal perception.

KEY WORDS: Bimodal perception • McGurk effect • Audio-visual integration • Speech-reading

RIASSUNTO

Quando un soggetto vede la faccia di un parlante che articola una /ga/ e contemporaneamente sente la sillaba /ba/, la percezione più frequente è quella di una /da/, ovvero una sillaba che non è stata presentata né acusticamente, né visivamente. Questo è un classico esempio del fenomeno McGurk, che è stato a lungo studiato per capire meglio come, dove e quando il sistema nervoso centrale integra l'informazione uditiva con quella visiva. Sino ad oggi il fenomeno è stato dimostrato per l'inglese, lo spagnolo, ed il tedesco, mentre in giapponese e in cinese l'effetto è più debole. Scopo di questo lavoro era quello di valutare l'entità dell'effetto McGurk nella lingua italiana. I nostri risultati dimostrano un evidente effetto McGurk anche per la lingua italiana, particolarmente quando una consonante vocale bilabiale uditiva è doppiata con una consonante vocale apico-dentale o velare visiva. I risultati sono discussi sulla base di alcune recenti ipotesi sull'integrazione bimodale e sulle differenze inter-linguistiche.

PAROLE CHIAVE: *Percezione bimodale • Effetto McGurk • Integrazione audio-visiva • Lettura labiale*

Acta Otorhinolaryngol Ital 2009;29:203-208

Introduction

Even if language perception is usually considered as an auditory process, it has already been demonstrated that visual information has a strikingly potent effect on perception, also in normal hearing subjects [1]. When presenting a movie in which acoustic signal has been degraded, typical results are as follows: 1% of word recognition with audio off; 6% of word recognition with video off; 45% of word recognition with video and audio on.

The effect of visual information is generally studied either with speech stimuli degraded by noise, by different speech manipulation (filtering, interruption, compression, etc.), or by testing hearing-impaired patients. Nevertheless, there is a phenomenon which permits evaluation of the effect of bimodal integration even in normal-hearing subjects and with clear acoustic signal. This situation is represented by the McGurk phenomenon: in the classic example, when subjects see a speaker say /ga/ and are simultaneously presented with an acoustic /ba/, they typically perceive /da/, a syllable that was not presented either acoustically, or visually (fusion response). A perceptive response of combination /bga/ can be obtained when a visual (V) /ba/ is administrated with a auditory (A) /ga/ (combination response) [2].

The McGurk phenomenon, even if representing a non-natural situation, has been studied in order to better understand how, where and when the central nervous system (CNS) processes and integrates visual and auditory signals. Furthermore, various factors can influence this integration: age, stimulus type and native language, discrepancies between stimuli, speech contest, hemispheric advantage [3-7].

Various Authors have studied the McGurk phenomenon among young children [3 4 7-9]. Burnham & Dodd [10] demonstrated the presence of fusion responses until the first few months, in prelinguistic age. Nevertheless, children aged of 3 to 5 years, and 7 to 8 years have less visual influence

in bimodal perception, i.e., a weaker McGurk effect, than adults. The developmental increase in visual influence over age is possibly related to experience in articulatory speech sounds. In fact, preschool children who make substitution errors in articulation are less influenced by visual cues than those who can correctly produce consonants [11]. Thus there is evidence that articulation experience is related to the degree of visual influence in auditory-visual speech perception.

To evaluate the McGurk phenomenon in different phonological conditions, various studies have been performed on listeners of several native-languages: English, Japanese, German, Spanish, Thai, Chinese Mandarin and Hungarian [5 6 12-16].

Munhall et al. [17] evaluated the effect of temporal asynchronies between A and V stimuli, observing that the McGurk phenomenon maintains its robustness up to asynchronies of 250 msec.

Several Authors have studied the effect of different discrepancies between the A and V modalities: Green et al. [18] analysed the effects of a female voice on a male face and vice versa.

Moreover, several Authors have described a stronger McGurk effect when it is correlated to vocals /i/ and /a/, while it resulted weaker with vocal /u/. This may be due to the type of lip movements when articulating /u/, that could reduce the quality of the visual information given [12 19].

Diesch [20] studied the modality of the CNS processing in the two cerebral hemispheres, demonstrating an advantage of the left side when elaborating fusion responses.

Aim of this study was to evaluate whether a McGurk effect can be elicited also for the Italian language. In fact, as far as we know, no data have been reported in the literature in this regard.

Material and Methods

Stimuli

Visual stimuli were edited from a digital video recording of a female speaker articulating the following eight Consonant-Vowel (CV) syllables: /ba/, /da/, /ga/, /pa/, /ta/, /ka/, /ma/, /na/. All video movie files with the articulations were edited in order to begin and end at the same neutral resting position. Each visual CV stimulus was dubbed with one of the 8 acoustic stimuli of the same CV uttered by an Italian female speaker and recorded through a Digital Audio Tape (DAT) in order to obtain all the 64 possible combinations. To permit a better alignment of the discrepant Audio-Visual (AV) stimuli, the voice bar of the 3 stop consonants /b,d,g/ was reduced from approximately 100 ms to 30 ms. This procedure, which seems to be necessary because of the long voice onset time (VOT) of Italian plosives, has been proved not to modify their identification[21]. With the use of dedicated software it was possible to accurately align the visual consonant release with the acoustic plosion. Eight lists of eight AV stimuli have been randomized for each testing session.

The audio signals were presented, at a comfortable hearing level, through two loud-speakers positioned at 1.5 m from the subject at ± 45° with respect to the azimuth, in a 3x3 m soundproof booth, while the movie files were presented on a 20" monitor, positioned at 1 m from the subject.

Subjects

The study population comprised 4 male and 6 female subjects of Italian language, age range 18-27 years (mean 23.2 yrs). All subjects had normal visual skills and normal hearing.

Test Procedures

Subjects were asked to watch and listen to each AV stimulus, and then to write down (in an open set paradigm) what they perceived. The stimuli included the so-called "combination presentation" (A/da/ vs. V/ba/; A/ga/ vs. V/ba/), which sometimes produce "combination responses" (e.g. /bda/ or /bga/). Four test blocks have been proposed, on two different days, for a total of 256 AV stimuli administered to each subject (i.e., 64 AV CV couples of stimuli repeated for the 4 sessions).

Results

Responses that were not identical to the A stimulus were regarded as "visually influenced". All responses were plotted in a confusion matrix for each subject examined. Overall results, in percentage from all subjects, are outlined in Table I. The size of visual influence was used in the analyses: this is an established method of reporting results [22]. Along the off-diagonal cells the percentage of correct responses are reported for the match AV syllables (i.e., A/ba/ vs V/ba/). Mean error for these congruent stimuli was 1.5% (range 0-4%). These data were used as the basis for all further analyses. Table I is divided into four quadrants, corresponding to a binary division, on each modality, into labial and non-labial consonants. The binomial distribution was used in the analysis of off-diagonal cells with 98.5% and 1.5% as the values of the parameters for p and q. A percentage of visually influenced responses of 80% or less is significantly different from chance ($p < 0.05$).

In the upper left quadrant, i.e., when V/da,ta,ga,ka,na/ are presented with the same A syllables, the visually influenced responses are very few, with the possible exception for the apico-dental /na/. Few visually influenced responses occur also in the lower right quadrant: when auditory /ba, pa, ma/ are dubbed, with a different bilabial CV, the auditory perception is not influenced. On the other hand, when the bilabial CV are dubbed with visual non-labial CV, the percentage of visually influenced re-

Table I. Type (and mean percentage) of responses for audio-visual CV stimuli, for all subjects.

	Visual signal							
	da	**ta**	**ga**	**ka**	**na**	**ba**	**pa**	**ma**
da	da 100	da 100	da 100	da 100	da 100	da 58	da 82	da 83
						ba 17	bda 18	bda 17
						bda 17		
						pda 8		
ta	ta 100		ta 100	ta 100	ta 58	ta 58	ta 100	
		ta 98				pta 25	pta 36	
		da 1				ba 9	pa 9	
		ka 1				bta 8		
ga	ga 100	ga 100	ga 100	ga 100	ga 100	ga 83	ga 100	ga 91
						bga 17	ga 91	
ka	ka 100	ka 100	ka 100	ka 100	ka 100	ka 82	ka 82	ka 75
						pka 18	pa 9	pa 8
							pka 9	pka 8
								bka 9
na	na 100	na 100	na 100	na 91	na 96	na 64	mna 50	na 55
				la 9	la 4	mna 18	na 25	mna 27
						ma 9	ma 25	ma 9
						bna 9		bna 9
ba	da 37	da 82	da 64	ga 58	da 58		ba 100	ba 100
	ga 37	ba 18	ga 27	da 25	ga 25	ba 100		
	ba 26		ba 9	ba 17	ba 17			
pa	pa 50	pa 50	pa 55	pa 70	pa 64	pa 83	pa 96	pa 100
	ka 50	ka 50	ka 45	ka 30	ka 27	ba 17	ba 2	
					pka 9		ta 2	
ma	na 91	na 82	na 92	na 91	na 80	ma 91	ma 100	
	ma 9	ma 9	nma 8	ma 9	ma 20	ba 9	la 9	ma 100

sponses becomes significantly high. In fact, in the lower left quadrant, all examples yield highly significant visually influenced response rates, from 30 to 100% with a mean of 73% ($p < 0.001$).

In the upper right quadrant, the percentage of these kinds of responses varies between 0% and 75% with a mean of 25% ($p < 0.05$). Identification of the non-labial auditory CV (/da, ta, ga, ka, na/), matched with bilabial visemes (/ba, pa, ma/), causes less visually influenced responses with respect to the opposite condition of A bilabial with V non-labial.

Discussion

Natural perceptions are rich with experiences from the auditory and visual modalities. Many experiments have shown that although the combination of acoustic and visual information goes unnoticed by the perceiver, (i.e., it does not evoke a distinct "multisensory" sensation), it has a strong effect on acoustic perception and can enhance: a) recognition of the speech in noise and degraded speech [1]; b) orientation [23]; c) classification [24] and d) reaction time [25].

The McGurk effect is a further and particular example of the role that visual information may have on perception and its robustness has been demonstrated for English, German, and Spanish, while in Japanese and Chinese languages it seems weaker [12-15 26 27]. In fact, Kuhl [26] and, recently, Sekiyama and Burnham [27] have demonstrated a weaker visual influence for Japanese, than for English language adults, while Massaro et al. [28] reported the same trend for Japanese with respect both to Spanish and to English speakers.

Our results demonstrate a marked McGurk effect for the Italian language, which has never been described before. The phenomenon is highly significant when an A bilabial CV is dubbed with a V apico-dental or velar CV. In the opposite condition, when a bilabial V CV is dubbed with an A non-labial CV, the effect is still significant, but to a lesser degree. Visual presentation of /da, ta, ga, ka, na/

does not influence the perception of the same auditory CV. Thus, the different places of articulation of /da, ta, ga, ka, na/ are not visually recognizable, with one possible exception for the apico-dental /na/. Finally, when bilabial CV are dubbed with the homologous visual CV, their perception is not modified. In other words, lipreading cannot alter the perception of the mode of articulation. Nevertheless, it is worthwhile pointing out that recent proposals arguing that the visual speech signal is rich in informational content, much more so than traditionally held, based solely on visemic confusion matrices would predict [29].

The Fuzzy Logical Method of Perception (FLMP), proposed by Massaro[30], could be more successful than other models in accounting for these experimental data. Briefly, when perceivers integrate auditory and visual sources of information, each source is more influential, to the extent that the other source is ambiguous.

The typical situation of V/ga/ vs. A/ba/ is reported, as an example, in Figure 1. Both modalities support /da/, to some degree, which would account for this alternative, even though one of the modalities supports another alternative to a greater degree. Each feature match is represented by a common metric of fuzzy logic truth values that range from 0 to 1 [31].

Thus, assume that:

nothing like = 0.1;

mostly nothing like = 0.3;

somewhat like = 0.7;

a lot like = 0.9.

Using multiplicative integration of FLMP:

support for /ga/ = 0.9 x 0.3 = 0.27

support for /ba/ = 0.1 x 0.9 = 0.9

then support for /da/ = 0.7 x 0.7 = 0.49.

As can be seen in this example, /da/ gets almost twice as much support as any other alternatives.

This analysis can be extended to include all the possible alternatives. Hence, our results are in good agreement with this explanation.

The present results obtained on visual influence in AV perception are similar to those observed in English and Spanish adult listeners. Thus, visual influence, in the present subjects is stronger with respect to that observed in Japanese adults. In a recent cross-language McGurk study, Sekiyama and Burnham [27] observed that Japanese adults are 100 msec faster than English language adults in the A condition, while English adults are faster in the visual only condition. At the same time, in uni-modal perception accuracy no difference is observed between these languages. As stressed by these Authors [27] "there is no essential difference in the availability of auditory and visual information for Japanese and English participants, but in the relative time course for the availability of auditory and visual unimodal information". Compared to Japanese, both English as well as Italian have more consonant contrasts and several visually, but not so many auditorally, distinct contrasts. Languages presenting more phonological complexity may require more attention to visual cues: in other words, languages with simpler phonetic cues may not demand the same AV integration as languages with more complex phonological aspects.

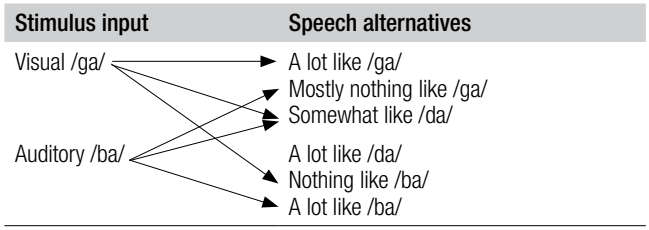

Fig. 1. Degree of support for speech alternatives when a V/ga/ is dubbed with an A /ba/.

On the opposite hand, Chen & Massaro [6] underline the fact that "Although the results of experiments with native English, Spanish, Japanese, and Dutch talkers showed substantial differences in performance across the different languages, the application of the FLMP indicated that these differences could be completely accounted for by information differences, with no differences in information processing" [6]. In other words, according to Massaro, there is no evidence to support the hypothesis of different types of audiovisual processing for speakers of different languages.

The McGurk phenomenon was considered as evidence of gestural theories: in fact, the motor theory of speech [32], as well as some more recent hypotheses [33] seem to demonstrate that the acoustic phonological cues are intrinsically linked with the articulation pattern of the visuo-facial movements.

On the other hand, following the auditory theories, visual information can influence speech perception since visual features and phonological representations have been acquired through the experience of watching the mouth of the talkers while listening to them speak. Whether audio-visual speech perception is best accounted for auditory or gesture-based theories of perception is still a debatable question. Furthermore, the classic model of audiovisual integration proposed by Smeele et al. [34] hypothesizes that information from different modalities is processed in a hierarchical fashion along unisensory streams, which converge in high-order structures. The combined representation is then processed in a feed-forward fashion that does not affect downstream processing. More recently, this model has been criticised and we cannot tell whether audiovisual integration occurs before (in early afferent processing) or after cortical processing and is a result of corticofugal modulation. In fact, the sites of convergence for acoustic speech and facial movements were previously considered to be structures as portion of the superior temporal sulcus [35], intraparietal [36], prefrontal and premotor cortices [37]. Evidence is

mounting that audiovisual integration begins in lower-order structures as the superior colliculus [38] or generally in the brainstem [39], thus requiring the modification of previous models.

At the same time, the McGurk effect was initially considered as a further example supporting the "special" nature of speech. Nevertheless, analogous illusions have been reported with non-native speech [13 40] and musical stimuli. In fact, non native stimuli induce more visual influence than native stimuli [5 14 26]. Furthermore, Saldaña and Rosenblum [41] evoked a robust McGurk effect by dubbing the visual movie of a cello player while producing pluck notes, with the audio reproduction of contrasting bow sounds.

In conclusion, the matter of how, where and when audiovisual integration is processed in the brain is still debatable. Nevertheless, the demonstration of the McGurk effect in several languages, among which, Italian, and the observed differences in the visual influence may add new insights into the mechanisms of AV integration. Most probably, more complex phonological languages, such as English, Spanish and Italian, possibly require a stronger AV integration, with respect to Japanese and Chinese, i.e., languages with simpler phonological cues.

References

1 Middelweerd MJ, Plomp R. *The effect of speechreading on the speech-reception threshold of sentences in noise*. J Acoust Soc Am 1987;82:2145-7.

2 McGurk H, MacDonald J. *Hearing lips and seeing voices*. Nature 1976;264:746-8.

3 Kuhl PK, Meltzoff AN. *The bimodal perception of speech in infancy*. Science 1982;218:1138-41.

4 Dodd B. *Lip-reading in infants: attention to speech presented in- and out-of-synchrony*. Cogn Psychol 1979;11:478-84.

5 De Gelder B, Bertelson P. *Multisensory integration, perception and ecological validity*. Trends Cogn Sci 2003;7:460-7.

6 Chen TH, Massaro DW. *Mandarin speech perception by ear and eye follows a universal principle*. Percept Psychophys 2004;66:820-36.

7 MacKain K, Studdert-Kennedy M, Spieker S, Stern D. *Infant intermodal speech perception is a left-hemisphere function*. Science 1983;219:1347-9.

8 Burnham D. *Visual recognition of mother by young infants: facilitation by speech*. Perception 1993;22:1133-53.

9 Massaro DW, Thompson LA, Barron B, Laren E. *Developmental changes in visual and auditory contributions to speech perception.* J Exp Child Psychol 1986;41:93-113.

10 Burnham D, Dodd B. *Auditory-visual speech integration by prelinguistic infants: perception of an emergent consonant in the McGurk effect*. Dev Psychobiol 2004;45:204-20.

11 Desjardins RN, Rogers J, Werker JF. *An exploration of why preschoolers perform differently than do adults in audiovisual speech perception tasks*. Exp Child Psychol 1997;66:85-110.

12 MacDonald J, McGurk H. *Visual influence on speech perception*. Percept Psychophys 1978;24:253-7.

13 Sekiyama K, Tohkura Y. *McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility*. J Acoust Soc Am 1991;90:1797-805.

14 Fuster-Duran A. *Perception of conflicting audio-visual speech: an examination across Spanish and German*. In: Stork DG, Hennecke ME, editors. *Speech-reading by humans and machines.* New York: Springer-Verlag; 1996. p. 135-43.

15 Werker JF, Frost PE, McGurk H. *La langue et les lèvres: cross-language influences on bimodal speech perception*. Can J Psychol 1992;46:551-68.

16 Grassegger H. *McGurk effect in German and Hungarian listeners.* In: Proceedings of the International Congress of Phonetic Sciences. Congress organizer at KTH and Stockholm University, Stockholm 1995. p. 210-3.

17 Munhall KG, Gribble P, Sacco L, Ward M. *Temporal constraints on the McGurk effect*. Percept Psychophys 1996;58:351-62.

18 Green KP, Kuhl PK, Meltzoff AN, Stevens EB. *Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect*. Percept Psychophys 1991;50:524-36.

19 Green KP, Gerdeman A. *Cross-modal discrepancies in coarticulation and the integration of speech information: the McGurk effect with mismatched vowels*. J Exp Psychol Hum Percept Perform 1995;21:1409-26.

20 Diesch E. *Left and right hemifield advantages of fusions and combinations in audiovisual speech perception*. Q J Exp Psychol A 1995;48:320-33.

21 Martini A, Bovo R, Agnoletto M, Da Col M, Drusian A, Liddeo M, et al. *Dichotic performance in elderly Italians with Italian stop consonant-vowel stimuli.* Audiology 1988;27:1-7.

22 Massaro DW, Cohen MM, Smeele PM. *Perception of asynchronous and conflicting visual and auditory speech*. J Acoust Soc Am 1996;100:1777-86.

23 Zambarbieri D. *The latency of saccades toward auditory targets in humans*. Prog Brain Res 2002;140:51-9.

24 Ben-Artzi E, Marks LE. *Visual-auditory interaction in speeded classification: role of stimulus difference*. Percept Psychophys 1995;57:1151-62.

25 McDonald JJ, Ward LM. *Involuntary listening aids seeing: evidence from human electrophysiology.* Psychol Sci 2000;11:167-71.

26 Kuhl PK. *Learning and representation in speech and language*. Curr Opin Neurobiol 1994;4:812-22.

27 Sekiyama K, Burnham D. *Impact of language on development of auditory-visual speech perception.* Dev Sci 2008;11:306-20.

28 Massaro DW, Tsuzaki M, Cohen MM, Gesi A, Heredia R.

Bimodal speech perception: an examination across languages. J Phonetics 1993;21:445-78.

[29] Soto-Faraco S, Navarra J, Weikum WM, Vouloumanos A, Sebastián-Gallés N, Werker JF. *Discriminating languages by speech-reading*. Percept Psychophys 2007;69:218-31.

[30] Massaro DW. *Perceiving talking faces: from speech perception to a behavioural principle*. Cambridge, MA: MIT Press; 1998.

[31] Zadeh LA. *Fuzzy sets*. Information & Control 1965;8:338-53.

[32] Liberman AM, Mattingly IG. *The motor theory of speech perception revised*. Cognition 1985;21:1-36.

[33] Rizzolatti G, Craighero L. *The mirror-neuron system*. Ann Rev Neurosci 2004;27:169-92.

[34] Smeele PM, Massaro DW, Cohen MM, Sittig AC. *Laterality in visual speech perception*. J Exp Psychol Hum Percept Perform 1998;24:1232-42.

[35] Giard MH, Peronnet F. *Auditory-visual integration during multimodal object recognition in humans: a behavioural and electrophysiological study*. J Cogn Neurosci 1999;11:473-90.

[36] Calvert GA, Campbell R, Brammer MJ. *Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex*. Curr Biol 2000;10:649-57.

[37] Bushara KO, Grafman J, Hallett M. *Neural correlates of auditory-visual stimulus onset asynchrony detection*. J Neurosci 2001;21:300-4.

[38] Calvert GA. *Crossmodal processing in the human brain: insights from functional neuroimaging studies*. Cereb Cortex 2001;11:1110-23.

[39] Musacchia G, Sams M, Nicol T, Kraus N. *Seeing speech affects acoustic information processing in the human brainstem*. Exp Brain Res 2006;168:1-10.

[40] Sekiyama K, Kanno I, Miura S, Sugita Y. *Auditory-visual speech perception examined by fMRI and PET*. Neurosci Res 2003;47:277-87.

[41] Saldaña HM, Rosenblum LD. *Visual influences on auditory pluck and bow judgments*. Percept Psychophys 1993;54:406-16.

Received: May 30, 2009 - Accepted: July 22, 2009

Address for correspondence: Dr. R. Bovo, Unità Operativa e Cattedra di Audiologia e Foniatria, Università di Ferrara, c.so Giovecca 203, 44100 Ferrara, Italy. Fax: +39 0532 237447. E-mail: roberto.bovo@unife.it