A COMPARATIVE ACOUSTIC STUDY OF SPONTANEOUS AND READ ITALIAN SPEECH

Emanuela Magno Caldognetto, Claudio Zmarich, Franco Ferrero

Centro di Studio per le Ricerche di Fonetica, C.N.R.
Via G. Anghinoni 10, 35121 Padova, Italia
Tel. 049- 8274409 ; Fax 049- 8274416; E-mail Magno@csrf00.csrf.pd.cnr.it

ABSTRACT

This paper presents an acoustic study of spontaneous and read Italian speech based on the analysis of monologues and corresponding read transcribed texts, each produced by three different subjects. The speaking styles were examined in terms of articulation, speech, fluency and word rate indices; typology of pauses and their cooccurrence; mean and range of F0 values; classification of phonetic events resulting from adjacency of vowels situated at word boundaries.

1. INTRODUCTION

In recent years the attention of linguists and psycholinguists has focused on the study of spontaneous speech [1], the former mainly to provide a systematic description of the phonetic and phonological variability that depends on speaking style [2], [3], and the latter to gain a better understanding of the mechanisms behind the linguistic planning and execution processes [4], [5] [6]. Furthermore, the study of speech styles is significant even from the point of view of its application in the field of speech technologies [7]: in speech synthesis programs, in order to achieve speech that is not only intelligible but also natural [8], or into automatic speech recognition systems [9], in order to introduce the ability to detect hesitation phenomena. Data referring to Italian speech could be useful for comparing data collected in similar studies on other languages [10], [11], [12], [13] and for pointing out the effects due to the particular syllabic and lexical features of the Italian language. For example, the very frequent CV syllable configuration tends to favour the adjacency of vowels situated at word boundaries [14], thus giving rise to events involving *hiatus* (maintaining of the two original segments), dipthongization, deletion of one of two different segments, deletion of one of two equal segments, *sinaloephe* of the two original segments into a new one, with effects on syllabic counts.

2. METHOD

For this purpose, three monologues were recorded, each produced by three different subjects (students of University of Padova), in presence of a listener non-interacting . During this recordings, the subjects were asked to give descriptions of some personal experience. The duration of these monologues varied between 34 and 100 seconds. These same three subjects were later asked to read the ortographic transcription of their speeches (after elimination of disfluencies). The recordings (total duration=354.30 s) were then converted to digital form at 16 kHz and 16 bits on a PC equipped with an CSL 4300 A/D converter board and then analysed with DSP CSL 4300 software produced by Kay Elemetrics Corp. The graphical representation of acoustic data in terms of waveform envelopes, wide band spectrograms, F0 contours and amplitude contours enabled a phonetic transcription of all the recordings in IPA symbols. Furthermore, unfilled pauses and a number of verbal phenomena such as filled pauses (hesitations), vowel lengthenings, glottalizations, disfluencies (interruptions and repetitions), speech repairs and slips of the tongue were detected, measured and hence classified. As to lengthening, this was measured on the initial and final positions of phonetic chains, by subtracting the vowel duration value of read text from the value of the corrisponding phone of spontaneous text. In this way, we were able to calculate the extra-lengthenings deriving specifically from hesitation phenomena, exceeding lengthenings due to prosodic boundaries common to both texts.

For each speaker and each speaking style the following units were identified:

Total text: the whole speech production, i.e. the sum of speech chains, unfilled pauses, filled pauses, disfluencies;

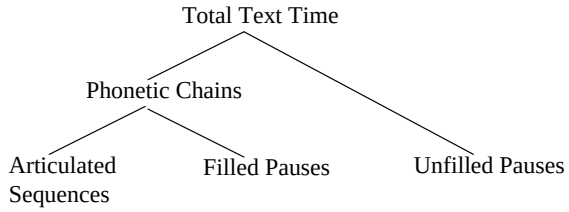
Unfilled pauses: any silent or breathing interval between two successive phonetic chains, excluding the closure phase of any voiceless plosive consonant starting the successive phonetic chain;

Phonetic chains: the sequence of phonetic segments (including filled pauses and disfluencies) delimitated by two silent pauses ;

Filled pauses: any occurrence of hesitations, interjections, abnormally lengthened vowels, repetitions etc.;

Articulated sequences: any phonetic chains excluding all filled pauses.

The hierarchical relations among these units are illustrated in the following graph, that highlights the difference between our approach and those of other



DURATION and NUMBER of PHONETIC UNITS										
		SPK. 1		SPK. 2		SPK. 3				
		READ	SPON.	READ	SPON.	READ	SPON.			
s	Articulat. sequen. time	28.72	28.39	58.72	69.99	39.04	46.25			
s	Filled pauses time	0	10.87	0	12.55	0	6.15			
s	(Phonetic chains time)	28.72	39.26	58.72	82.54	39.04	52.40			
s	Unfilled pauses time	5.22	5.82	11.25	17.27	5.26	8.81		TOTAL MEAN	
s	Total text time	33.94	45.08	69.96	99.81	44.30	61.21		READ	SPON.
%	*Articulat. sequen. time*	*84.6*	*63.0*	*83.9*	*70.1*	*88.1*	*75.6*		*85.5*	*69.5*
%	*Filled pauses time*	*0*	*24.1*	*0*	*12.6*	*0*	*10.0*		*0*	*15.6*
%	*(Phonetic chains time)*	*84.6*	*87.1*	*83.9*	*82.7*	*88.1*	*85.6*		*85.5*	*85.1*
%	*Unfilled pauses time*	*15.4*	*12.9*	*16.1*	*17.3*	*11.9*	*14.4*		*14.5*	*14.9*
%	*Total text time*	*100*	*100*	*100*	*100*	*100*	*100*		*100*	*100*
n.	Articulat. sequen. num.	11	14	25	33	14	16			
n.	Filled pauses num.	0	14	0	26	0	14			
n.	(Phonetic chains num.)	11	14	25	33	14	16			
n.	Unfilled pauses num.	10	13	24	32	13	15			
n.	Syllable number	182	180	379	386	231	236			
n.	Word number	96	105	172	180	127	127			

Fig. 1: see text for explanations

Authors [4] [5] [9] [15].

3. RESULTS

Quantitative (indices of fluency and F0 characteristics) and qualitative (pauses constituency and vowels adjacency at word boundaries) analysis were performed on all speech productions.

3.1. Indices of fluency.

For each phonetic chain and each total text, measurements were made regarding absolute and percentage duration, after which the number of syllables and words actually produced was computed (see fig.1). The total of all the units considered always turned out to be greater in spontaneous as compared to read speech texts. The percentage assessment demonstrates the fact that, though the duration of unfilled pauses was the same in the two styles, filled pauses were present (15%) only in spontaneous speech. The number of articulatory sequence events, unfilled and filled pauses, words and syllables was always greater in spontaneous than in read speech.

Next, the following indices were calculated:

Articulation rate i.: the number of syllables divided by the articulated sequence time (syll./s);

Speech rate i.: the number of syllables divided by the phonetic chain time (syll./s);

Fluency rate i.: the number of syllables divided by the total text time (syll./s);

Word rate i.: the number of words divided by the articulated sequence time (words/s).

Figure 2 illustrates the mean values of these indices, for each task and for each subject.

On the basis of the first three indices considered,

FLUENCY INDICES									
	SPK. 1		SPK. 2		SPK. 3		TOTAL MEAN	
	READ	SPON.	READ	SPON.	READ	SPON.	READ	SPON.
Articulation rate mean	6.2	6.0	6.4	6.3	5.7	5.4	6.1	5.9
Articulation rate C.V.	*0.13*	*0.24*	*0.09*	*0.50*	*0.10*	*0.19*	*0.11*	*0.31*
Speech rate mean	6.2	4.5	6.4	4.7	5.7	4.4	6.1	4.5
Speech rate C.V.	*0.13*	*0.33*	*0.09*	*0.26*	*0.10*	*0.28*	*0.11*	*0.29*
Fluency rate mean	5.2	3.8	5.3	3.9	5.0	3.7	5.2	3.8
Fluency rate C.V.	*0.16*	*0.35*	*0.16*	*0.32*	*0.15*	*0.29*	*0.16*	*0.32*
Word rate mean	3.4	3.7	2.9	3.2	3.2	3.3	3.2	3.4
Word rate C.V.	*0.16*	*0.28*	*0.23*	*0.93*	*0.12*	*0.45*	*0.17*	*0.55*

Fig. 2: see text for explanations

F0 PARAMETERS									
		SPK. 1		SPK. 2		SPK. 3		TOTAL MEAN	
		READ	SPON.	READ	SPON.	READ	SPON.	READ	SPON.
Hz	F0 mean	203	234	206	180	200	194	203	203
	F0 C.V.	*0.07*	*0.13*	*0.07*	*0.07*	*0.03*	*0.06*	*0.07*	*0.09*
Hz	F0 min	119	125	100	100	124	100	114	108
Hz	F0 max	400	470	380	390	301	333	360	398
Hz	F0 range	281	345	280	290	177	233	246	289

Fig. 3: see text for explanations

Percentages of OCCURENCE of various types of PAUSES broken down per Speakers & Styles							
		SPK. 1		SPK. 2		SPK. 3	
n	TYPOLOGY	READ	SPON.	READ	SPON.	READ	SPON.
1	silence	10		28	3.2	38.5	
1	inspiration			8	3.2	15.4	6.7
2	silence+inspiration	40		12	9.8	15.4	
2	inspiration+silence			4			
2	vowel lengthening+silence	10	15.4	20	22.6	15.4	26.6
2	vowel lengh.+inspiration		7.7	16	13	7.7	
2	silence+ vowel lengthening				3.2		
2	silence+glottal stop			8	3.2		
2	inspiration+ vowel lengthen.	10					
3	vow.lengt.+silen.+inspir.	30	7.7	4	3.2	7.7	20
3	vow.lengt.+silen.+hesitat.				13		13.3
3	vow.lengt.+silen.+laringal.				3.2		
3	vow.lengt.+ hesitat.+silen.				3.2		
3	vow.lengt.+insp.+ vow.lengh.		7.7				
3	vow.lengt.+insp.+ hesitat.		30.8				13.3
3	silen.+insp.+ hesitat.				3.2		6.7
3	silen.+ hesitat.+silen.		7.7				6.7
3	insp.+silen.+ hesitat.				3.2		
4	vow.len.+insp.+sil.+ vow.len.		7.7				
4	vow.len.+insp.+sil.+ hesitat.		7.7				6.7
4	vow.len.+sil.+insp.+ vow.len.		7.7				
4	vow.len.+sil.+insp.+glottal.				3.2		
4	vow.len.+sil.+glottal.+ hesitat.				3.2		
4	sil.+insp.+sil.+ vow.len.				3.2		
4	insp.+sil.+ hesitat.+sil.				3.2		
TOTALS %		100	100	100	100	100	100

Fig. 4: see text for explanations

spontaneous speech appeared to be less fluent than read speech. However, it must be noted that the coefficient of variation (CV) of these three indices is much greater in spontaneous rather than in read speech, thus testifying the non uniformity of production conditions along the various phonetic chains in spontaneous speech. The word rate index instead provided greater values in spontaneous speech, due to the greater incidence of reduction processes, not only at word boundaries, but also and especially within the words themselves.

3.2. F0 Characteristics

For each phonetic chain and each whole text, measurements relating to F0 trends were carried out, hence reporting the following characteristics (fig. 3): maximum and minimum, mean (and coefficient of variation), and range of values.

Under conditions of equal total average F0 values, spontaneous speech generally exhibits greater dynamics, as indicated by coefficient of variation. This feature reflects the greater extension of both minimum and maximum values. This special feature of spontaneous as opposed to read speech may be due to the greater number of phonetic chains with differing relative intonation boundaries, focusing phenomena, and paralinguistic, i.e. emotional and attitudinal, characteristics.

Percentages of OCCURENCE of PAUSES based on Number of constituents							
		SPK.1		SPK.2		SPK.3	
		READ	SPON	READ	SPON	READ	SPON
%	One-el. pauses	10		36	6.4	53.8	6.7
%	Two-el. pauses	60	23.1	60	51.6	38.5	26.7
%	Three-el. pauses	30	53.8	4	29.1	7.7	60
%	Four-el. pauses		23.1		12.9		6.6
%	TOTALS	100	100	100	100	100	100

Fig. 5: see text for explanations

3.3. Frequency and typology of pauses

Figure 4 provides a general view of the different pausing strategies adopted (percentage values), broken down per speakers and speech styles (including vowel lengthening at the initial and final positions of phonetic chains).

Figure 5 provides a summary classification based on the number of the elements characterising the

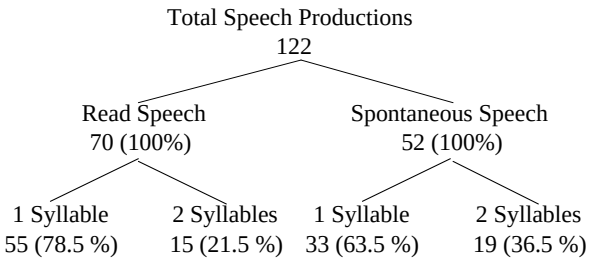


Fig. 6a: see text for explanations

various pause strategies.
The data illustrated in figures 4 and 5 show that the most complex pauses (i.e. those composed of four elements) were produced only during spontaneous speech, though with differing percentage scores for each speaker. In read speech, pauses consisting of either one or two elements were the most common.

3.4. Classification of results of vowel adjacency

Figures 6a and 6b provide a percentage distribution of the various different results from vowel adjacency within each single phonetic chain. With respect to the target vocalic sequences these results have been classified as dipthongization, deletion of one of the two (equal or different segments), *sinaloephe* of the two original segments into a new one and *hiatus* (maintaining of the two original segments). It is evident that only in this latter case the two separate syllables are actually maintained, whereas in all other cases there is a reduction to a single syllable. The distinction between hiatus and dipthongue was made in the presence of any vowel lenghtening or glottal stop between the two vowels, or if not, by applying the Lehiste and Peterson criteria [16]. Results show that occurrences of vowel adjacency are more frequent in read speech than in spontaneous speech (Fig. 6a). In both styles there is a general tendency to favour the formation of a single syllable from the two initial syllables, whereas the two speaking styles differ with respect to the presence of specific phenomena (Fig. 6b).

	Read Speech 100%	Spont. Speech 100%
Deletion of 1 of 2 different vowels %	50.9	54.5
Deletion of 1 of 2 equal vowels %	12.8	0.0
Sinaloephe %	21.8	30.3
Dipthongization %	14.5	15.2

Fig. 6b: see text for explanations

CONCLUSIONS

The parameters chosen for our study (i.e. fluency indices, F0 dynamics, pause composition, vowel adjacency results) provided good distinction between spontaneous and read speech. Spontaneous speech trials were less fluent, due to the presence of filled pauses and complex pauses, whereas it was typically characterised by a greater dynamic trend of F0, not only in the total texts but also in the production of single phonetic chains. With regard to vowel adjacency results, which turned out to be in greater number in read speech than in spontaneous speech, in both styles there was a general tendency to favour the formation of a single syllable from the two initial syllables.
Further research will have to evaluate trends of F0 boundaries, analyse the properties of F0 in full pauses and identify the prosodic and syntactic condition that regulate the location of pauses and vowel lengthenings, as well as the constraints underlying syllabic simplification phenomena.

REFERENCES

[1] M. Eskenazi, "*Trends in Speaking Styles Research*", Proc. EUROSPEECH'93,Vol. 1 , pp. 501-509, Berlin, 1993.
[2] W.J. Barry, "*Phonetics and Phonology of Speaking styles*", Proc. ICPhS'95, vol.2, pp. 4-10, Stockholm, 1995.
[3] K.J. Kohler, "*Articulatory Reduction in Different Speaking Styles*", Proc. ICPhS'95, vol. 2, pp. 12-19, Stockholm, 1995.
[4] F. Goldman-Eisler (1968), "Psycholinguistics. Experiments in Spontaneous Speech", Academic Press, London, 1968.
[5] F.Grosjean and A.Deschamps, "Analyse contrastive des variables temporelles de l'anglais et du francais: vitesse de parole et variables composantes, phénomènes d'hesitation", *Phonetica*, 31, pp. 144-184, 1975.
[6] J. Hirschberg, "*Prosodic and Other Acoustic Cues to Speaking Style in Spontaneous and Read Speech*", Proc. ICPhS'95, vol. 2, pp. 36-43, Stockholm, 1995.
[7] Y.Sagisaka, N.Campbell, N.Higuchi (Eds.), "Computing prosody Computational Models for Processing Spontaneous Speech", Springer-Verlag, New York, 1997.
[8] W.N. Campbell, "*From Read Speech to Real Speech*", Proc. ICPhS'95, Stockholm, vol. 2, 20-27, 1995.
[9] N.A.Daly-Kelly, "*Linguistic and Acoustic Characteristics of Pause Intervals in Spontaneous Speech*", Proc. EUROSPEECH'95, vol.2, pp. 1023-1026, Madrid, 1995.
[10] A.Butcher, "Aspects of The Speech Pause: Phonetic Correlates and Communicative Function", Arbeitsberichte 15, Institut für Phonetik, Universität Kiel, Kiel, 1981.
[11] M. Eskenazi, "*Changing Speech Styles: Strategies in Read Speech and Casual and Careful Spontaneous Speech*", Proc. ICSLP'92, vol. 1, pp. 755-758, Edmonton, Alberta, 1992.
[12] G. Bruce, "*Modelling Swedish Intonation for Read and Spontaneous Speech*", Proc. ICPhS'95, vol. 2, pp. 28-35, Stockholm, 1995.
[13] M.E. Van Donzel & F.J. Koopmans-Van Beinum, "*Pausing Strategies in Discourse in Dutch*", Proc. ICSLP'96, Vol. 2 , pp. 1029-1032, Philadelphia, 1996.
[14] P.L.Salza, G. Marotta, D.Ricca, "*Duration and Formant Frequencies of Italian Bivocalic Sequences*", Proc. ICPhS'87, Tallinn, vol. 3, 113-116, 1987.
[15] D. Duez, "Silent and Non-silent Pauses in Three Speech Styles", *Language and Speech*, 25, pp. 11-28, 1982.
[16] I. Lehiste and G.E. Peterson, "Transitions, Glides, and Diphthongs", *J.Acoust.Soc.Am.*, 32, pp. 693-702, 1960.