On the Perception of Visual Durational Speech Features: A Comparison between Native and Non Native Speakers

A. Esposito*, M. Esposito*, M. Giagkou***, A.Vatakis**]*, A. Vinciarelli**
* Seconda Università di Napoli, Department of Psychology and IIASS, Italy
** University of Glasgow, Department of Computing Science, Glasgow, UK
***Institute for Language and Speech Processing, Athens, Greece
****Cognitive Systems Research Institute (CSRI), Athens, Greece
iiass.annaesp@tin.it, mirko.esposit@gmail.com , mgiagkou@gmail.com, argiro.vatakis@gmail.com, Alessandro.Vinciarelli@glasgow.ac.uk

***Abstract*— Some languages as Italian and Hindi, are endowed with a set of consonants that are discriminated among them only by the length of the consonant closure (single versus geminate consonants). Consonant closure is a speech temporal feature. This work investigates on visual durational cues to the point they help native and non native speakers in discriminating between single and geminate consonants, in order to ascertain proto multimodal interactive processing mechanisms of time in languages. The comparison is made between a group of Italian subjects, since Italian has geminate consonants, and a group of non natives in order to check if this ability can be attributed to the language expertise. The results show that non natives are able to discriminate between single and geminate consonants even though their performance is not as good as in native speakers, suggesting an innate ability to process visual temporal features in language tasks for supporting the effort to decode the exchanged messages.**

I. INTRODUCTION

To the best of our knowledge, there are no studies investigating on the perception of visual durational speech features. For this reason, this work must be considered as an exploratory research still running for collecting evidence among different cultures and languages. The exact timing of sentences, words, and language phones plays an important role in the production and the efficient transmission of a message. The effects of an appropriate timing sweep from changing the prosody and the emphasis of a message to distinguishing the beginning and end of a clause, changing the meaning of two similar words and, in some languages, making a distinction among language phones (both vowels and consonants depending on the language, as for example single and geminate consonants in Italian and Hindi or long and short vowels in Swedish (see [10, 11] for a review).

Historical findings shows that a word extracted from fluent speech should last at least 800 ms in order to be correctly decoded by a native speaker [7], suggesting that timing in language is more than the mere consequence of the articulation time and that it must be more deeply investigated in order to understand and explain its effects on the properly semantic understanding and comprehension of a message. In Italian, at phone level, timing is the most important feature to discriminate between Italian single and geminate consonants [1] and it seems to function as a categorical feature in sweeping the perception of Italian voiced and voiceless perceptual cues (producing on the perception the same effects of the Voice Onset Time (VOT) [6, 2]). This claim has not been confirmed by an experimental setting being the results of an informal test lead by one of the authors during the conference on *Update on Specific Language Impairment* held in Urbino, Italy, in 2005. On that occasion, about 200 native Italian speakers and language experts were asked to listen to two signals. The former was the word */dato/ (data)* where the [*t*] closure lasted 138ms. The latter was the same signal where the [*t*] closure was artificially shortened to 15ms. The "informal" result was that "subjects", in total agreement, attributed the voiceless feature to the long [*t*] (interpreting the word as */dato/*) and the voiced one to the short [*t*] (interpreting the word as */dado/ (dice)*). Despite their exactitude, these findings reinforce the idea that a correct timing is basic for an accurate message understanding. Even though recent findings attribute to a specific neural circuit in the cerebellum [4] the processing of temporal relations among events up to milliseconds, we are doubtful that it could be adequate to explain the complex relationship that ties timing to spoken languages. The complex timing processing required by linguistic tasks must require a convoluted neural circuitry involving many specialized neural paths rather than a single amodal internal stopwatch.

There are plenty of papers tying time and language processing from an auditory perspective (see [2] for a review). From a multi-modal interactive perspective it would be useful to have more information on the visual counterpart. This information can be basic to rebuild from scratch timing models for speech synthesis and/or speech recognition technologies. The present work aims to define an experimental setting useful to explore the role of timing in the execution and comprehension of language tasks from a visual point of view and possibly to support, as it has been show for the well know McGurk effect [8], the idea that multi-modal interactive brain mechanisms concurrently contribute to the processing of time in language tasks.

A. *Testing the visual perception of time in language*

The perception of speech visual durational cues was tested asking participants to watch a set of mute videos where an actor was pronouncing either a single or a

978-1-4673-5188-1/12/$31.00 ©2012 IEEE

geminate consonant and attribute to it the correct consonant category. The Italian language helps in this setting because of the high number of minimal pairs that change meaning according to the consonant (single vs. geminate or short vs. long consonant). For example the words /*papa*/ (pope) and /*pap:a*/ (baby food) are differentiated only by the geminate /p:/. From an acoustic point of view the two consonants /p/ and /p:/ mainly differ in the length of the consonant closure, and to some extent from the length of the preceding vowel, which if followed by a geminate consonant will be slightly shortened [1]. A correct subject categorizing (both native and non native speakers) of single vs. geminate consonants would bear to hypothesize the existence of proto linguistic visual mechanisms for speech timing resulting from the interaction of visual and auditory neural pathways. The main questions posited are:

- Should we assume the existence of domain specific components (time versus spatial) within mechanisms for encoding time in general and/or timing in spoken languages?
- Are these components cultural specific and therefore learned and language dependent?

II. TESTING ITALIAN SPEAKERS

A. Materials and Methods

In order to test the abovementioned hypotheses two tasks are defined, where subjects watch mute videos one at the time and are asked to judge the consonant quantity (single vs. geminate) uttered either at a *Conversational* (20 subjects) or at a *Hyper-articulated*[1] (20 subjects) speaking rate. The dependent measured variables are the *number (#) of correct answers*, and independently of the correctness, the *number (#) of stimuli perceived as single or geminate*.

Video clips (~ 4 sec long) were collected from one male Italian native speaker producing conversational and hyper-articulated single and geminate utterances as in the following example:

- **Single case**: Dico /*papa*/ chiaramente (I am clearly saying /pope/)

- **Geminate case**: Dico /*pap:a*/ chiaramente (I am clearly saying /baby food/)

The subjects were trained before gathering the experimental data using 10 mute videos (5 minimal word pairs of single and corresponding geminate consonants). Each experimental task (*Conversational* versus *Hyper-articulated* exploits 14 videos equally balancing the single with the corresponding geminate consonant (the lists of the 7 Italian minimal word pairs used in the *Training*, *Conversational*, and *Hyper-articulated* conditions are reported in the APPENDIX). A total of 40 Italian native speakers with no familiarity with the tasks (aged from 18 to 30 years, and equally balanced for gender) participated in the experiments.

[1]The hyper-articulated speech, according to the H and H theory [9], is defined as the speech for which clarity tends to be maximized in contrast to the hypo-articulated speech (*conversational*) which is produced with minimal efforts.

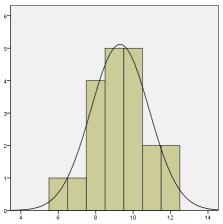

Figure 1. Distribution of the # of correct answers (x-axis) over the # of Italian participants (y-axis) for the conversational task.

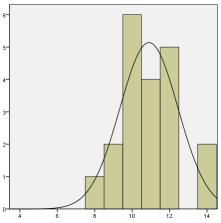

Figure 2. Distribution of the # of correct answers (x-axis) over the # of Italian participants (y-axis) for the hyper-articulated task.

TABLE I.
"STIMULI PERCEIVED AS" BY THE ITALIAN SUBJECTS

# stimuli perceived as	single	geminate	Mean single	Mean geminate	Binomial distribution
Conversational	136	144	6.8	7.2	ρ=0.705
Hyper-articulated	118	162	5.9	8.1	ρ=0.005

B. Results

Results showed that Italian native speakers discriminate single versus geminate consonants in both *Conversational* and *Hyper-articulated* tasks. TABLE I reports the gathered # of *correct answers* and the corresponding binomial distributions showing that, in both conditions, the subjects' choices are not due to chance. Figures 1 and

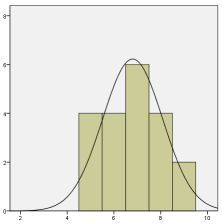

Figure 3. Distribution of the # of stimuli perceived as a single consonant (x-axis) over the # of Italian participants (y-axis) in the conversational task.

2 illustrate the distributions of the # of the Italian subjects' *correct answers* in the *conversational* and *hyper-articulated* task respectively.

A repeated ANOVA measurement on the # of stimuli *perceived as* single or geminate (*within dependent variable*) computed for both the *Conversational* ($F(1,19) = .487$, $p = 0.494$) and *Hyper-articulated* condition ($F(1,19) = 13.604$, $p = 0.002$) suggested in the latter case a tendency to perceive stimuli more as geminate than single consonants.

The # of stimuli *perceived as single* over those *perceived as geminate* consonants is summarized in Table II. Figures 3 and 4 illustrate the distributions of the # of stimuli perceived as a single consonant in the conversational and hyper-articulated conditions.

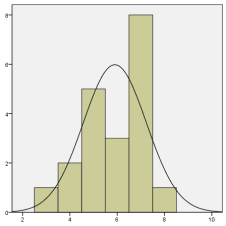

Figure 4. Distribution of the # of stimuli perceived as a single consonant (x-axis) over the # of Italian participants (y-axis) in the hyper-articulated task.

Finally, a one way ANOVA testing the differences among the # of stimuli *perceived as geminate* in both the Conversational and Hyper-articulated group (*between dependent variable*) confirmed ($F(1,38) = 4{,}735$, $p = 0.036$) that a slower speaking rate affects the perception of the consonant quantity towards an overestimation of the geminate consonants in the *hyper-articulated* condition.

III. TESTING NON NATIVE SPEAKERS

The second experiment aims to test if Italian geminates can be visually discriminated by foreign language speakers (having English as second language and belonging, as native speakers, to different language groups) who do not speak and understand Italian. The experimental hypotheses driving the research are to check the foreign subjects' ability to decode visual durational features, and therefore, correctly discriminate the consonant quantity both in the *hyper-articulated* and *conversational* task, as well as to investigate if the speaking rate affects their performance. The results should shed light on the language dependence question posed in the introductive section.

A. Materials and Methods

The experimental set-up was appropriately modified to allow non native speakers to perform the tasks. A total of 40 equally gender balanced subjects (aged from 18 to 34 years) are tested. They are divided into 2 groups, each of 20, both watching mute videos of *hyper-articulated* and *conversational* speech productions according to three main procedural aspects:

TABLE II.
RESULTS ON THE TESTING OF ITALIAN SUBJECTS

# of *correct answers* on 280 samples in	# *correct answers*	Mean	Binomial distribution
Conversational speech	186	9.3	ρ=0.000
Hyper-articulated speech	219	10.9	ρ=0.000

- ***Words*** *vs.* ***Phrases*** *(Between)*: subjects watched videos of mute productions of isolated vs. sentence embedded words (20 subjects were assigned to isolated words and 20 to sentences). The words are the same in the two main groups (*Words vs. Phrases*) and the same used for the Italian subjects.
- ***Conversational*** *vs.* ***Hyper-articulated*** *(Within)*: subjects watched videos of mute *conversational* vs. *hyper-articulated* productions (either of words or sentences);
- ***Sequence 1*** *vs.* ***Sequence 2*** *(Between)*: To check the sequencing effects, for each group, 10 subjects tackled the conversational and then the hyper-articulated productions and for the remaining the sequencing was reversed.

The dependent measured variables are the *number (#) of correct answers*, and independently of the correctness, the *number (#) of stimuli perceived as single or geminate*.

Before starting the experimental tasks, subjects were trained on the same material used for the Italian subjects, with a different procedural approach. There was first an *Audio Training* procedure tackled by both the groups (*Words vs. Phrases*). There, subjects are first asked to listen to word samples while the correct label on the consonant quantity is shown and then to listen to the same word samples and label them on their own. Then, the *Phrase group* (but not the *Word group*) tackled a *Video Training* procedure where subjects are asked to watch segmented word by word videos. Two of the segmented video words (/*dico*/ and /*chiaramente*/) formed the carrying utterance, while the third word belonged to the set of minimal pairs used for training the Italian subjects. This was done in order to allow the non natives to familiarize with the Italian sentence structure and be able to focus on the target word. Finally, both the *Word* and *Phrase* groups were trained to label isolated or sentence embedded words as single or geminate using only the visual information. The lists of words used for training and testing are the same ones used for the Italian subjects and are reported in the APPENDIX. All the data were gathered using "Super lab" software. Video clips (~ 4 sec long) were collected from the same male Italian native

TABLE III.
RESULTS ON THE TESTING OF NON NATIVE ITALIAN SUBJECTS

# of *correct answers* on 280 samples in	# *correct answers*	Mean	Binomial distribution
Conversational words	155	7.75	ρ=0.041
Hyper-articulated word	191	9.55	ρ=0.000
Conversational speech	162	8.1	ρ=0.005
Hyper-articulated speech	179	8.7	ρ=0.000

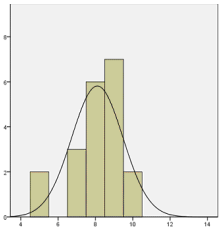

Figure 5. Distribution of the # of correct answers (x-axis) over the # of non native participants (y-axis) in the conversational task.

speaker producing *conversational* and *hyper-articulated* isolated single and geminate words while the stimuli for the word embedded utterances were the same used for the Italian subjects.

A total of 40 gender balanced subjects (aged from 18 to 34 years) are tested, divided into 2 groups each of 20, according to the main manipulated variables, i.e. the production of isolated single and geminate Italian words versus the same words embedded in a phrase. Both groups watched mute videos of *hyper-articulated* and *conversational* speech productions.

Moreover, within the groups (*Word vs. Phrase group*) two more groups were created, where the first group of 10 subjects are asked to answer on *conversational* and then *hyper-articulated* speech, while for the second group the presentation order was reversed.

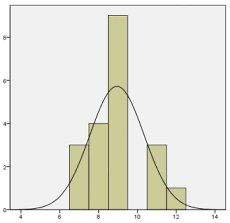

Figure 6. Distribution of the # of correct answers (x-axis) over the # of non native participants (y-axis) in the hyper-articulated task.

B. Results

Results showed that non native speakers discriminate mute versions of single versus geminate consonants in both *Conversational* and *Hyper-articulated* tasks, since, as reported in TABLE III, the gathered number of *correct answers* is above the chance in both conditions. In addition, no significant differences were found between the groups (*Word* vs. *Phrase* group - $F(1,78) = .104$, $\rho = .748$) as well as no sequencing effects (*Sequence 1* vs. *Sequence 2* - $F(1,38) = 2.119$, $\rho = .154$). For these reasons, the data gathered from the *word group* are not reported in the following. Figures 5 and 6 illustrate the distributions, in the *Phrase group* of the number of non native participants' correct answers in the c*onversational* and *hyper-articulated* condition respectively.

A repeated ANOVA measurement ($F(1,19) = 4.534$, $\rho = .047$) performed on the *Phrase group* showed that also for non natives there seems to be a tendency of the speaking rate to affect the # of correct *answers*. This tendency was checked computing the # of stimuli correctly identified either *as single* or *geminate* consonants. Data are summarized in TABLE IV and illustrated in Figures 7 and 8 for the non native speakers.

A repeated ANOVA measurement on the # of stimuli *perceived as* single or geminate (*within dependent variable*) computed in both the *Conversational* ($F(1,19) = .322$, $\rho = 0.577$) and *Hyper-articulated* condition ($F(1,19) = 8.435$, $\rho = 0.009$) confirmed for the non native, as for the Italian subjects, a tendency, at a slower speaking rate, to perceive consonant stimuli more as geminate than single ones.

A one way ANOVA shows that the performance of Italian and non native participants is significantly different both for the *Conversational* ($F(1,38) = 6.673$, $\rho = 0.014$) and the *Hyper-articulated conditions* ($F(1,38) = 17.462$, $\rho = 0.000$) suggesting that native speakers are more accurate in both conditions.

IV. Discussion and Conclusions

The main results carried out on Italian speakers suggest that they can visually perceive and correctly judge slight time variations in language tasks, allowing to hypothesize an implicit and spontaneous capacity of visual timing analysis, both in ecological (*conversational*) and artificial (*hyper-articulated*) settings. The Italian perception accuracy is affected by the s*peaking rate* suggesting that Italian speakers exploit a *"visual timing threshold"* to distinguish between short and long consonants. It is reasonable to assume that below this threshold the consonant will be perceived as single. The results indicate a fine tuning of this threshold at a natural (*conversational*) and a biased tuning at an artificial (*hyper-articulated*) speaking rate since a tendency to perceive more geminates has been systematically observed only in the *hyper-articulated* condition. The bias can be attributed to the impossibility to adjust the geminate threshold to a natural value in the artificial setting. Therefore, participants would perceive more geminates because more consonant closures are above the single consonant visual duration.

TABLE IV.
"Stimuli perceived as" by the non native Italian subjects

# stimuli perceived as	single	geminate	Mean single	Mean geminate	Binomial distribution
Conversational Phrases	144	136	7.2	6.8	0.705
H-articulated Phrases	123	157	6.15	7.85	0.024

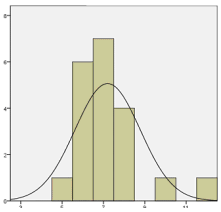

Figure 7. Distribution of the # of stimuli perceived as a single consonant (x-axis) over the # of non native participants (y-axis) in the conversational task.

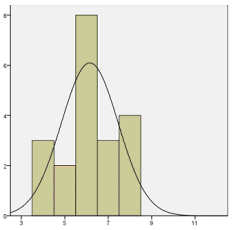

Figure 8. Distribution of the # of stimuli perceived as a single consonant (x-axis) over the # of non native participants (y-axis) in the hyper-articulated task.

Is this a natural consequence of visual time processing or is it due to the language expertise of the involved subjects, being Italian native speakers?

In the following experiments it was observed that also non native speakers can visually infer the quantity of an Italian single and/or geminate consonant both in isolated and sentence embedded word productions, suggesting a *bottom-up processing* [3, 5] of the visual durational

features, i.e., an innate ability to process visual temporal features in language tasks. Perhaps, the observed articulation is internally simulated producing a pattern of activation quickly processed by proto multimodal linguistic processing mechanisms operating within a very fine temporal threshold of milliseconds and encoding visual durational speech features. Again, the *speaking rate* is a prominent factor for the subject's effectiveness in her/his response accuracy suggesting that these mechanisms are tuned to process visual duration in an ecological setting and suffers of the artificial variations imposed by the *hyper-articulated* condition. **N**on natives are affected by the very same *"threshold bias"* observed for Italian subjects, leading them to the inability to adapt the discriminating threshold and consequently to detect more geminate than single consonants in the artificial language task condition.

The overall superior degree of accuracy (in both the *conversational* and *hyper-articulate* condition) observed for the Italian when compared with the non native speakers can be explained considering that they are eased in the task because of their language expertise, i.e., they can rely on cognitive processes (Top-Down processing [12]) allowing them to enhance their performance in the ecologic setting and better adapt to an artificial speaking rate.

More experimental data are needed to assess these hypotheses and comparisons with different language speakers are planned. However, these data are extremely useful for future Info Communication Technologies in order to model the synthesis of multimodal facets of timing in language tasks and implement virtual avatars and intelligent interactive dialogue systems which exploit either visual/vocal or visual feedback only for human-machine interaction. To our knowledge, there are no, to date, research works tackling this issue. This paper opens new perspectives in modelling time in speech.

ACKNOWLEDGMENTS

This work has been supported by the European COST Action ISCH TD0904 "TMELY: Time in MEntal activitY (www.timely-cost.eu). Acknowledgements go to two unknown reviewers for their useful comments and suggestions and to Miss Tina Marcella Nappi for her editorial help.

APPENDIX

In the following are reported the lists of minimal word pairs used in the above reported experiments:

Minimal word pairs used for the subject's Training :

Fumo *(Smoke)* - Fummo *([we] Were)*
Lese *(Injuried)* - Lesse *(Boiled)*
Cari *(Dears)* - Carri *(Carts)*
Beve *([he] Drinks*) - Bevve *([he] Drunk)*
Tufo *(Tuff)* - Tuffo (Dive)

Minimal word pairs used for the Conversational set-up:

Cane (*Dog*) - Canne *(Reeds)*
Casa *(House)* - Cassa *(Chest)*
Dita *(Fingers)* - Ditta *(Company)*
Faro *(Lighthouse)* - Farro *(Spelt)*
Pala *(Shovel)* - Palla *(Balloon)*
Papa *(Pope)* - Pappa *(Baby food)*
Tuta *(Tracksuit)* - Tutta *(All)*

Minimal word pairs used for the Hyper-articulated set-up:

Nono *(the Nineth)* - Nonno *(Grandfather)*
Mese *(Month)* - Messe *(Masses)*
Fata *(Fairy)* - Fatta *(Build, Done)*
Poro *(Pore)* - Porro *(Leek)*
Vile *(Coward)* - Ville *(Residences, Mansions)*
Capa *(Head)* - Cappa *(Cape)*
Moto *(Motorbike)* - Motto *(Saying)*

REFERENCES

[1] A. Esposito and N. Bourbakis "The role of timing in speech perception and speech production processes and its effects on language impaired individuals," in Proceeding of the IEEE 6th Symposium on BioInformatics and BioEngineering (BIBE'06).

[2] A. Esposito and M.G. Di Benedetto, "Acoustical and perceptual study of gemination in Italian stops" *JASA*, vol. 106(4), pp.2051-2062, 1999.

[3] G.J. Feist, T.E. Bodner, J.F. Jacobs, M. Miles, and V. Tan "Integrating top-down and bottom-up structural models of subjective well-being: A longitudinal investigation," *Journal of Personality and Social Psychology*, vol. 68(1), pp. 138-150, 1995.

[4] R.B. Ivry, T.C. Justus, and C. Middleton, C (2001). "The cerebellum, timing, and language: Implications for the study of dyslexia". In *Dyslexia Fluency and the Brain*, M. Wolf ed., Timonium, MD: York Press, 198-211, 2001.

[5] R.A. Kinchla and J.M. Wolfe "The order of visual processing: Top-down, bottom-up or middle-out*," Attention, Perception & Psychophysics*, vol. 25(3), pp. 225-231, 1979.

[6] L. Lisker, and A.S. Abramson "A cross-language study of voicing in initial stops: acoustical measurements," *Word*, vol. 20, pp. 384-422, 1964.

[7] J.M. Pickett and I. Pollack "Intelligibility of excerpts from fluent speech: Effects of rate of utterance and duration of excerpt," *Language & Speech*, vol. 3, pp. 151-164, 1963.

[8] A.R. Nath and M.S. Beauchamp "A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion," *NeuroImage,* vol. 59(1), 781-787,2011.

[9] B. Lindblom "Explaining phonetic variations: A sketch of the Hand H theory". In *Speech Production and Modelling*, W. J. Hardcastle, A. Marchal eds., Kluwer Academic Publishers, 403-409,1990.

[10] M. Kenstowicz "On the notation of vowel length in Lithuanian," *Papers in Linguistics* 3,73–113, 1970.

[11] M. Kenstowicz, *Phonology in generative grammar*. Blackwell Publishers, 1994.

[12] W. Tyler "The effects of expectations on perception: Experimental design issues and further evidence," Working paper series Federal Reserv Bank of Boston, 07-14, http://hdl.handle.net/10419/55632, 2007.