Prof. Luciano Romito, Laboratorio di Fonetica, Dipartimento di Linguistica, Università della Calabria. Ponte attrezzato Pietro Bucci nr.20 A, Arcavacata di Rende 87036 Cosenza Italia. Tel. 0984494078; 0984494097, Fax 0984494114, E-mail luciano.romito@unical.it

Dott.ssa Maria Blefari, Laboratorio di Fonetica, Dipartimento di Linguistica, Università della Calabria. Ponte attrezzato Pietro Bucci nr.20 A, Arcavacata di Rende 87036 Cosenza Italia. Tel. 0984494078; Fax 0984494114, E-mail *****

Title: Towards a new parameter in the Speaker Recognition

Nr words: ****

Bytes: *****

Inserire un curricul professionale 50-75 words.

Towards a new parameter in the Speaker Recognition

Abstract

The subject of this study is the two indexes of fluency Articulation and Speech Rate.

After a brief review of confused definitions in linguistics literature , we have calculated the index of samples of voices of spontaneous speech using speakers with the same diatopic and diastratic features.

Making a comparison with data from other laboratories, we suggest that there is not a homogeneous method for this type of internal analysis . Lastly, we have evaluated the importance of both indexes as parameters to use in Speaker Recognition, by means of the verification of the identity of a speaker in a closed-set of known voices.

Keywords: Speaker Recognition, Articulation Rate, Speech Rate, fluency, parameter.

Introduction

Over the years, many and various solutions have been proposed for the problem of the identification of offenders.

The human capacity to recognize, in daily life, the voice of a known person, has led to the belief that vocal emission can have a possible application in judicial enquiries.

On account of its variable nature, the voice would not appear to be suitable for such a complex field.

However, scientific knowledge of phonetic parameters, that depend on the speaker (such as the fundamental frequency F_0, the vowel formants' frequency, the resonance of nasal consonants F_0, F_{N1}, F_{N2}, and the Speech Rate), allows for the use of vocal emission for speaker recognition in forensics.

In this field, the term Speaker Recognition is generally used to cover the variety of procedures adopted for the identification, or the verification of identity, of a speaker by means of the voice.

The attribution of a voice to a speaker, requires analysis, observation, evaluations and comparison of parameters.

Since this study concentrates on the search for new parameters in Speaker recognition, the following characteristic procedures of the semi-automatic method have been adopted:

- choice of acoustic parameter to analyse;
- evaluation of the statistical importance of the considered variable;

- estimation and comparison of samples by means of a statistical test (using the results of previous measurements).

The goal of this study is to verify the reliability of the indexes of fluency Articulation Rate and Speech Rate, through the verification of the identity of a speaker in a closed-set of known voices.

The characteristic of the fluency indexes is that they are *prosodic* phenomena or, more frequently , over segmental, that is, above the segments, because they refer to the entire sequence.

The audio-verbal communication is realized through the production and reception of phonic blocks of varying size (sentence, word groups, single words or fragments of a word) and these blocks, are internally modeled by a certain pitch, by variation of velocity and by position of accents. The over-segmental features of phones differ from the segmental because of their *relative* nature with respect to the surrounding phones: a phone is considered in terms of more or less duration, intensity, volume, not in absolute, terms but always in reference to the other sequence of phones in which it finds itself.

The Articulation Rate depends on the intrinsic duration of the various phones, on the rate of articulation movement and on the rules of co-articulation.

The Speech Rate depends on specific speaker features, on the communicative situation (diatopic variable), but it varies also in a phrase, with accelerations and contractions, which indicate, among other things, the degree of attention that

the speaker asks of the listener (speed for less attention, slowness for more attention).

Furthermore, the rhythm of speech, in a similar situation, depends on the phonetics alphabet used[i].

This study elucidated below is divided into two steps:

1. Study of reliability of the index of fluency Articulation Rate;
2. Study of reliability of the index of fluency Speech Rate.

Articulation and Speech Rate: some definitions

First, it is important to stabilize, according to the index definition, what parameter is necessary to consider and how calculate it.

By means of a brief review of the relevant linguistics literature, it is possible to observe how, over the years, there have been various attempt to define the index of fluency Articulation Rate and Speech Rate.

For example, for the Articulation Rate, there is:

- *«La vitesse d'articulation. Véritable vitesse de phonation, quisqu'on lui retrace le temps de pauses, elle est exprimée en nombre de syllabe/sec et s'obtient en divisant le nombre de syllabes émises par le temps d'articulation du locateur»* (Grosjaen and Deschamps 1975: 148);
- "*The articulation rate of each group('utterance') was computed by dividing the total number of syllables in the group by the cumulative duration of the runs comprising the group (excluding any pause time)*", where *run* is "*the*

stretch of speech that contains no pauses, with a pause defined as a silent interval of 250 ms or greater" (the calculus of Articulation Rate is made on a group of *runs* with a minimum of 30 syllables, Miller *et al.* 1984: 218-219);

- "*il rapporto tra il numero delle sillabe e la durata della catena fonica*[ii]"(Sorianello 1996:95);
- "*il numero delle sillabe diviso per la durata della catena fonica*[iii]" (Magno Caldognetto and Vagges 1993:101 recalling Duez 1982);
- "*espresso come il numero di sillabe al secondo della sequenza articolata*[iv], *risultante dalla formula: numero delle sillabe della sequenza articolata/durata della sequenza articolata*" (Zmarich *et al.* 1996:120);
- "*average number of (phonetic) syllables per second of the articulation phase of speech*[v] *(number of syllable/[duration- combined duration of all pause])*" (Künzel 1997:1358);
- '*la velocità di articolazione è data dal rapporto tra il numero delle sillabe e la durata delle catene foniche. Nel computo delle sillabe vengono di norma inclusi tutti quei fenomeni udibili quali pause piene e prolungamenti vocalici*." (Giannini 2000:253);
- "*equivale al rapporto tra il numero delle sillabe realmente pronunciate e il tempo impiegato per realizzarle*" (Pettorino 2003:228).

The first difficulties that we encounter in the course of this brief, but important review of definitions are those related to the different considerations regarding

what parts of the signal to examine: temps d'articulation du locateur, run, phonic chain, articulated sequence, the articulation phase of speech, the time used to realize the syllable and the different definitions of articulated sequence and phonetic chain. Moreover, the different definitions of silent pauses[vi] and the various limits given when silence occurs followed by an occlusive phone[vii].

Some definitions of Speech Rate are:

- "*il rapporto tra il numero delle sillabe e la durata dell'enunciato*" (Sorianello 1996:95);
- *"Speech Rate: espresso come numero di sillabe al secondo della sequenza articolata, in relazione alla durata della catena fonica comprensiva di esitazioni e disfluenze, risultante dal rapporto tra il numero delle sillabe della sequenza articolata e la durata dell'intera catena fonica"* (Zmarich *et al.* 1996:120);
- "*rapporto tra il numero di sillabe e il tempo totale dell'enunciato*"(Pettorino 2003:228).

In general, the result is a certain degree of confusion.

In this study to calculate the Articulation Rate we refer to Zmarich *et al.* (1996:120), while to calculate the Speech Rate we used the definition of these writers with some changes.

To the definition of phonetic chain, we add here the definition of "breathing group", that is, the portion of signal between one inspiration and another.

Breathing is fundamentally composed by two phases: inspiration and expiration.

If the second phase is realized during the production of the sequence, the first can be realized as an isolated inspiration, that is audible and visible on the spectrogram as a zone of noise at low strength, frequently with traces of formants, or through the execution of some phones, above all vocalic, that can be recognized by perception and acoustically, from the progress of the energy curve.

To calculate the Speech Rate index we used the formula: number of syllables of phonetic chain/duration of phonetic chain, expressed in syll/sec.

In this way the definition of Speech Rate is the same as that given by Sorianello (1996) and Giannini (2000).

Materials and Methods

The material used are audio signals taken from the FOCUS[viii] database. The corpus that we have used is composed of recordings in PSTN (telephony network) of spontaneous speech, passages of structured dialogue elicited in an interview, and by repetitions of the same dialogue by the same speakers.

Our having defined as spontaneous speech the samples analyzed, could be criticized, because they are "structured" interviews, and for this reason constitute semi-spontaneous speech.

The considerations and the results of this study, however, do not change. In effect, a study conducted by Künzel (1997), on ten speakers (five male and five female), verified that the differences between spontaneous and semi-spontaneous speech are so small as not to be relevant[ix].

The database includes recordings of a group of speakers, a homogeneous sample from the point of view of geographical origin, dialect, age, degree of instruction, work: all features that exclude possible diatopic and diastratic perturbations.

The segmentation[x] of signals was made by means of the visualization of three windows: of wave form, sonogram, energy curve and with the help of auditory feed-back.

To correctly identify syllabic boundaries, it is essential to know how the phones are represented on the sonogram.

The computation of syllable was carried out by individualizing the phonetic syllable[xi] and the relative rules of syllabification[xii].

The characteristic fragmentation of on-line planning for spontaneous speech is linked to the realisation of articulated sequences and/or phonetic chains of differing length.

We have considered only the articulated sequences and the phonetic chains with a number of syllables superior or equal ten, to remove the possibility of misleading data.

We have not established limits on the duration of pauses.

Data analysis

Articulation Rate

The index of Articulation Rate calculated for each sample of voice and for the articulated sequences that we have considered are presented in Table 1.

ARTICULATION RATE																					**μ**	**σ**	**n**
A	6,0	6,4	6,3	6,2	6,2	6,9	5,9	5,3	8,8	5,7	6,3	6,6	4,7	5,8	5,8	6,2	6,4	5,0			6,1	0,9	18
B	6,2	5,6	6,2	7,2	4,7	6,1	7,2	6,0	6,7	5,0	4,5	5,0	6,4	5,2							5,9	0,9	14
C	4,3	5,5	5,3	4,0	5,0	5,0	6,5	4,4	5,8	4,9	5,3	5,7	3,7	6,4	5,5	5,3					5,2	0,8	16
D	6,9	6,6	6,1	8,5	6,0	5,3	6,9	6,9	6,7	7,4	7,5										6,8	0,8	11
E	4,9	6,5	5,9	5,4	7,8	6,6	6,0	5,9	6,8	6,0	7,3	5,9	6,8	6,4	5,7	6,9	7,2	6,4			6,4	0,7	18
F	4,6	6,0	4,6	4,4	6,3	4,2	5,1	6,6	6,9	5,7	6,3	5,6	8,1	5,1	4,4	6,9					5,7	1,1	16
G	5,5	4,9	5,9	6,7	5,8	6,0	8,2	6,2	7,0	6,8	8,1	5,9	6,6	4,3							6,3	1,1	14
H	7,0	6,4	7,2	7,3	5,6	6,4	5,8	6,2	5,9	6,6	5,5	4,9	8,3	7,0	6,3	6,4	7,0				6,5	0,8	17
I	6,0	4,9	5,9	6,6	5,9	6,9	4,8	5,4	5,0	7,6	6,1	7,0	5,9	6,0	3,6	7,8	6,5				6,0	1,1	17
L	6,3	6,6	6,7	6,5	6,5	8,2	8,2	8,4	7,3	7,8											7,2	0,8	10
M	5,2	4,4	6,5	4,4	5,5	7,9	6,5	6,6	6,2	8,8	7,1	7,7	7,5	7,6	6,4	6,7	4,6	5,3	7,7	5,3	6,4	1,3	20
N	5,6	6,0	6,7	4,9	4,8	4,9	5,6	5,7	6,2	5,7	5,4	6,1	5,2	5,4	6,4	5,2	6,0				5,6	0,5	17
μ: Mean Speech Rate mean for each sample; **σ:**Speech Rate's deviation standard for each sample; **n:** number of values																							

Table 1 Index of Articulation Rate, calculated for each articulated sequence

Comparing the results that we have, with others from Mori *et al.* (2004) there are some differences because of different methods, and in particular, the difference in having considered "only the articulated sequences" that are not

characterized by the presence of voiced pauses and segmental breathing and that are composed of a minimum of six syllables" (Mori and Paoloni 2004).

We report the comparison in table [xiii].

Speaker	Experiment	Mori-Paoloni
A	6,1	7,7
B	5,9	7,97
C	5,17	6,66
D	6,8	7,3
E	6,4	7,28
F	5,7	6,57

Speaker	Experiment	Mori-Paoloni
G	6,3	7,42
H	6	7,35
I	5,98	7,64
L	7,25	8,2
M	6,4	7,71
N	5,6	6,89

Table 2 Comparison of Mean Articulation Rate for each speaker

This difference in results is due not only to the difficulty in making an internal analysis but also the use of different methods of analyse. This is the cause of difficulty in obtaining homogeneous measures in different laboratories, with the analysis carried out by different operators.

We have used the database to calculate the Articulation Rate, to observe the reliability of this index as a parameter in Speaker Recognition, using the samples A, B and C.

We have applied the statistical test (test-t of student), and then calculated the dissimilarity percentages, which provide the results in Table 3 for each comparison.

A-B	A-C	B-C
62,704%	99,846%	97,012%

Table 3 Dissimilarity percentages of Articulation Rate for comparisons A-B, A-C and B-C

There are middle-high percentages; which do not permit us to state that Articulation Rate, thus understood, can be utilized as a parameter in Speaker Recognition.

On the other hand, similar result are obtained in another study (romito and Blefari 2004), using samples from the same database. In this study the authors suggest that the comparison of the index of Speech Rate of voice samples from the same linguistic community, results in bigger dissimilarity percentages, among the fluency indexes considered.

Our intention therefore, is to analyze the index of fluency Speech Rate and investigate its reliability in Speaker Recognition.

Speech Rate

To calculate the index of Speech Rate for the samples A, B and C we have used the method described previously.

The computation of syllables, thanks to the help of the spectrograph, had taken into account all of the features of spontaneous speech[xiv]: the syllables of the so - called non-silent pauses[xv], vowel germination[xvi], the cancellations and the diphthongization[xvii] and the fall of syllables or parts of them.

If the goal of this study is to individualize a temporal parameter that characterizes the speaker and that describes production using the syllable number and temporal length, it is necessary to consider all of the phonetic production, including the syllables of non-silent pauses.

This choice is also suggested by results in Duez (1982)[xviii]: for this author, the phenomenon of hesitation is a specific speaker feature, above all in the tendency to use, or not use full pauses.

The index of Speech Rate for the three voice samples, A, B and C is in the Table 4.

	Speech Rate (sill/sec)																													μ	σ	n
A	5,8	7,8	6,3	6,1	6,1	7,5	6,5	5,2	7,4	5,0	5,0	5,5	6,6	7,5	6,8	9,1	6,7	8,7	4,9	7,7	5,0	4,5	5,5	5,1	6,6	7,0	6,4	3,6		6,3	1,3	28
B	5,8	4,6	6,8	6,1	3,9	5,6	4,6	6,6	6,0	5,3	4,1	4,9	5,3	6,7	5,0	4,0	4,7	4,6	4,0	4,0	6,1									5,2	0,9	21
C	3,9	6,5	4,4	2,6	5,7	3,6	5,2	3,6	8,1	3,3	4,8	5,7	3,7	4,4	3,4	5,1	3,3	5,5	3,8	4,9	4,7									4,6	1,2	21
μ: Mean Speech Rate mean for each sample; **σ:**Speech Rate's deviation standard for each sample; **n:** number of values																																

Table 4 Index of Speech Rate of samples A, B and C

The dissimilarity percentages for the comparison are:

A-B	A-C	B-C
99,872%	99,996%	90,552%

Table 5 Speech Rate Dissimilarity percentages: samples A, B and C

The fact that the analysis gives such high percentages, suggests that the method used for distinguishing voice samples is suitable for use in Speaker Recognition.

We have also studied to see if the same method is good for seeking similariy in samples from the same speaker, using the recordings of repetitions of A, B and C, named A', B' and C'.

The Speech Rate index for these samples is presented in Table 6.

Speech Rate (sill/sec)																							μ	σ	n
A'	5,9	3,9	7,9	6,0	7,7	7,0	8,0	7,8	4,7	4,5	5,6	4,8	4,8	5,8	7,0	6,0	7,3						6,2	1,3	17
B'	5,4	5,2	4,9	7,1	4,9	5,8	7,9	4,1	7,6	4,1	4,4	4,8	5,9	8,2	7,0	5,3	6,2	6,1	7,1	6,1	4,4	8,1	5,9	1,3	22
C'	5,0	3,8	5,0	5,2	4,3	5,2	3,7	6,7	5,0	5,3	3,6	5,6											4,9	0,9	12
μ: Mean Speech Rate mean for each sample; **σ:**Speech Rate's deviation standard for each sample; **n:** number of values																									

Table 6 Index of Speech Rate of samples A', B' and C'.

The application of the similarity test gives percentages that are not really sufficiently reliable to enable us to say that the used method can be applied to similar voice samples for the same speaker, as is shown in Table 7.

A-A'	B-B'	C-C'
78%	4%	47%

Table 7 Similarity percentages for comparisons A-A', B-B', C-C'

Conclusions

At he end of this study we can derive, on the basis of a careful analysis of the used methods and of the obtained results, some important considerations.

- For the Articulation Rate we have suggested that the different results, of different laboratories and operators, are due to the lack of a homogeneous definition and analysis method.
- For Speech Rate we indicate the mean production of syllables for each phonetic chain, where for phonetic chain we mean the sequence of phonetic segments (including the non-silent pauses) bounded by two silent pauses, and / or portions of signal included between one inspiration and another.
- To analyse the importance of the index of fluency, Speech Rate, as a parameter in Speaker Recognition, it is necessary to consider the whole phonetic realization of a speaker and make statistical comparisons with samples of voice of the same, and / or of different speakers, to support the used method. The need to consider the speaker's entire phonetic realization, means identifying the phonetic syllables that build the segmental sequence (including the non-silent pauses).
- The index calculated using these parameters gives high dissimilarity percentages, which permit us to say that the method is good for differentiating samples of spontaneous speech of speakers with the same diatopic and diastratic features.
- The indisputable degree of subjectivity to which the method is subject : in the individuation of the parameter, in the ability of the operator to identify the vowel germination and the phonetic chain.

[i] Studies on three different variety (Bari, Napoli and Pisa) demonstrate that the production of syllables in less time isn't a voluntary strategy, but it depends on the different phonetics alphabets, (Pettorino 2003).

[ii] Here the writer uses the definition of phonetic chain by Pettorino and Giannini 1994: "La catena fonica è la porzione di un enunciato compresa tra due pause vuote".

[iii] The phonetic chain is for the authors: "Total articulation time: durata globale della produzione verbale del soggetto, costituita da catene foniche e dai silenzi"

[iv] That is the phonetic chain without the non silent pauses.

[v] In this case it is the duration of task of each speaker, equivalent to half a minute.

[vi] Miller *et al* **********:"a silent interval of 250 ms or greater"; Sorianello **********"una momentanea sospensione dell'attività fonatoria e conseguente assenza di rumore spettrografico, di almeno 100 ms nel parlato spontaneo", Künzel ********* give the limit 100 ms.

[vii] Magno Caldognetto and Vagges ********* give 100 ms to the fase of consonantic occlusion, whilst Zmarich *et al.*, give 50 ms.

[viii] Focus (FOrensic CorpUS), is a corpus of samples of similar voices, realized to study Speaker Recognition, a database used to analyse and compare results, collected in the ISCTI laboratory, cfr Falcone and Barone (on line).

[ix] "[…].Results show that differences in tempo-related parameters are found almost exclusively between spontaneous and semi-spontaneous samples on the one hand and read samples on the other. There are hardly any differences between the first two speaking conditions. ", H.Künzel 1997:78.

[x] A phonetic transcription with informations on the temporal setting of unit boundaries, cfr. Salza 1990:24-25.

[xi] The notion of syllable has been object of study for a long time because of its phonetic and phonological "double nature"; in this study we refer respectively to: "la struttura elementare che sta alla base d'ogni raggruppamento di fonemi"(cfr. Jakobson 1974:94.) e "l'unità prosodica costituita da uno o più foni, agglomerati intorno ad un picco d'intensità"(cfr. Leoni and Maturi 1998: 74-5).

[xii] Differences between phonetic and phonological syllabification are not so great, only in the case with the group /s/ and /z/ + consonantal or liquid phoneme (for example "*festa*" is phonetically divided in *fɛ's-ta* and phonologically in*fɛ́-sta),* and some connections, not autochthonous, such as /tm/ and /tl/ (for example 'atmosfera, atleta'), cfr Muljačič 1969: 471.

[xiii] The comparison was made on the mean Articulation Rate for each speaker.

[xiv] Spontaneous speech is characterized by a tendency towards the reduction of articulatory effort, causing phenomena of assimilations, centralizations, deaccentuations, elisions, and, from the point of view acoustic, medium shortening of all the phones, fall of phones and/or syllable tendency to omit segments of the sequence, cfr. Kohler, 1995, op. cit. in Zmarich *et al.* 1997.

[xv] "Pause piene: esitazioni, interiezioni, allungamenti di vocale, disfluenze, ecc": Zmarich *et al.* 1997.

[xvi] The spectroacustical analysis of some vowel "length" at the end of a word or in hesitations, suggests the production of two, or more, phonetic syllables.

[xvii] Because of vowel meeting and/or adjacent words.

[xviii] Duez D. (1982): in a study on three style of speech in French.

References

Duez D. (1982), Silent and non-silent pauses in three speech styles, *Language and Speech*, 3:179-192.

Falcone M., Barone A, *F O C U S: un corpus vocale di voci simili per lo studio della identificazione del parlatore in ambito forense*.(on-line).

Grosjaen F., Deschamps A. (1975), Analyse contrastée des variables temporelles des l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation, in *Phonetica*, 31:144-184.

Jakobson R. (1974), *Saggi di linguistica generale*, a cura di L.Helmann, Milano, Feltrinelli.

Künzel H. J. (1997), Some general phonetic and forensic aspect of speaking tempo, in *Forensic Linguistics*, 4(1):1350-1771.

Leoni F.A., Maturi P. (1998), *Manuale di fonetica*, Roma, Carocci.

Magno Caldognetto E., Vagges K. (1993), Le pause quali indici diagnostici per lo stile del parlato spontaneo, in *Atti delle 2^{e} Giornate di Studio del G.F.S.*, Calabria, 28-29 novembre 1991, 19:97-106.

Miller J.L., Grosjaen F., Lomanto C. (1984), Articulation Rate and its variability in spontaneous speech: a reanalysis and some implication, in *Phonetica*, 41: 215-225.

Mori L., Paoloni A., (2004), Sulla sociolinguistica forense:la costituzione di corpora vocali per l'analisi della velocità di articolazione in italiano, in *Atti delle XIV Giornate del GFS, Viterbo, 4-6 Dicembre 2003.*

Muljačič Z. (1969), *Fonologia generale e fonologia della lingua italiana*, Bologna, il Mulino.

Pettorino M. (2003), Caratteristiche prosodiche dell'italiano dialogico in *Voce Canto Parlato, studi in onore di F.Ferrero*, a cura di P. Cosi, E. Magno Caldognetto, A. Zamboni, Unipress, Padova, pp.227-230.

Romito L. Blefari M., (2003) **TITOLO ********** unpublished thesis, University of Calabria.

Salza P. L. (1990), La problematica della segmentazione del segnale vocale, in *Atti delle 1^{a} Giornata di Studio del G.F.S*, Padova, 3-6-novembre.

Sorianello P. (1996), Dal parlato letto al parlato spontaneo: indici prosodici a confronto,in *Atti delle 7^{e} Giornate di Studio del G.F.S*, Napoli, 14-15 novembre, 7:89-110.

Zmarich C., Magno Caldognettto E., Ferrero F. (1997), Analisi confrontativa di parlato spontaneo e letto: fenomeni macroprosodici e indici di fluenza, in *Quaderni del CNR*, 16:266-290.