In: M. González González, Elisa Fernández Rei, Begoña González Rei (eds.), *Actas do III Congreso da Sociedade Española de Acústica Forense* (Santiago de Compostela, 27-28 Out. 2005), Santiago de Compostela: Xunta de Galicia, 339-351.

ACOUSTIC CUES FOR VOICE CHARACTERIZATION: AN EXPERIMENT ON WORDS IN ISOLATION UTTERED BY A SAMPLE OF TURIN STUDENTS

ANTONIO ROMANO
Dip. di Scienze del Linguaggio - Università di Torino, Italia

INTRODUCTION[1]

This work is the result of a brand new attempt (at least in Italy) to introduce in the University the discussion about the recourse to acoustic methods as a means for voice identification technique in courtrooms.

As in previous similar experiments abroad, the Faculty of Foreign Languages of Turin started up a training course (consecrated to a selection of students with a satisfactory linguistic background) whose duration was of about one month. During the following three months the students tested their skills in manipulating phonetic data by means of raw acoustical and statistical analyses.

The experiment (nowadays we are at its third edition) had recourse to an acoustic *DB* collecting the linguistic productions of the students (all coming from the Piedmont area) and carrying out formant analyses and time measu-

[1] This work has been planned with the help of three *PhD* students. The first assessments on the original acoustic *DB* have been discussed in ROMANO-MANCO-TOMATIS (2003). Analyses are also partially based on further research carried out by Lucia Nicoletti who also provided an early English version of the present text.

rements as in some Italian voice identification expertises. The analysed data come from their realisations of a corpus of 150 words uttered in isolation. The results, which are here discussed, are based in particular to a selection of ten feminine voices. Such selection is quite consistent – speaker of the same age and origin, and same conditions of recording were used.

The *DB*'s sounds were collected and analysed in an equipped classroom were the students had the opportunity to make their listening tests and measurements on independent personal computers[2].

The measurements applied to the duration of consonants and vowels and to the formant values in the central and most stable part of stressed vowels. Attention was particularly drawn to the importance of distinguishing the different phases of closure and release for plosives and affricates and to the relationship between vowel lengths in open vs. closed syllables[3].

SOME GENERAL RESULTS

The principal observations which can be made on the words selected and analysed are related to their general structure: mainly vocalic and consonantal length -in accented/unaccented positions, in open/closed syllables-, formant patterns and accent position.

The issue of the temporal aspects in the production of vowels and consonants, which has already been dealt with in previous field studies, is here reconsidered and verified. Moreover, the most significant contribution this work can offer - with regard to phenomena occurring in the read-spoken language - refers primarily to the duration relationships amongst segments in prominent/non-prominent positions, and to the verification of the presence of a (iso-)syllabic effect for the vowels in open vs. closed syllables.

[2] Recordings were made directly on *PC* with a *Soundblaster*™ card and a *SONY*™ *ECM 907* microphone. Utterances were digitized in *.wav* format following a 16kHz/16bits *PCM* coding. Measurements were originally performed with *CoolEdit* (*demo version* 1.52), *Wavesurfer* and finally with *PRAAT* (v. 3.8.64 and later 4.1.25). The *PC*s were provided with *SPSS*™ for statistical analyses.

[3] All the measurements have been assessed by two trained phoneticians (formant values were affected by error rates between 2 and 13%). As for durations, an important instruction was to not include drawls in the length estimate of final vowels, bounding it to the end of the visible F_2 track. For more details and references see ROMANO *et alii* (2004).

As for the possibility of distinguishing speakers on the base of a different timing of closure and release for plosives and affricates, we have discussed our data in ROMANO *et alii* (2004). Even though these cues could help in a large scale comparison between regional varieties, we found a general agreement within our speakers' group[4].

A general agreement also applied to vowel lengths: 1014 stressed vowels have been measured against 385 unstressed (not final) vowels.

In particular, we found a stable contrast between stressed and unstressed vowels (not depending on vowel quality - differences are not significant; see fig. 1). Mean durations of stressed vowels vary between 172 and 201 ms whereas unstressed vowels show mean values spanning from 82 to 95 ms.

Another important contrast is maintained in our sample: the length relation between stressed vowels in open vs. closed syllable that describes vowels in open syllable longer than in closed syllables (that leads to a complementary distribution of long vs. short vowels)[5].

Among the 1014 stressed vowels which have been analysed, 460 were in open syllables and 554 in closed syllables.

The phenomenon was systematic for each speaker and, on the whole, for all the words showing similar contrasts. But, as shown in fig. 2, probably due to intraspeaker variability, only a weak statistical significance supports that claim. Nevertheless, open syllable vowels present higher mean values (between 195 and 218 ms), whereas closed syllable vowels are generally shorter (156÷178 ms). That determines an average length ratio between vowels in closed and open syllables of about 0.80[6].

[4] Other region-dependent rules not enough well studied are related to strategies of reducing consonant length from post-stressed positions to pre-stressed positions (see ROMANO *et alii*, 2004: 244-245).

[5] Since FERRERO (1972), various studies highlighted the presence of this phenomenon in different varieties. Measures are available in FAVA & MAGNO-CALDOGNETTO (1976), BERTINETTO (1981), FARNETANI & KORI (1986), MAROTTA (1995) and -in a different perspective- VOGEL (1982).

[6] This figure places the productions of our sample closer to varieties reducing the contrast (cp. ROMANO, 2003). Despite a number of exceptions, figures for this ratio are usually assessed around 0.60÷0.70 for more standard varieties (and often southern varieties) and towards 0.90 for a number of peripheral varieties.

Like in previous studies comparing speakers from various regions, only a partial confirm has been found for the common claim that Northern Italian presents a generalized lengthening of stressed vowel in closed syllable.

Individual performances also presented similar statistics (in no case *p* raised over 0.20). The length ratio between vowels in closed and open syllables was also surprisingly very stable: it spanned from 0.78 to 0.81.

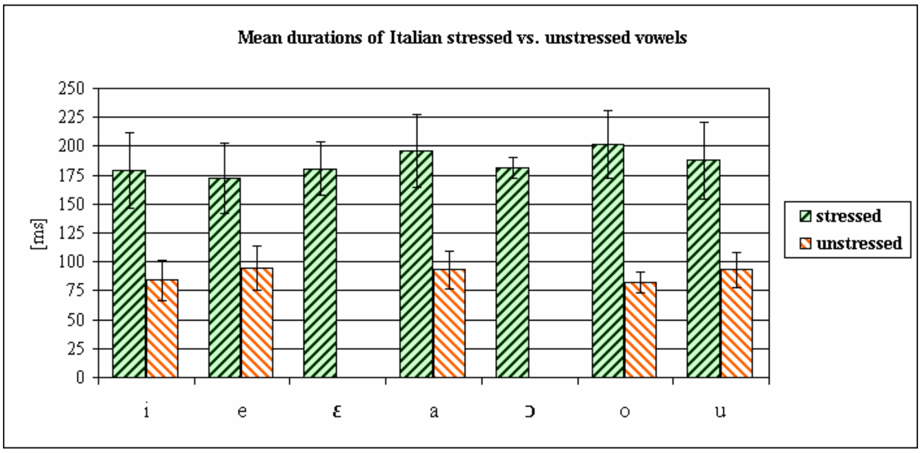

Figure 1. Mean duration of Italian stressed vs. unstressed vowels for all the productions of the ten students (over 1300 items analysed).

In the framework of the parametric approaches used in Italian courtrooms, other important cues -generally claimed to help discriminating between speakers- are formant patterns. Scatter plots and statistics of F_0 and first three formants were assessed on the base of 100 stressed vowels per speaker. We selected three of them who gathered the most similar voices in the sample (even though easy to recognize at listening, after a short training).

On the whole, data on the three diagrams are well overlapping and attest a good consistence among the individual vowel systems. Anyway, at a first glance the filling of the articulatory space of the 5 vocalic areas looks enough different from the three speakers. The first one (BAR) is especially diverging from the other two (above all for a more narrow F_2 data dispersion – then roughly along the front-back axis; notice the position of [i] occurrences; realisations of /o/ and mid-open /ɔ/, not significantly different one each other, are lower than in the other cases). However similar considerations are useless in order to distinguish the other two voices (in spite of differences detectable at listening). No statistical measure on these few data could objectively assess a difference between the two voices (not even in the case of /a/ who has the greater number of occurrences)[7].

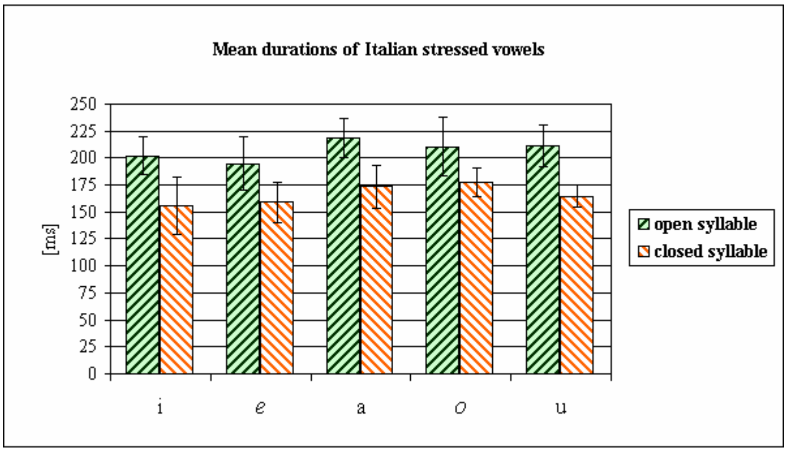

Figure 2. Mean duration of Italian stressed vowels for all the productions of the ten students (over 1000 items analysed). The distinction of the syllable type shows the complementary distribution of long vs. short vowels (Italic *e* and *o* represent the front and back merges of close-mid and open-mid vowels).

Our sample allows us to conclude that formant measures may induce to assume as different the productions of the same speaker (thus frequently leading to what are called cases of false rejection). The comparison above shows instead that statistical distances can be very small for different speakers (thus leading to frequent cases of false detection). Nevertheless we ascertained that in Italy, many voice expertises for legal purposes are still based on such a kind of measures (often carried over on poorer and more degraded samples of connected speech).

[7] 59 [a] for BAR, 40 for STA and 41 for ZAN. A simple *t*-Student test (presuming Gaussian and disjointed the distributions of these variables) may easily prove that the latter two series are statistically not distinguishable (the two samples belong to the same population; *t*=0.314, *p*<0.50 with a degree of freedom not lower than 80). The same test, applied to all the cross-comparisons, gave just three positive results and revealed other two cases of separation only for a single formant. Furthermore an application onto two random samples of formant measures for the same speaker yielded a positive result.

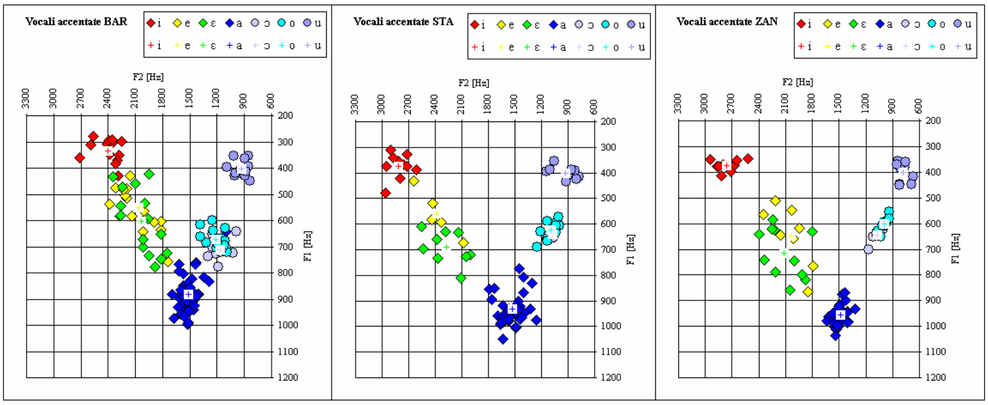

Figure 3. F_1-F_2 scatter plots for vowels of three students (100 items analysed). Very similar vocalic spaces characterise the three voices (The expected distinction of standard Italian between close-mid and open-mid vowels is clearly not present in this variety).

Similar considerations stand of course for duration measures, though they might help in a joint analysis[8].

As other Italian authors are prompting since long ago (see for instance IBBA *et alii*, 1979 and TRUMPER, 1979), our reflections mainly point out that studies in this domain should better take into account geolinguistic features as well as other parameters such as *VOT* and consonant duration and more general voice characteristics not constrained by linguistic structural properties (see, among others, PIAZZA *et alii*, 1998).

The very problem is that, in forensic practice, just simplified formant analyses are carried over without the prerequisite knowledge of vowel systems.

Many “experts” have recourse to Bayesian models, with sophisticated probabilistic tools, but they use to found their statistics on just 10 measures of vowel formants (not distinguishing stressed vowels from unstressed). We had the opportunity to examine a number of expertises where we observed that some measures (which already are usually less than an acceptable number) were based not on real occurrences of vowels but only on graphic

[8] Informal experiments revealed that the acoustic cues we analysed on statistical basis give a poor information as compared to the listeners’ ability in identification tasks of the same voices.

vowels (such as the <i> in “sufficiente”, which is not pronounced in Italian) or on approximant realisations (such as the <i> in “parliamo”)[9].

On the other hand, more theoretical reasons should discourage to use shorthand formant analyses.

Languages are codes that individuals share to communicate one each other. It seems reasonable that a basic conventional agreement (depending on the linguistic system) leads them towards similar choices as for symbolic elements and rules governing the message production. That is an ultimate reason why formant patterning and segmental timing show a considerable degree of convergence within homogeneous speaker groups. So, linguistic models of group can greatly affect individual productions and lead to the neutralisation of differences which could be considered discriminating at a first glance (cp. NOLAN, 1997).

As we think to have shown with our experiment, sets of parameters not properly selected may not determine the correspondence between two voices, and close similarity in some features is sometimes not a convincing argument for itself. This is especially true when such parameters are affected by ‘universal’ phonetic properties.

An indication that could be given instead, in the light of this experience, is that, when attributing specific features to a voice, attention should be focused on those elements possessing higher degree of freedom, rather than on those showing a general convergence to a given model (such as the first formants of a vowel).

SOME REASONS OF DOUBT

As far as we have observed in the Italian courtroom procedures for instance, in the case of a threat message intercepted through a telephone call, “experts” deal with a recording made on the linguistic production of a suspect who recites the same words of the original message. The examiner then run a spectrographic and statistical analysis on both recordings and compare the ‘voiceprints’.

[9] That rises a sensible doubt, at least for a trained professional, on the position where the measure has been performed. Other shocking details about the Italian situation are published in ROMITO (2000).

As we stated above, what is unacceptable is that usually these experts - which are sometimes very skillful in engineering methods - have generally a poor linguistic background.

Dealing with *a priori* (or *a posteriori*) guiltiness probability of a person (a matter that should be revisited in an ethic perspective[10]) by ignoring the complexity (the articulatory constraints which bound the sound production in the vowel space and the structural constraint governing the articulation) is simply a matter of straight roughness.

The current attitude to concentrate the attention on - let us say - "likelihood ratios" evaluated on what are considered merely raw data, without taking into account the linguistic variables and without knowing the possibility for e.g. an Italian speaker to use different vowel systems on the base of geo- and socio-linguistic criteria is definitely a heavy loss (that is the reason why some researchers are now working on that, see MORI & PAOLONI, 2004).

The statistical implementations we can observe, at least in some cases, are not counterbalanced by an account of these aspects that is fine-grained at the same extent.

Statistical models used in ordinary expertises, include Bayesian methods and all kind of cross-correlations and multivariate analyses, but ignore dialectological principles and basic phonetic properties (as for instance that the general Italian vowel system has seven vowel qualities in stressed position and not five as the spelling hints and as only a limited number of regional systems has)[11].

Furthermore, given the objective difficulties, this kind of practice generally neglects to estimate the probability for a voice to be imitated[12] and several other sources of variability such as disguise, signal deterioration and so on[13].

[10] As suggested abroad; see BRAUN & KÜNZEL (1998) and BOË *et alii* (1999).

[11] This crucial point has been finally acknowledged in a recent review (see PAOLONI, 2003).

[12] About this critical topic see for instance ERIKSSON & WRETLING (1997) and PAOLONI & PETTORINO (2003).

[13] A common form of disguising a voice is using a whispery or distorted voice. Another problem rises from the fact that some voices, especially those of family members, may be very similar and easy to confuse. Eminent phoneticians have doubts about the reliability of these methods. Spectrographic analysis and parametric assessments, while accurate if carried over by trained phoneticians under ideal laboratory conditions, are not reliable enough when recordings are degraded by background noise, telephone encoding and general poor quality, and are compared without a suitable set of linguistic precautions.

Even if nowadays, after extensive research and experimentation, the common techniques are considered as extremely reliable, we go back to the early position of many real experts testifying against the use of instrumental voice identification evidence in courtrooms. In our sceptical view, generally sharing the same spirit of BOË *et alii* (1999), they are still applied in a simplified way by people without the prerequisite linguistic knowledge and without a satisfactory training in phonetics.

CONCLUSIONS

The results, which are here discussed, are based to a consistent selection of ten feminine voices sharing several sensible linguistic properties.

They have been analysed and discussed in a training course at the Faculty of Foreign Languages of Turin in order to let students test their skills in manipulating phonetic data by means of raw acoustical and statistical analyses. The analysed variables were temporal and spectral characteristics of a corpus of 150 words uttered in isolation by ten speakers, in laboratory conditions.

With our results we have shown, on the one hand, the relevance of temporal parameters in the socio-geolinguistic characterization of speakers but, on the other hand, the difficulties to use similar parameters for voice identification within homogeneous speakers groups.

Most of the acoustic parameters observed in this study are correlated to linguistic features which show a considerable degree of convergence for speakers belonging to homogeneous linguistic groups. That sometimes includes voice settings and other paralinguistic or extralinguistic variables. Specific models can greatly affect individual productions and lead to the neutralisation of many elements, which are often considered discriminating for different voices.

In these conditions, close similarity in some features is sometimes not a convincing argument and sets of parameters not properly selected may not determine the correspondence between two voices. Moreover a better attention should be paid to those characteristics of the system which are subject to the largest conditions of variability.

Bibliography

BALDWIN J. & FRENCH P.: *Forensic Phonetics*, London-New York, Pinter, 1990.

BATTANER E., GIL J., MARRERO V., LLISTERRI J., CARBÓ C., MACHUCA M.J., de la MOTA C., RÍOS A.: “VILE: Estudio acústico de la variación inter e intralocutor en español”, *Actas del II Congreso de la Sociedad Española de Acústica Forense* (Barcelona, Spain, 2003), 59-70.

BERTINETTO P.M.: *Strutture prosodiche dell’italiano*, Firenze, Accademia della Crusca, 1981.

BOË L.J.: “Forensic voice identification in France”, *Speech Communication*, 31 (2-3), 205-224, 2000.

BOË L.J., BIMBOT F., BONASTRE J.F. & DUPONT P.: “De l’évaluation des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique”, *Langues*, 2, 270-288, 1999.

BRAUN A.: “Procedures and perspectives in forensic phonetics”, *Proc. of the 13th Int. Congress of Phonetic Sciences* (Stockholm, Sweden, 1995), vol. III, 146-153, 1995.

BRAUN A. & KÜNZEL H.: “Is forensic speaker identification unethical - or can it be unethical *not* to do it?”, *Forensic Linguistics*, 5(1), 10-21, 1998.

CERRATO L. & PAOLONI A.: “Utilizzo dei parametri acustico-fonetici nella identificazione del parlante in ambito forense”, In: P. Cosi *et al.* (eds.), *Voce Canto Parlato. Studi in onore di Franco Ferrero*, Padova, Unipress, 59-66, 2003.

ERIKSSON A. & WRETLING P.: How Flexible is the Human Voice? A Case Study of Mimicry. *Proc. of EuroSpeech’97 - 5th European Conference on Speech Comm. and Technology* (Rhodes, Greece, 1997), 175-178, 1997.

FALCONE M., PAOLONI A. & DE SARIO N.: “IDEM: a software tool to study vowel formant in speaker identification”, *Proc. of the 13th Int. Congress of Phonetic Sciences* (Stockholm, Sweden, 1995), vol. III, 294-297, 1995.

FARNETANI E. & KORI Sh.: “Effects of syllable and word structure on segmental durations in spoken Italian”, *Speech Communication*, 5, 17-34, 1986.

FARNETANI E. & VAYRA M.: “Word- and phrase-level aspects of vowel reduction in Italian”, *Proc. of the 12th Int. Congress of Phonetic Sciences* (Aix-en-Provence, France, 1991), 2, 14-18, 1991.

FAVA E. & MAGNO-CALDOGNETTO E.: “Studio sperimentale delle caratteristiche elettroacustiche delle vocali toniche e atone in bisillabi italiani”, In: R. Simone *et al.* (eds.), *Studi di Fonetica e Fonologia, Atti del Conv. Int. della SLI* (Padova, Italy, 1973), Roma, Bulzoni, 35-80, 1976.

FEDERICO A.: “Le basi statistiche della decisione bayesiana nella identificazione del parlatore”, *Materiali presentati e discussi al “Seminario di Fonetica Forense” del 28° Convegno Nazionale dell’Ass. Italiana di Acustica* (Trani, Italy, 2000).

FEDERICO A. & PAOLONI A.: “Parametric speaker recognition over large population of telephonic voices”, *Proc. of EuroSpeech’95 - 4th European Conference on Speech Comm. and Technology*, (Madrid, Spain, 1995), 1995.

FERRERO F.E.: “Caratteristiche acustiche dei fonemi vocalici italiani”, *Parole e Metodi*, 3, 87-96, 1972.

FERRERO F.E.: “Problemi spettroacustici di classificazione e di misurazione delle vocali: un contributo”, In: F. Cutugno (ed.), *Fonetica e fonologia degli stili dell’italiano parlato*, *Atti delle VII Giornate di Studio del “Gruppo di Fonetica Sperimentale” dell’Ass. Italiana di Acustica* (Napoli, Italy, 1996), Roma, Esagrafica, 235-264, 1996.

IBBA G., PAOLONI A. & SAVERIONE B.: “Significatività della durata delle consonanti occlusive ai fini del riconoscimento del parlatore”, *Rivista Italiana di Acustica*, 3/1, 23-39, 1979.

KÜNZEL H.J.: “Some general phonetic and forensic aspects of speaking tempo”, *Forensic linguistic - Int. Journal of Speech; Language and the Law*, IV/1, 48-83, 1997.

MAROTTA G.: “La sibilante preconsonantica in italiano: questioni teoriche ed analisi sperimentale”, In: R. Aiello & S. Sani (eds.), *Scritti linguistici in onore di Tristano Bolelli*, Pisa, Pacini, 393-438, 1995.

MORI L. & PAOLONI A.: “Sulla sociolinguistica forense: la costituzione di corpora vocali per l’analisi della velocità di articolazione in italiano”, In: A. De Dominicis *et al.* (eds.), *Atti delle XIV Giornate di Studio del “Gruppo di Fonetica Sperimentale” dell’Ass. Italiana di Acustica* (Viterbo, Italy, 2003), Roma, Esagrafica, 75-80, 2004.

MULLER Ch.: *Initiation à la statistique linguistique*, Paris, Larousse, 1968.

NOLAN F.: *The Phonetic Bases of Speaker Recognition*, Cambridge, Cambridge University Press, 1983.

NOLAN F.: "Speaker recognition and forensic phonetics", In: W.J. Hardcastle & J. Laver (eds), *A Handbook of Phonetic Sciences*, Oxford, Blackwell, 1997.

PAOLONI A.: "La voce come elemento di identificazione della persona", In: A. De Dominicis (ed.), *La voce come bene culturale*, Roma, Carocci, 125-139, 2002.

PAOLONI A.: "Note sul riconoscimento del parlante nelle applicazioni forensi con particolare riferimento al metodo parametrico IDEM", *Rivista Italiana di Acustica*, 27/3-4, 113-128, 2003.

PAOLONI A.: *Appunti di fonica*, Roma, Fondazione Ugo Bordoni, 2000.

PAOLONI A. & PETTORINO M.: "La voce imitata: un'analisi acustica-percettiva", In: A. Regnicoli (ed.), *La fonetica acustica come strumento di analisi della variazione linguistica in Italia*, *Atti delle XII Giornate di Studio del GFS dell'Ass. Italiana di Acustica* (Macerata, Italy, 2001), Roma, Il Calamo, 219-226, 2003.

PIAZZA A., IORIO M., BENZI P., ROBETTI I., BERTINETTO C., MASCARO V., CELLI R. & COCHIS M.L.: "La perizia fonica", *Minerva Medicolegale*, 119(4-suppl.), 3-21, 1998.

REYNOLDS D.A. & HECK L.P.: *Speaker Verification: from Research to Reality*. MIT-Lincoln Laboratory - Nuance Communications (*ICASSP Tutorial*, Salt Lake City), 2001.

ROMANO A.: "Statistiche di frequenza fondamentale per uno stesso locutore in diverse condizioni di produzione", *Atti del 28° Convegno Nazionale dell'Ass. Italiana di Acustica* (Trani, Italy, 2000), 249-252, 2000.

ROMANO A.: "Geminate iniziali salentine: un contributo di fonetica strumentale alle ricerche sulla geminazione consonantica", In: R. Caprini (a cura di), *Parole romanze. Scritti per Michel Contini*, Alessandria, Dell'Orso, 349-376, 2003.

ROMANO A., MANCO F. & TOMATIS M.: "Caratterizzazione del parlato sulla base di indici temporali: un esperimento su parole isolate di un campione di studenti torinesi", *Bollettino dell'Atlante Linguistico Italiano*, 27 (2003), 237-251, 2004.

ROMITO L.: *Manuale di fonetica articolatoria, acustica e forense,* Cosenza, Università della Calabria - Centro editoriale e librario, 2000.

TOSI O.: *Voice Identification: Theory and legal applications*, Baltimore, Univ. Park Press, 1979.

TRUMPER J.: *Sociolinguistica giudiziaria: preliminari di metodi e applicazioni*, Padova, CLESP, 1979.

VOGEL I.: *La sillaba come unità fonologica*, Bologna, Zanichelli, 1982.

WOODS A., FLETCHER P. & HUGUES A.: *Statistics in Language Studies*, Cambridge, Cambridge Univ. Press, 1986.