Acoustic analysis and perception of classes of sounds (vowels and consonants)

Maria Gabriella Di Benedetto° and Anna Esposito*
°INFOCOM Dept., Rome University "La Sapienza"
Via Eudossiana 18, I00184 Rome, Italy
*International Institute for Advanced Scientific Studies "E. R. Caianiello"
Via G. Pellegrino 19, I84019 Vietri sul Mare (SA), Italy

1 Introduction

Speech application systems such as recognizers, synthesizers, and coders require as a necessary input that some fundamental questions about the structure of speech sounds be understood. For example, actual recognition systems are not robust in terms of noise and variations of voice quality and speaker. Synthesizers suffer for being characterized by poor and unnatural quality of speech, and by a lack of flexibility in terms of changes in voice gender and reflection of emotional states. Finally, coders at low bit rates perform poorly in terms of subjective quality. Moreover, at medium bit rates, an improvement in coding algorithms could be obtained if the effects of noise introduced by the transmission channel on the acoustic features of speech segments were better understood.

One way for gaining insight in how speech sounds are structured is to analyze the acoustic signal in a very detailed manner. This procedure allows to define a number of acoustic attributes, which characterize, in a significant way, the speech units of a language. However, as well known, the acoustic attributes of a given sound vary as a function of many factors, among which the context in which the sound is embedded, and the speaker, have the strongest effects. Therefore, a careful analysis of the speech signal requires sophisticated experiments in which a large number of tokens of a given sound are recorded to form the experimental data base. Accurate measurements of the acoustic parameters must be performed on all collected data, and their significance must be evaluated. However, even by doing so, a variety of factors such as speaking rate, emotional state, and suprasegmental features may be neglected anyway. It should be noted, in addition, that the acoustic attributes also depend upon the language under examination, and thus the findings obtained on one language can hardly be extended to other languages. All these aspects make the analysis complex and time-consuming, a feature which is often not compatible with the timing of speech application systems development.

The properties which are found to be significant at the acoustic level, may or may not be so at the perceptual level. For example, an acoustic parameter such as the third formant (F3) may significantly vary from one vowel to the other, and therefore exhibit values which are peculiar of a vowel, while perceptually the F3 information might be integrated with the information contained in F2. Therefore, the acoustic analysis must be supported and integrated by a perceptual analysis in order to evaluate the perceptual relevance of the acoustic attributes.

The aim of this paper is to present a set of experiments designed in the above view for a variety of speech sounds. Results of acoustic and perceptual analyses carried out on both consonants and vowels will be reported. The output of these analyses are the acoustic and perceptual attributes of classes of sounds.
The paper is organized as follows. Section 2 will be dedicated to vowels and section 3 to consonants. A general discussion on the results reported and on their significance will conclude this paper.

2 Vowels

2.1 Introduction

The present study focuses on characterizing vowel segments in terms of acoustic parameters. Therefore, the problem is to find acoustic properties of vowels which prove to be both speaker independent (normalization problem) and context independent (coarticulation problem). It is clear that these two problems are directly related to a similar issue that consists in finding properties in the acoustic vowel waveform which are invariant with respect to speaker, language, and phonetic context variations.
As well known, when a vowel is produced, the vocal tract can be considered as a sequence of acoustic tubes which resonate at particular frequencies, called formants. The position of these formants depends essentially upon the length and the cross-sectional area of each tube. If the vocal tract is modeled by a linear system, then the formant frequencies F1, F2, F3, F4 are the imaginary parts of the poles of the transfer function of this system. The losses due to the non-rigid wall characteristics and to the non-ideal junctions and interfaces at the extremities of the tract constitute the real parts of the poles. When a vowel is produced, the source signal is formed by the pressure of the air expirated by the lungs and modulated by the opening and closing of the vocal folds. The spectrum of the source, which is a sequence of pulses of triangular shape with period T0, is then characterized by the harmonics of the fundamental frequency F0=1/T0.
The fundamental frequency, F0, is proportional to the tension of the vocal folds. The variation of the tension of the folds is related to the tension of the surrounding muscles and to the movement of the hyoid bone [1], so that the source and the filter are coupled during the articulation. The acoustic signal is the result of a convolution between the source signal and the impulse response of the filter, which models the vocal tract behaviour. This signal is periodic and in its spectrum the F0 harmonics which are next to the formants are emphasized. Depending upon which vowel is pronounced, the tongue position varies and, consequently, the size of each of the acoustic tubes, the rigidity of the walls, and the tension of the vocal folds are modified, influencing the F0, F1, F2, F3...values. The acoustic model predicts the

relative invariance of the formants of the extreme vowels [i,a,u], when, changing the speaker, the dimensions of the vocal tract are varied [2].
A variation of the sound pressure generates a variation of the hydro-mechanic pressure inside the ear. Many auditory nerve fibers on the basilar membrane of the cochlea are excited and output a signal formed by a sequence of electric pulses which is transmitted to the brain. Each fiber has an action similar to that of a non linear low-pass filter centered on a characteristic frequency CF. These filters cover the spectrum according to a scale which is approximately logarithmic, and which has been formalized in the Bark scale. On the basis of physiological experiments [3], it was shown that, in response to vocalic stimuli, the electric pulses generated by all the fibers code the position of the formants. In addition, on the basis of perceptual experiments, it was shown that two peaks in the spectrum of a vowel closer than 3-3.4 Bark are integrated in one peak in intermediate position ('spectral center of gravity effect', [4]). As regards the first formant F1, perceptual experiments showed that this formant is perceived relatively to F0 [5]. Later studies [6] showed that the relation between F1 and F0 may not be a simple linear relation, implying a non-uniform vowel normalization in agreement with [7].

Vowel normalization, i.e. the problem of finding acoustic properties of vowels which prove to be speaker independent, can be based on either no a-priori knowledge on speaker's characteristics (intrinsic vowel normalization), or on the use of some a-priori knowledge on each speaker's vowel system (extrinsic vowel normalization). In the first category of methods, the perception of a vowel is supposed to be dependent only upon the signal itself, and, in particular, upon few parameters characterizing it. In particular, the formant frequencies and the fundamental frequency are crucial factors in identifying a vowel [8,9]. The normalization consists in finding a function of these parameters which proves to be invariant with respect to the speaker. Several functions, having a normalization effect, have been proposed; most often, the formant and fundamental frequencies are expressed in Mels, Barks, or logarithmic units, and formants ratios or formant differences are considered [10]. In the second category of methods, the perception of a vowel is supposed to be largely influenced by the context in which it is included, implying the existence of an adaptation time to the specific speaker [11].
In the present paper, a method which falls in the above first category will be described and analyzed in section 2.2. The study follows an approach by which the information contained in vowels is coded in a way similar to that used by the auditory system. The acoustic parameters selected for representing vowels, independently of the speaker, were tested by using a statistical and a neural classifier. The performance of these two classifiers will then compared.

The problem of finding acoustic properties of vowels which prove to be context independent (coarticulation problem) has received particular attention over the last decades. However, although traditionally formant frequencies have been directly associated to dimensions characterizing vowels [12,13,14], such as height, backness, and tenseness, they have rarely been taken into account as acoustic parameters which vary in time. When formants are sampled at one or more instants of time in

the vocalic portion, an important part of the information might be neglected. Previous studies have analyzed the possibility of taking into account the whole formant trajectories, in particular the F1 trajectory for representing vowel height [15,16]. In the present paper (section 2.3) the problem of representing the feature [tense] by analyzing properties related to formant trajectories will be analyzed. The reported study describes perceptual experiments carried out in order to specify perceptual correlates of the tense/lax distinction in two languages (American English and Italian) [21].The feature [tense] has been broadly used in models of generative phonology to indicate the linguistic distinction between consonants (for example [t] vs. [d]) and vowels (for example [e] Vs [ɛ] in Italian, as can be found in Muljacic [17]). Various attempts have been made to correlate this distinctive feature to some specific property. From an articulatory point of view, the question is controversial. One hypothesis is the major tenseness of some muscles of the vocal tract during the production of the sounds defined as tense. For example, the tension of the tongue against the hard palate appears as one property related with the production of the [t,d] distinction, but there is no clear evidence of an effectively well-defined muscular distinction in the production of tense and lax vowels. From an acoustic point of view, properties based on measurements of the time-varying spectrum may be related to the feature [tense], namely some properties of the formant trajectories [18] and the variation of the relative distance between some of the formants, with respect to the central acoustic position of the vowel schwa. It appears then that the feature [tense] might be characterized by different acoustic and articulatory manifestations. As proposed in [19], this makes this feature a possible candidate for being called a "cover feature" in distinctive feature theory. From a perceptual point of view, the variation of F1 and F2 trajectories of synthetic vowel stimuli led to a perceptual distinction between tense and lax vowels [20]. Furthermore, the evolution of languages shows that different languages have made different use of the tense/lax distinction, so that sounds effectively linguistically distinct along the tense/lax dimension have been created or progressively deleted. For example, in Italian, minimal pairs based on the difference between two vowels of a tense/lax pair ([e] vs. [ɛ], in the words [peska], 'fishing', vs. [pɛska], 'peach') have disappeared in most dialects and the distinction is rarely made in natural speech.

2.2 Vowel normalization

The aim of this study was to verify whether, in Italian, the use of auditory parameters, such as the Bark-transformed formant differences, is more appropriate to represent vowels pronounced by different speakers than the traditional formant values expressed in Hertz. In addition, the performance of a statistical and of a neural net recognizer, based on the above input parameters was compared [22].
Spectral measurements and statistical analyses of all Italian vowels uttered by 25 male and 11 female speakers will be described in section 2.2.1. In agreement with a model of American English vowels [10], it will be shown that the difference between the first formant and the fundamental frequency (F1-F0) and the difference between successive formants on the Bark scale (F2-F1 and F3-F2) are effective in

normalizing male and female spectral differences and in better clustering vowel areas. The results will be discussed on the basis of a model of speech articulation, and experimental theory of speech perception.
Two vowel recognition methods, based on the results of the acoustic analysis, will then be compared. The input to both recognizers was a vector X the elements of which represented the spectrum of the acoustic signal according to a perceptual model presented in [10]. The first classifier, described in section 2.2.2, was based on a statistical approach applying discriminant analysis. The second classifier, described in section 2.2.3, was based on a neural network approach and was an implementation of a multi-layer perceptron. In section 2.2.4, the classification error rates will be compared and discussed, in relation to the different underlying assumptions that the two algorithms make on the input data. Further detail about this experiment can be found in [23].

2.2.1 Experimental conditions and procedure

The Italian vowel system consists of seven vowels [i,e,ɛ,a,o,ɔ,u]. Three vowels [i,e,ɛ] are front vowels, while four vowels [a,o,ɔ,u] are non-front vowels. In the present study, the Italian vowels, pronounced by 25 male and 11 female speakers, were analyzed. The vowels, extracted from the data-base developed in [24], were pronounced in pV# syllables by male speakers and in isolation by female speakers. This data-base constitutes a reference point for many acoustic studies of Italian vowels (see for example [25]), and was also used in applications such as a text-to-speech synthesis system. For each vowel, a short temporal window was considered (4 periods long, located around the maximum of the signal envelope). For each vowel, the fundamental frequency F0 was computed using an algorithm based on the cepstrum of the signal, and the first four formants were found by manual comparison of the local maxima in the Fourier transform (FFT) of the signal and the maxima of the autoregressive analysis (AR) of the spectrum (linear prediction with 16 coefficients found with the autocorrelation algorithm).
The comparison between the two spectra was necessary, as, frequently, the peaks in the AR spectrum under 2000 Hz are slightly lower than those found from the examination of the FFT, when F0 is high (this happened for most of the female speakers). The average values of F0, F1, F2, F3 and F4 expressed in Hertz, for each vowel, for the 36 speakers, are shown in Fig.1.

The values of F0, F1, F2, F3, and F4 were the object of a statistical investigation which showed that the measurements expressed in Bark reduced the difference between male and female speakers. The above difference was quantified by computing the Mahalanobis distance which is the Euclidean distance between the average values of two groups, divided by the variance of each group. Table I reports the Mahalanobis distance between male and female groups, for different vowels, in the case of the F1 vs. F2 (in Hz) and (F1-F0) vs. (F2-F1) (in Bark) representations, and shows that this distance, except for the vowel [u], decreased when the measurements were in Bark. As reference, distances lower than 10 indicate a

significant overlap between groups.

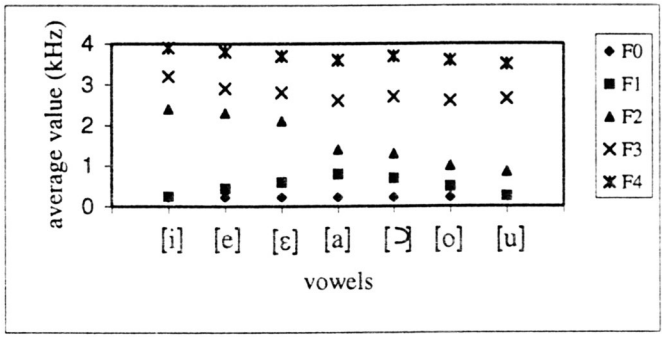

Figure 1 - Average values of F0, F1, F2, F3, and F4, for all Italian vowels, considering 25 male and 11 female speakers

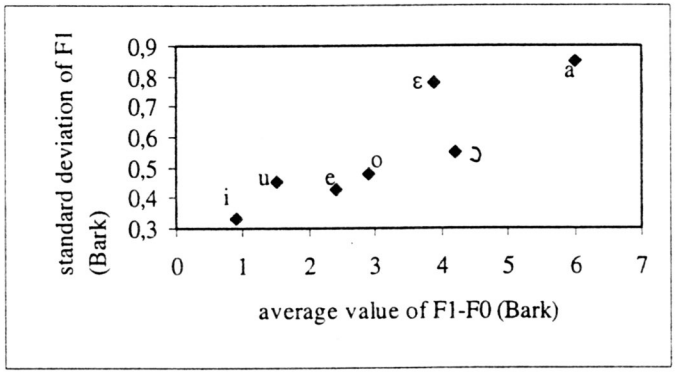

Figure 2 - Variability of the first formant F1 as a function of the F1-F0 distance

Figure 2 shows the relation between the standard deviation of F1 and the average values of F1-F0, for different vowels. One can notice, from Fig.2, that a correlation existed between these two parameters, and that in particular when F1 was close to F0, F1 was relatively unchanged for different speakers, in agreement with [6]. In addition, the distance F1-F0 seems to be correlated to the phonological classification of vowels according to vowel height, in agreement with the experiments reported in [5].
The relations between the standard deviations of F2-F1 and of F3-F2 with respect to the average values of F2-F1 and F3-F2, all expressed in Bark, showed that, when two formants are far apart, the variability of the distance between the two formants is higher than when the formants are close. In addition, as shown in Fig.3, all front

vowels verified F3-F2<3 Bark, while all non front vowels verified F3-F2>3 Bark.

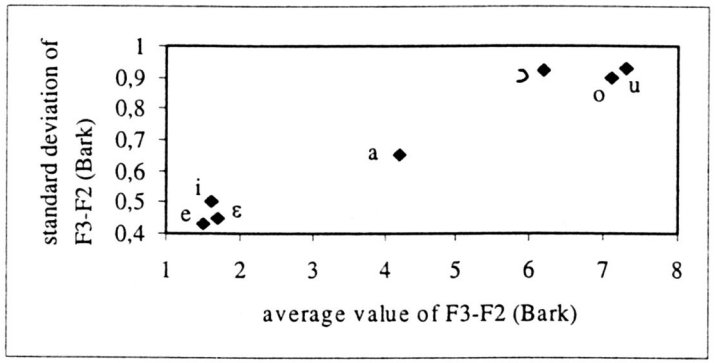

Figure 3 - Variability of the F3-F2 distance as a function of the average value of F3-F2

2.2.2 Statistical classifier

Each vowel was represented by the values of F1-F0, F2-F1, and F3-F2, all expressed in Bark. Given a significant sample formed by a number of utterances of the vowels, each new utterance was classified as a particular vowel according to the following procedure .
First, the new utterance was labeled as front or back, according to the value of F3-F2. Secondly, the two vowels which had the lower Euclidean distance in terms of F1-F0 and F2-F1 values from the utterance analyzed, were found.
The choice between the two selected vowels was based on a linear discriminant analysis. The linear discriminant analysis determined the linear combination of the inputs which maximized the difference between the average values of the two groups with respect to the variance of each group. The results obtained by the application of this classifier are shown in Table II. These results are satisfactory for the cardinal vowels [i,a,u] and for front vowels, while they are less satisfactory for back vowels.

2.2.3 Neural classifier

Two 2-layers perceptrons, one for front vowels and the other for non front vowels, were trained. A similar architecture was used for Swedish vowel classification [26]. Each node in a perceptron computes the weighted sum of the inputs. The result goes through a sigmoidal threshold device. In this way, each node creates two regions, in the input space, separated by a hyper-plane. The output of each node of a layer is connected to the input of the nodes of the following layer, in order to inhibit or excite their response. In the case of vowel classification, the input to the first layer of N "hidden" nodes was formed by the vector [X1,X2], where X1=F1-F0 and

X2=F2-F1. The N outputs of the hidden nodes were connected to the second layer formed by M output nodes, one for each vowel.
The training of the nets was obtained by using the back-propagation error technique, which consists in presenting to the net many samples of the vowel to be recognized. In each cycle, the input [X1,X2] and the desired output were presented to the net. The desired output was high (equal to 1) while for all the other nodes it was low (equal to 0). The output of the perceptron was then computed and the weights were modified proportionally to the error using a constant gain g. After each cycle, the training was speeded up by adding a quantity proportional to the correction made during the preceding cycle, according to a constant value a. During the first cycles, all the outputs were around 0.5. Slowly, after a number of cycles, the outputs tended to the values 0.1 or 0.9.
Different perceptrons were tested each being characterized by a different number of hidden nodes N, and different initial weights. The perceptron for front vowels [i,e,ε] was correctly trained after 1080 cycles using N=12, g=0.1, and a=0.9, and all the initial weights being small and of random value. Perceptrons with a lower number of hidden nodes tended to distinguish only extreme vowels. The classification rates and confusion matrix obtained after the training are shown in Table III.
The behaviour of the perceptron for non front vowels [a,o, ɔ,u] was less satisfactory due to the significant overlap between the vowels [o,u] and [o,ɔ]. Independently of the number of nodes, the perceptrons tended to confuse more than 25% of the utterances of [o] and [ɔ], as they were in a way cheated during the back-propagation of the error by ambiguous pronunciations of [o,ɔ], in the region of intersection between [u,o] and [o,ɔ]. In a second series of experiments, the two vowel classes [o] and [ɔ] were confused in one class. In this case, the training of the perceptron was correct with N=20, g=0.1, a=0.4, after 1512 cycles. The classification errors and confusion matrix are shown in Table IV.

2.2.4 Conclusion

The statistical analyses and the behaviour of the two classification methods verified that the acoustic and perceptual parameters selected were significant for the distinction of Italian vowels. It is important to point out, however, that it is necessary to carry out studies in which the effects of prosody and coarticulation are also taken into account.
As regards the acoustic correlates of distinctive phonetic features for Italian vowels, the overlapping between [o,ɔ] shows that the tense/lax dimension is not well represented by the selected parameters. In fact, this category seems to be related to temporal properties of the first formant trajectory, as will be further analyzed in section 2.3. In addition, the overlapping between [u,o] could be related to the fact that F1-F0 does not represent properly vowel height when F1 and F0 are close and F1 is low, confirming what reported in [6].
As regards applications, a normalization between vowels pronounced by male and female speakers was obtained. The comparison between the statistical and the neural classifiers verified the equivalence of the two methods when the number of input

variables was low and the groups can be separated by a linear combination of the input data. The training of the neural net with the back-propagation error algorithm was time-consuming and depended upon the number of hidden nodes of the first layer. When the overlapping between two groups was significant (Mahalanobis distance lower than 10), the behaviour of the perceptron was less satisfactory than the behaviour of the linear discriminant analysis. Nevertheless, if the number of variables was to be high (>10) the statistical procedure would loose control on data (computational difficulties in the inversion of the covariance matrix) and consequently it would be difficult to discriminate among groups on the basis of a linear combination of the inputs.

2.3 Acoustic correlates of the feature tense

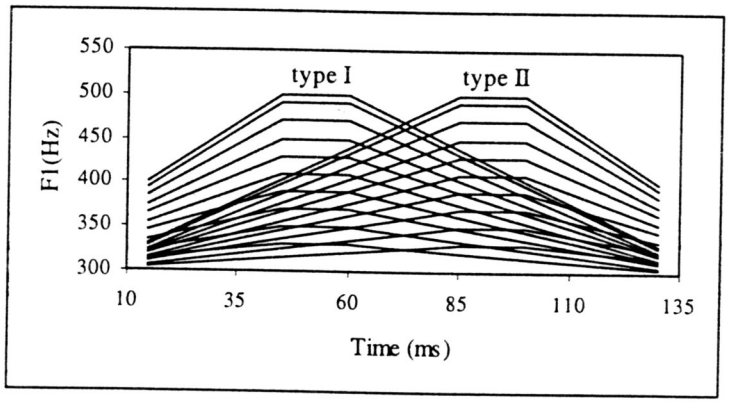

Figure 4 - Schematic F1 trajectories for long type I and type II stimuli

Perceptual experiments carried out in order to specify perceptual correlates of the tense/lax distinction will be described in the present section. These experiments investigated the influence of the first formant frequency trajectory shape on the perception of the tense/lax distinction of American English and Italian vowels. It should be noted that the Italian vowel system includes two pairs [e,ɛ] and [o,ɔ] which are described phonologically as tense/lax pairs.

2.3.1 Description of the experiment

The stimuli used were all dVd synthetic syllables in which the vowel could have two standard trajectory shapes. The stimuli were described to the subjects as being dVd synthetic syllables, and the subjects were asked to identify the vowel in the stimuli as any vowel of the vowel system of their language. The subjects were also asked to note whether the perception of the consonant of the dVd syllable disturbed the perception of the vowel. All the stimuli were synthesized with the Klatt synthesizer [27]. The F1 trajectory was the only parameter by which two stimuli, with the same

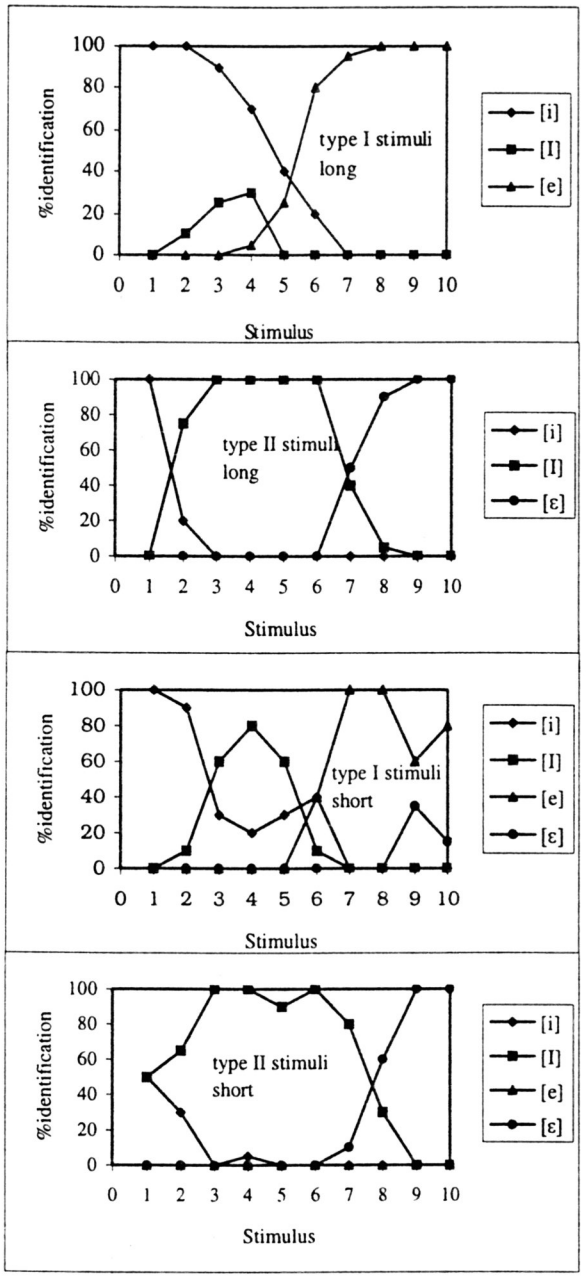

Figure 5 - Results of the experiment with long and short stimuli in terms of percent identification of tense vowels [i] and [e] vs lax vowels [I] and [ɛ], for one American subject.

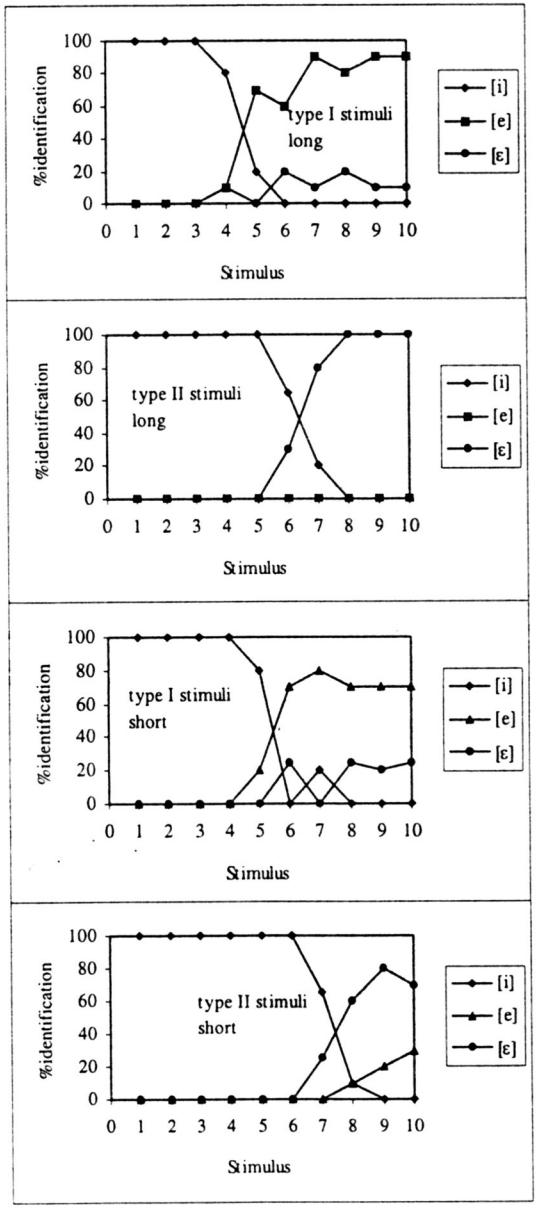

Figure 6 - Results of the experiment with the long and short stimuli in terms of percent identification of tense vowels [i] and [e] vs the lax vowel [ɛ], for one Italian subject.

Table I. Mahalanobis distance between male and female groups in the F1 vs. F2 space (F1 and F0 are expressed in Herz), and in the F1-F0 vs. F2-F1 space (F0, F1, and F2 are expressed in Bark)

Vowel	distance (kHz)	distance (Bark)
[i]	14	4.5
[e]	19	4.2
[ɛ]	11.9	2.1
[a]	10.3	2.8
[ɔ]	2.25	0.11
[o]	1.32	0.79
[u]	0.34	1.62

Table II. Classification results obtained with the statistical classifier. Confusion matrix and percentage of correct classification

	[i]	[e]	[ɛ]	[a]	[ɔ]	[o]	[u]	%
[i]	34	2	0	0	0	0	0	94.4
[e]	0	35	1	0	0	0	0	97.2
[ɛ]	0	3	33	0	0	0	0	91.7
[a]	0	0	0	36	0	0	0	100
[ɔ]	0	0	0	1	31	4	0	86.8
[o]	0	0	0	0	4	26	6	72.2
[u]	0	0	0	0	0	4	32	88.9
Average: 90.2								

Table III. Classification results obtained with the neural classifier, for front vowels. Confusion matrix and percentage of correct classification

	[i]	[e]	[ɛ]	%
[i]	36	0	0	100
[e]	1	34	1	94.4
[ɛ]	0	1	35	97.2
Average:				97.2

Table IV. Classification results obtained with the neural classifier, for non front vowels. Confusion matrix and percentage of correct classification

	[u]	[o, ɔ]	[a]	%
[u]	31	5	0	86.1
[o, ɔ]	3	68	1	94.4
[a]	0	0	36	100
Weighted average:				93.7

F1 maximum, differed; depending on the shape of the F1 trajectory, the stimuli were identified as of type I or II. Two sets of stimuli were used. In the first set (long stimuli), the duration of the stimuli was 115 msec and of the steady-state 15 msec, for all stimuli, while the onglide duration was 30 msec for type I and 70 msec for type II stimuli. In the second set (short stimuli), the duration of the stimuli was 95 msec (20 msec shorter than the stimuli of the first set) and of the steady-state was 15 msec, for all stimuli, while the onglide duration was 30 msec for type I and 50 msec for type II stimuli. In each set, 10 stimuli of each type were considered, with an F1 maximum ranging from 330 to 500 Hertz in steps of 20 Hertz. Schematic F1 trajectories for stimuli of the first set are shown in Fig.4.
Four subjects participated in the experiment; two were native speakers of American-English, and two were native speakers of Italian. All the subjects had profound knowledge only in their native language. None of the subjects knew about the purpose of the experiment.
Corresponding to each set of stimuli, the experiment consisted of three phases. In the first phase, each type I stimulus was presented ten times. The ten type I stimuli were ordered to make each stimulus follow another only once. In the second phase, the same procedure was used with type II stimuli. In the third phase, stimuli of both types were presented. On the whole, each stimulus was presented twenty times. At the end of each phase, the subject could rest for few minutes. The test was approximately 45 minutes long.

2.3.2 Results of the experiment

The vowels of the stimuli were identified as [i,I,e,ɛ] by the American subjects and as [i,e,ɛ] by the Italian subjects. The American subjects characterized the vowel [e] as the non-diphthongized [e^{y}]. None of the subjects declared that the perception of the consonant of the stimuli (perceived as [d]) disturbed the perception of the vowel.

Figure 5 shows the results obtained for one of the American subjects, with the first set (long stimuli) and with the second set (short stimuli). These results were representative of the results of the two listeners, and show that stimuli with the F1 maximum in the low frequency range were mainly associated with the tense vowel [i]. Very possibly this was due first to the low value of the F1 maximum and secondly to the very slight difference between the F1 maximum and the onset and offset values (flat shape of the F1 trajectory). Stimuli with the F1 maximum in the medium frequency range were perceived as [I] if of type II, while there was an uncertainty area for type I stimuli (perceived as [i], [I] or [e]). This uncertainty was more evident with short than with long stimuli. In the high frequency range, the distinctions between the perception of type I and type II stimuli were sharp: type I stimuli were perceived as [e] and type II stimuli as [ɛ].

The results obtained with the first set (long stimuli) for Italian subjects are presented in Fig.6, which shows that type I stimuli were mainly perceived as the tense vowels [i,e] and type II stimuli as [i] when F1 maximum was in the low frequency range

and as [ɛ] when it was in the high range. The results of the experiment with the second set (short stimuli), also presented in Fig.6, were similar to those obtained with long stimuli. Figure 6 also shows that short type II stimuli with high F1 maximum were sometimes perceived as [e], while this did not occur with long stimuli. Possibly this effect was due to the fact that, in short stimuli of type II, the F1 maximum was reached more rapidly than in long stimuli of type II. Thus, short type II stimuli had an F1 trajectory shape more similar to the F1 trajectory shape of short type I stimuli than did long stimuli. Consequently, the subjects tend to perceive short type II stimuli as tense vowels.
Note that the Italian vowel system does not include the vowel [I] so that the vowel areas of [i] and [e] are contiguous in the F1 dimension, causing the Italian subjects responses not to show the same uncertainty found for the American subjects, when F1 maximum was in the medium frequency range.

2.3.3 Discussion

Results of the experiments presented showed that in general stimuli characterized by a fast rising of the F1 trajectory tended to be associated with tense vowels while stimuli with a slow F1 rising were in general perceived as lax vowels, except when the F1 maximum values were low and consequently the F1 trajectory shape was almost flat. For type I stimuli, there was an uncertainty area in the responses of the American subjects, when F1 maximum was in the medium frequency range.
Huang [20] investigated the influence of the relative duration of onglide, steady-state and offglide on the perception of the tense/lax distinction of American English vowels using synthetic stimuli with three different F1 and F2 trajectories; standard (onglide time:20 msec, steady-state:195 msec, offglide time:20 msec), tense (onglide time:50 msec, steady-state;135 msec, offglide time: 50 msec) and lax (onglide time:50 msec, steady state: 85 msec, offglide time:100 msec). Longer offglide duration was chosen for lax stimuli, according to Lehiste and Peterson [18]; these investigators noted, on acoustic measurements made on monosyllabic CVC words, that tense vowels had a steady-state longer than lax vowels, and that tense vowels had an offglide shorter than for lax vowels. Huang's findings were that vowel responses shifted towards tense vowel judgments when stimuli with tense formant trajectories were considered, in the case of either tense/tense or tense/lax non-high vowel pairs. This shift was also present going from tense to standard stimuli for non-high vowel pairs. On the results of Huang's experiment, we observed that the shift obtained going from lax to tense and from tense to standard were similar.
In another experiment, Huang investigated the role of duration in the perception of vowels of American English. Huang considered stimuli with similar onglide and offglide durations (20 msec) but different steady-state durations (0,50,100,195 msec). Huang's findings were that a shift towards tense vowel judgments was found going from the shortest to the longest duration. A shift was also present in the case of lax/lax pairs, and it was in the direction of the vowel characterized by a higher value of F1.
The results of the experiments of the present study showed that stimuli characterized

by a long offglide and a short onglide of F1 tended to be perceived as tense vowels while stimuli with a short offglide and a long onglide tended to be perceived as lax vowels, and that shortening the stimuli had little influence on the perception of the synthetic vowels by the American subjects. The divergencies found with the studies mentioned above could be explained as follows: 1) different components were considered in the present study (same steady-state duration and different onglide and offglide durations), and 2) the difference in duration between the stimuli in the two sets was 20 msec and the results of Huang's experiment showed that in the case of the [i]-[I] pair the shift was evident going from the shortest to the longest stimuli but that it was small going from the 90 msec to the 140 msec stimuli (which were similar in duration to the stimuli considered in the experiments of the present study). The shift of judgments found by Huang in the case of tense/standard stimuli could be interpreted, on the basis of the results of the present study, as due to the different initial slope of the F1 trajectory. However, this explanation is not unique; an alternative interpretation could be that the determining factor of the effect is the duration of the steady-state, which assumes different values in the two cases. Both interpretations could be valid. In addition, in Huang's study the F2 values of the synthetic vowels varied along a continuum between adjacent vowels, while in the present study the F2 values were the same for all stimuli.
Future work should tend to specify acoustic correlates of the tense/lax distinction, which would be possibly related to the effect observed in the perceptual experiments described. If this link existed, it would be possible to hypothesize that these properties may be used by the auditory system to encode the information about tenseness or laxness of vowels.

3 Consonants

In this paragraph, consonant properties will be analyzed. In particular, stop consonants will be described in terms of:

- spectrum properties for identifying the place of articulation;
- duration as an acoustic parameter for characterizing stop consonants in single and geminated forms.

The results reported focused on the analysis of consonants in Italian. References to results for consonants in other languages will also be considered. Furthermore, the above acoustic features will be evaluated in terms of their perceptual relevance.

3.1 The amplitude of the peaks in the spectrum as acoustic attributes of the place of articulation

3.1.1 The problem

Phonetic studies on consonants have developed a classification according to their articulatory features. Articulatory features are determined by the action of the glottis, tongue root, tongue body, tongue blade, velum, jaw, and lips to produce speech

sounds. Thus, the description of the acoustic patterns of consonants follows a classification according to the articulators involved in their production. For example, if a consonant is produced with a complete closure at the lips, it is classified as labial.

The articulatory features of consonants are of three types: 1) features of place of articulation, 2) features of manner of articulation, and 3) voiced/voiceless distinction (see [28] for details). A set of acoustic attributes (such as formant transitions, duration, energy, shape of the spectrum, etc.) is associated to the above mentioned features. The acoustic attributes describe the broad acoustic properties of consonants (and, in general, of all phonetic units) and are derived from spectral measurements. It would be desiderable to characterize an articulatory feature using acoustic attributes. For example, if the transition from a consonant to a vowel shows (in the speech spectrogram) a rising of all the formant frequency values, it is possible to hypothesize that the consonant is produced with a closure at the lips.
There is no one-to-one correspondence between articulatory features and acoustic attributes, rather, the acoustic information for successive linguistic units overlaps and interacts. Therefore more than one acoustic attribute could be required to characterize an articulatory feature. Some acoustic attributes are robust and could accompany a particular consonant in any vowel environment. Some others are context-conditioned; i. e., for a particular consonant, they are determined by the vowel. For example, in Italian, the presence (in the speech spectrogram) of a strong energy at low frequencies is a robust acoustic attribute which characterizes voicing [29, 30]. Instead, formant transitions are context-conditioned attributes (characterizing place of articulation) since formant transitions follow different trajectories depending on the adjacent vowel.

The features of place of articulation are related to the place in the vocal tract at which a constriction to produce a consonant is formed. There are three general places of articulation - labial, alveolar, and velar – and at each general place there are two or three subsidiary places. Almost any manner (fricative, stop, nasal, etc.) or source (voiced, turbulent, transient) can be employed with any place of constriction. Thus, the possible patterns of acoustic characteristics that result from these many different combinations are varied and can be somewhat complex. However, in spite of such complexity, an investigation on the nature of acoustic attributes of consonant place of articulation will result useful for many applications, among which speech recognition and synthesis. Many research works have investigated on this topic, especially focusing on stop consonants [31, 32, 33, 34, 35, 36]. The starting point of these works is that the acoustic properties that define place of articulation for stop consonants can be derived from the analysis of the short-time spectrum sampled at the consonantal release. In particular, the spectra sampled in the initial few tens of milliseconds following the consonantal release can be used to separate stop consonants into categories according to the place of articulation. Blumstein and Stevens [37] hypothesized that the gross spectrum shape at the consonantal release should show acoustic properties which provide information about consonants place of articulation. These authors suggested that the gross

spectrum shape should be *diffuse-falling* or flat for labials, *diffuse-rising* for alveolar, and *compact* for velars [see 37 for details]. Therefore, they developed a series of templates designed to reflect each of these spectral properties and applied them on the short-time spectra of a large number of natural speech utterances. Overall, about 85% of the utterances were correctly accepted applying these templates and about the same percentage of utterances were correctly rejected. However, these results were not equally distributed across the various vowel contexts and it was shown that the vowel environment does have an effect on the classification of the appropriate place of articulation. Moreover, the acoustic measurements were performed using a fixed time window (26 msec) which may not provide the best measures. Blumstein and Stevens [37] suggested that it may not be desirable to postulate a single, fixed time window, and that the gross spectrum shape may be assessed examining successive spectral samples extending over 10-20 msec, each computed using a relatively short time window.

Esposito [38, 39] hypothesized that the amplitudes of the peaks in the spectrum may be used as acoustic attributes of the place of articulation of the consonants. She studied the properties of the sound spectra at the release of Italian stop consonants in vocalic contexts by applying a short window analysis [3 msec]. On the spectra sampled at the consonantal release, she measured the amplitudes of the peaks in different frequency ranges. The results of the analysis suggested a simple algorithm which can be used to separate stop consonants into categories according to the place of articulation by using simple data such as the values, in dB, of the maximum spectral peaks in different frequency ranges. Moreover, different window sizes (the spectra were computed over 3 msec after the release, averaged over 4 and 10 msec after the release, and using a smoothed spectrum) were also compared in order to determine which seemed to be more effective in keeping the hypothesized spectral shape.

3.1.2 Results on Italian stop consonants.

Esposito derived the set of amplitude attributes by examining the spectral shapes of several hundred utterances spoken by male and female speakers. These sets of amplitude attributes, for each place of articulation (labial, alveolar, and velar) and for the [i, a] vowel context are reported below.

Labial amplitude attributes in the [i] context
a1) The difference between the maximum peak in the 0-2kHz and the maximum peak in the 4-6kHz frequency ranges must be greater than 1dB;
b1) The difference between the maximum peak in the 1-7kHz and the maximum peak in the 0-2kHz frequency ranges must be lower than 9dB;
c1) The difference between the maximum peak in the 1-3kHz and the maximum peak in the 5-7kHz frequency ranges must be greater than 8dB.

Labial amplitude attributes in the [a] context
a2) The difference between the maximum peak in the 0-1kHz and the maximum peak in the 2-3kHz frequency ranges must be greater than 2dB;
b2) The difference between the maximum peak in the 0-1.2kHz and the maximum peak in the 5-7kHz frequency ranges must be greater than 6dB;
c2) The difference between the maximum peak in the 0-2kHz and the maximum peak in the 5-7kHz frequency ranges must be greater than 6dB.

Alveolar amplitude attributes in the [i] context:
a3) The difference between the maximum peak in the 1-3kHz and the maximum peak in the 5-7kHz frequency ranges must be lower than 9dB;
b3) The difference between the maximum peak in the 1-7kHz and the maximum peak in the 0-2kHz frequency ranges must below than 10dB;
c3) The difference between the maximum peak in the 3-5kHz and the maximum peak in the 4-6kHz frequency ranges must be lower than 0dB;
d3) The differences between the maximum peak in the 3-5kHz and the maximum peak in the 5-7kHz frequency ranges must be lower than 9dB.

Alveolar amplitude attributes in the [a] context:
a4) The difference between the maximum peak in the 2-3kHz and the maximum peak in the 3-4kHz frequency ranges must be lower than 8dB;
b4) The difference between the maximum peak in the 0-1kHz and the maximum peak in the 1-3kHz frequency ranges must be greater than -10dB;
c4) The difference between the maximum peak in the 0-1.2kHz and the maximum peak in the 4-6kHz frequency ranges must be lower than 13dB;
d4) The difference between the maximum peak in the 0-1kHz and the maximum peak in the 2-3kHz frequency ranges must be lower than 8dB;
e4) The difference between the maximum peak in the 0-1.2kHz and the maximum peak in the 1.5-3.5kHz frequency ranges must be lower than 6dB.

Velar amplitude attributes in the [i] context
a5) The difference between the maximum peak in the 0-2kHz and the maximum peak in the 4-6kHz frequency ranges must be lower than 2dB;
b5) The difference between the maximum peak in the 1-7kHz and the maximum peak in the 0-2kHz frequency ranges must be greater or equal to 9dB;
c5) The difference between the maximum peak in the 3-5kHz and the maximum peak in the 4-6kHz frequency ranges must be greater or equal to 0dB;
d5) The difference between the maximum peak in the 3-5kHz and the maximum peak in the 5-7kHz frequency ranges must be greater or equal to 0dB.

Velar amplitude attributes in the [a] context
a6) The difference between the maximum peak in the 0-1.2kHz and the maximum peak in the 1.5-3kHz frequency ranges must be lower than -1dB;
b6) The difference between the maximum peak in the 1-2kHz and the maximum peak in the 0-1kHz frequency ranges must be greater than 1dB;

c6) The difference between the maximum peak in the 2-3kHz and the maximum peak in the 3-4kHz frequency ranges must be greater than 1dB;
d6)} The difference between the maximum peak in the 1-2kHz and the maximum peak in the 6-7kHz frequency ranges must be greater than 11dB.

The attributes defined above were the same for both voiced and voiceless consonants. In order to determine the extent to which naturally produced stop consonants fit the hypothesized spectrum shape reflected by the set of amplitude attributes, the spectra of a number of voiced and voiceless consonants produced by several speakers were analyzed.

3.1.3 The data

The recording and measurements were made at the Research Laboratory of Electronics, Speech Communication Group, MIT, Cambridge, USA. The materials consisted in VCVC utterances produced by seven adult Italian speakers (three females and four males) in a sound-treated room and recorded on a high-quality magnetic tape recording system. The utterances were embedded in a carrier phrase. Subjects were asked to read a listing of VCVC utterances containing three repetitions of each of the stop consonants [p, t, k, b, d, g] in the context of the seven vowels [i, e, ∈, a, ⊃, o, u], producing a total of 882 utterances. The measurements were made for the intervocalic consonant. The results reported in the present paper are derived from the analysis of the stop consonants in the [i, a] context. The spectral representations included a DFT spectrum, a smoothed DFT, a spectral averaging method. The analysis window (Hamming window) was set to 3.1 msec. The spectrum at the consonant release, the spectrum averaged over 4 msec (for [b, d, g]) and over 10 msec (for [p, t, k]) after the release and, the *k-averaged* [1]spectrum was computed using a software program developed by Klatt [40]. All spectra were pre-emphasized, and the spectral amplitudes were enhanced modifying a spectral gain control parameter. The amplitudes of the maximum peaks in different frequency ranges were measured by visual examination.

3.1.4 The amplitude attribute analysis

The spectra of 252 utterances were individually tested against the sets of amplitude attributes. A conservative strategy was adopted for assessing whether the spectral shapes were accepted or rejected by a particular set of amplitude attributes. In order to fit a set of attributes, the spectrum had to meet all the conditions specified by the set. If it did not for any reason, the spectrum was rejected. A tabulation was made for each speaker of the proportion of spectra which fit or were rejected by each set of amplitude attributes. Tables V and VI show a summary of the results by applying

[1] The *k-averaged* spectrum was computed by measuring the VOT length of the voiceless consonant. The cursor was then placed on the waveform at the temporal sampling point corresponding to half the VOT length, and the spectrum was averaged over 5 msec to the left and 5 msec to the right of this sampling point.

the sets of amplitude attributes to voiced and voiceless consonants in the context of the vowels [i, a].

Table V: Amplitude results for labial, alveolar, and velar consonants in the [i] context. The entries give the mean percentage of utterances for each consonant that was correctly accepted or rejected by the set of amplitude attributes defined above. Also reported are the different spectra on which the amplitude attributes were applied.

Spectrum at release

Labial Attributes			**Alveolar Attributes**			**Velar Attributes**		
Correct acceptance	Correct rejection		Correct acceptance	Correct rejection		Correct acceptance	Correct rejection	
[p] 33%	[k]	95.2%	[t] 90.4%	[k]	80.9%	[k] 57.1%	[p]	100%
	[t]	100%		[p]	33.3%		[t]	95.2%
[b] 66.6%	[g]	100%	[d] 80.9%	[g]	100%	[g] 90.4%	[b]	100%
	[d]	95.2%		[b]	90.4%		[d]	85.7%
Averaged Spectrum								
[p] 90.4%	[k]	100%	[t] 85.7%	[k]	95.2%	[k] 95.2%	[p]	100%
	[t]	90.4%		[p]	90.4%		[t]	95.2%
[b] 57.1%	[g]	100%	[d] 66.6%	[g]	90.4%	[g] 80.9%	[b]	90.4%
	[d]	100%		[b]	90.4%		[d]	78.2%
k-Averaged Spectrum								
[p] 90.4%	[k]	100%	[t] 95.2%	[k]	90.4%	[k] 95.2%	[p]	95.2%
	[t]	100%		[p]	95.2%		[t]	95.2%

3.1.4 Discussion

The results reported showed that the amplitudes of the peaks in the spectrum obtained by examining successive spectral samples over 10 msec and computed using a short time window (3 msec) give useful information to discriminate among voiceless consonants [p, t, k] (see Tables V, VI). Overall, about 95% of the utterances were correctly accepted by these sets of amplitude attributes and about the same percentage of utterances were correctly rejected. The information obtained by the set of acoustic attributes can easily be translated into an automatic algorithm which discriminates successfully among [p, t, k], and which is simpler than the templates defined by Blumstein and Stevens [37]. Moreover, by applying the templates defined by Blumstein and Stevens on the same spectra, the discrimination performance obtained was less satisfactory [see 38 for comparisons]. This result was expected in the case of alveolar consonants, because of the different point of

constriction of Italian [t, d] with respect to American [t, d]. However, for labials and velars, the results obtained suggest some language specific influences on the gross shape of the spectrum. Moreover, in defining the above sets of amplitude attributes, Esposito considered the effects of the vowel contexts, whereas Blumstein and Stevens did not, since the templates were applied to consonants in different vowel contexts.

Table VI: Amplitude results for labial, alveolar, and velar consonants in the [a] context. The entries give the mean percentage of utterances for each consonant that was correctly accepted or rejected by the set of the amplitude attributes. For [b, d, g] in [a] context the discrimination results are referred to a new set of amplitude attributes, reported in [39].

Spectrum at release

Labial Attributes			**Alveolar Attributes**			**Velar Attributes**		
Correct acceptance	Correct rejection		Correct acceptance	Correct rejection		Correct acceptance	Correct rejection	
[p] 19%	[k]	100%	[t] 90.4%	[k]	95.2%	[k] 100%	[p]	85.7%
	[t]	100%		[p]	19%		[t]	85.7%
[b] 80.9%	[g]	100%	[d] 100%	[g]	100%	[g] 95.2%	[b]	100%
	[d]	85.7%		[b]	52.3%		[d]	100%
Averaged Spectrum								
[p] 90.5%	[k]	100%	[t] 100%	[k]	100%	[k] 95.2%	[p]	100%
	[t]	95.2%		[p]	90.4%		[t]	90.5%
[b] 95.2%	[g]	100%	[d] 85.7%	[g]	100%	[g] 95.2%	[b]	100%
	[d]	38%		[b]	85.7%		[d]	95.2%
k-Averaged Spectrum								
[p] 95.2%	[k]	100%	[t] 90.4%	[k]	85.7%	[k] 80.9%	[p]	100%
	[t]	85.7%		[p]	85.7 %		[t]	95.2%

In fact, the amplitude attributes defined for the consonants in the [i] context gave a good discrimination performance. However, when applied to the same consonants in the [a] context they gave very poor discrimination. In order to improve the discrimination performance, the amplitude attributes should be modified as a function of the vowel context. It is possible to hypothesize that it will be necessary to define a set of amplitude attributes for each vowel or for specific vowel classes sharing the same distinctive features. The above consideration is justified on the basis of the conjecture that since coarticulatory effects change the acoustic parameters from one vowel class to another, there is some anticipatory coarticulation effect which modifies the spectrum shape of the consonant under examination.

In the case of voiced consonants, the set of amplitude attributes allows to discriminate successfully [g] (95% of correct acceptance for both spectrum at the release and spectrum averaged over 4 msec) from [b, d] (about 100% of correct rejection). However, for [b, d] similar information does not work very well. Voicing, which is always present in Italian voiced consonants, causes pressure fluctuations which lead to variability in the peak amplitudes and causes shifts in the vocal tract resonances. Information about formant transitions is thus required.
With regards to the particular spectra computed, it is possible to say that the spectrum during the first 10 msec after the release and the *k*-averaged spectrum seem more useful to retain information about amplitude features when the consonants are voiceless. The spectra at the release retain more information about amplitude attributes of voiced consonants.

Perceptual relevance of acoustic properties based on short-time spectra sampled at consonantal release was reported for American English by Stevens and Blumstein [35] and by Blumstein and Stevens [41]. These authors showed, through a series of perceptual experiments that the characteristics of the gross spectrum shape are utilized by the human speech perception mechanism in order to extract information concerning the place of articulation.
An acoustic and perceptual study carried out on Italian speakers by Delogu [42] reported a number of (asymmetrical) confusions in identifying CV sequences in both synthetic and natural speech. In particular, [ki] and [pi] were found to be confused for [ti] but never the reverse. A preliminary work by Plauché *et al.* [43] attempting to isolate the specific acoustic features that cause these confusions showed that peak amplitudes in the spectrum do play a role; [ki] and [pi] syllables were always confused with a [ti] syllable when their spectral bursts were made similar to that of a [ti]. However, the above results are only preliminary, and further experiments are necessary to assess the acoustic and perceptual importance of these spectral attributes.

3.2 Duration as an acoustic parameter for characterizing stop consonants in single and geminate form.

3.2.1 The problem

Some languages allow the clustering of the same consonant in vowel contexts and this phenomenon is known as "consonant gemination". Gemination plays a particular role in the phonetic of such languages because several words change meaning as a function of singleton versus geminate consonants (minimal pairs). Phonetic theories agree in considering the gemination of a phoneme as a particular realization of the original one [17], which is modified in some of the acoustic parameters. Recent papers [44, 45, 46, 47] report that there is an acoustic relationship between consonant closure duration and gemination as well as between the length of the vowel preceding the consonant and gemination. Moreover, these

studies also report that there is a perceptual relationship between closure duration and gemination, whereas variation in the length of the vowel preceding the consonant does not seem to be perceptually relevant. Other acoustic parameters, which appear to be correlated with gemination [45], are the burst energy and the F0 values at the offset of the vowel preceding the consonant. However, an extensive work on this phenomenon is not available and the results reported are based on a small number of data. A recent work of Esposito and Di Benedetto [48] tried to individuate the acoustic parameters that play a role in the production of a geminate consonant and to validate their perceptual importance. To this end, they settled up a series of experiments for collecting the acoustic data and for synthesizing the stimuli for perceptual tests. A database of geminate and non-geminate utterances with no semantic meaning was built up and the acoustic analysis of such data was carried out. The perceptual properties were investigated through a perceptual experiment.

3.2.2 The Data

The recordings were made in the Speech Laboratory, INFOCOM Department, Rome University ``La Sapienza" (Italy). The measurements were performed using the UNICE version 1.6 by VECSYS speech analysis program, which accepts user commands to read in waveform files and generates spectral displays of various types. The spectral representation used was the DFT (Discrete Fourier Transform) magnitude spectrum. The analysis window (Hamming window) duration was set to a default of 256 samples which corresponds roughly to 26 msec at a sampling rate of 10 kHz. The data consisted of a set of VCV (the non-geminate case) and VCCV (the geminate case) utterances in which the consonant was [b,d,g,p,t,k] (the complete set of stop consonants in Italian) and the vowels [i,a,u]. These disyllabic utterances were chosen because in Italian, many minimal pairs, such as *papa* (pope) and *pappa* (baby food), *fato* (fate) and *fatto* (fact), are disyllabic words; therefore, the use of a disyllabic structure is justified by the natural attitude of the native speakers in producing it. Since different acoustic parameters, among which duration parameters were measured, the utterances were not included in a carrier phrase. Six Italian speakers (three male and three female) produced the utterances. Each utterance was repeated three times for a total of 324 utterances in single form and 324 utterances in geminate form.

3.2.3 Acoustic analyses: measurements and results

A set of measurements in the frequency and time domain were performed. The measurements in the frequency domain include:

a) V1 formant frequencies (F1, F2,F3) at the offset and in the middle of the vowel preceding the consonant;
b) the burst energy, and the VOT energy,
c) the burst power, and the VOT power;
d) the DFT spectra at consonant release.

The measurements in the time domain include:

a1) the duration of the vowel preceding the consonant (V1 duration) and the vowel following the consonant (V2 duration);
b1) the consonant closure duration;
c1) other duration measures, such as VOT and complete utterance duration.

The results obtained from the acoustic measurements showed that:

1) Formant frequency values of the vowel preceding the consonant are not related to gemination, suggesting that no extra vocal effort is needed in a geminate production;
2) There is no relationship between any representation of the energy at consonant release and gemination, in contrast with the general consideration that geminate consonants must show, at the release, greater energy than singletons.
3) The acoustic parameters that appeared strongly related to gemination were the duration of the intervocalic consonant and the duration of the vowel preceding it. Table VII reports averaged V1 and closure duration (for geminate and non-geminate cases) as a function of vowel context and consonantal place.

Table VII: Averaged V1 duration and closure duration in the geminate and non-geminate case. The mean values and the standard deviations (in brackets) for each vowel and place of articulation, computed over speakers (three males and three females) and repetitions (three repetitions for each vowel, for each consonant, and for the single and geminate form).

Context	**V1 duration**		**Closure duration**	
	non-geminate	*Geminate*	*non-geminate*	*geminate*
[a]	179 (26.78)	131 (29.84)	91 (20.75)	189 (37.04)
[i]	159 (26.52)	118 (25.02)	92 (20.38	186 (34.45)
[u]	167 (28.40)	126 (24.26)	89 (18.78)	173 (35.36)
Labials	161 (24.49)	122 (25.77)	102 (17.07)	187 (35.71)
Dentals	175 (26.06)	127 (29.31)	88 (21.17)	195 (34.60)
Velars	169 (32.46)	126 (25.51)	83 (16.56)	166 (34.54)
Total	168 (28.40)	125 (26.93)	91 (19.97)	182 (36.09)

The above parameters were found to be significant also in Hindi geminate consonants [45], suggesting that this effect is language independent. V1 duration in the geminate case was observed to be reduced by about 25% with respect to its duration in the non-geminate case. On the contrary, the closure duration in the geminate case was significantly elongated, by about 50% with respect to the non-geminate case. This result was present for all speakers, vowel contexts, place of articulation, and consonantal voicing, suggesting that closure duration can be

considered a primary acoustic cue for the geminate/non-geminate distinction, whereas, the role played by V1 duration must still be investigated on.

3.2.4 Perceptual experiment: measurements and results

In order to set the perceptual properties of gemination, a perceptual experiment was carried out. The aim of this experiment was to define an average closure duration that overtakes chance in the perception of gemination. Moreover, this experiment was also devoted to investigate the perceptual role played by V1 duration in geminate and non-geminate contexts. The experiment was carried out using /apa/ and /appa/ stimuli.

Stimuli. The stimuli were synthesized as follows: a natural /apa/ token, spoken by one subject, whose duration was close to the average duration of all the subjects, was extracted from the database. The vowel and closure duration in the original stimulus were 176 msec and 99 msec respectively. The digitized signal was then modified by means of a waveform editor (UNICE editor) to produce 2 stimuli, by decreasing the length of V1 from 176 msec (V1 duration in the original token) to 116 msec, in a step corresponding to $5F_0$ periods (60 msec). For each stimulus obtained, 10 new stimuli were produced by increasing the length of the silent portion of the intervocalic consonant ([p]) from 100 to 235 msec in steps of 15 msec. This yielded to a total of 20 stimuli (2 vowel duration x 10 consonantal duration).
Subjects were asked to identify the stimulus words as geminate or singleton. The experiment was run separately for the two vowel duration, with the order of presentation balanced across subjects. For each vowel duration, the 10 closure durations were presented such that each stimulus was preceded and followed once, by every other. Subjects listened to a total of 202 stimuli played in random order via a computer program and delivered through good quality headphones.

Results. The crossover points for each single listener and V1 duration (116 msec, 176 msec) are reported in Table VIII. The data show two different closure duration thresholds at which the perception of a consonant as geminate overtakes chance. The thresholds depend upon the duration of V1. Gemination was enhanced by a shorter V1 duration. In this case, the average closure duration at which the perception of a geminate overtakes chance was about 165 msec. However, a longer closure duration (about 183 msec) was required when V1 duration was longer. The displacement (about 18 msec) in the perceptual threshold for gemination, when V1 duration was longer, was representative of all listeners. As it can be noted, the difference among crossover points is positive for all listeners, which reinforces the observation that when V1 is lengthened the closure duration must be longer to perceive a geminate consonant. Moreover, a two tailed Student-t test applied to the data shows that the differences are statistically significant ($t(19) = 7.882$ at level $p < .001$).

3.2.5 Discussion

Closure duration was the most salient perceptual cue used by listeners to discriminate between geminate and non-geminate consonants. However, it was also an acoustic attribute of the signal (see results on the acoustic data), supporting the hypothesis that closure duration is a distinctive feature for gemination. Since all segment durations which have to act as a cue must be perceived to some baseline, this duration feature was relative rather than absolute. This is a very interesting finding because only duration parameters are found to play a role - whereas duration is often found to be relevant in connection to other acoustic attributes.

Table VIII: Crossover points for each listener and for the two different vowel durations. The change threshold differences (positive for each listener), the mean and the Standard Deviation are also reported.

Listeners	**V1=116 msec**	**V1=176 msec**	**Differences in msec**
AP	172.1	180.2	8.1
Ava	177.5	188.7	11.2
Avi	159.1	180.5	21.4
EC	177.5	188.4	10.9
EZ	159.9	167.5	7.6
FM	170.9	196	25.1
FR	160.1	167	6.9
FS	155.9	170.3	14.4
FT	171.2	188.8	17.6
GS	159	176.9	17.9
IR	166.4	178.8	12.4
Lpa	153.5	171.2	17.7
LPz	189.1	237.1	48
MB	168.4	178.9	10.5
MM	172.4	190.5	18.1
MS	168.6	191.1	22.5
PT	153.4	169	15.6
RR	167.5	193	25.5
RS	153.5	176.4	22.9
SR	159	163.5	4.5
Mean	165.8	182.7	16.9
Standard Deviation	9.5	16.1	9.6

Results (from other languages) on the perceptual role of the closure length in the geminate distinction are in agreement with our findings. Pickett and Decker [49] measured the phoneme boundary between a single and a double /p/ in the pair *topic* and *top pick*. At a speaking rate of six syllables per second, they found the phoneme boundary between *topic* and *top pick* to correspond to a closure duration of about 160 msec, in agreement with the reported crossover point when V1 duration was shorter. Furthermore, Rochet and Rochet [44], and Shrotriya *et al.* [45] showed that

for native speakers of different languages, closure duration is perceptually relevant in the geminate versus singleton distinction.
What about the reduction in the length of V1? From an acoustic point of view, it is expected that before a geminate V1 should be shorter than before a non-geminate because in the former case it is part of a closed syllable and in the latter it is part of an open syllable. However, the perceptual data show that the longer the V1 duration the greater the crossover value (measured in msec) in perceiving a consonant as geminate. This result suggests that the shortening of V1 could not only be attributed to syllable structure but it might be the result of two superimposed effects: one due to syllable structure and the other due to the presence of a geminate consonant. This conclusion is not in disagreement with the findings reported by Rochet and Rochet which showed that Italian listeners distinguished between *fato* and *fatto* on the basis of consonant duration but not on the basis of vowel duration. The conclusion was that V1 is shortened to balance the abnormal lengthening of the closure in order to keep the rhythm constant, and makes the utterance sound natural. This interpretation is also supported by Huggins [50] who found that subjects were much more sensitive to changes in vowel duration than to changes in closure duration. Therefore, these changes were perceived as changes in the sentence rhythm, when the duration of the other segments in the utterance remained unchanged.

4 Conclusions

The present paper contributes to the delineation of the role played by some acoustic attributes of the speech signal in characterizing vowels and consonants in Italian. For speech application problems, such as recognition and synthesis, this information plays a fundamental role in the design of preprocessing algorithms, classification methods, and system structure. The experiments reported are not exhaustive and only few acoustic attributes were taken in consideration. However, they constitute a first step toward the definition of acoustic and perceptual correlates of phonetic features of vowel and consonant segments in the Italian language.

Acknowledgments

Sincere thanks go to Giovanni Flammia with whom the experiment concerning the representation of Italian vowels along auditory dimensions was performed. The authors thank Riccardo Rossetti and Armando Vannucci who contributed, concerning the results reported for the gemination, to the acoustic measurement analyses and to the set-up of the perceptual experiment .

References

1. Honda K.: Relationship between pitch control and vowel articulation. In D.B.Bless and J.H.Abbs (eds) Vocal Fold Physiology: contemporary research and clinical issues. College-Hill Press, 1983, pp.286-297.

2. Stevens K. N.: The quantal nature of speech: evidence from articulatory acoustic data. In P.B.Denes and E.E.David Jr (eds) Human Communication: a unified view. McGraw-Hill, New York, 1972, pp.51-66.

3. Delgutte B. Codage de la parole dans le nerf auditif. Thèse de Doctorat d'Etat, Université Pierre et Marie Curie, University of Paris 6, chap.2, pp.33-78, 1984.

4. Chistovich L. A, Sheikin R, Lublinskaja V.V. Centres of gravity and spectral peaks as the determinants of vowel quality. In B.Lindblom and S.Ohman (eds) Frontiers of speech communication research. Academic Press, London, 1979, pp.143-157.

5. Traunmüller H. Perceptual dimension of vowel openness in vowels. J. Acoust. Soc. Am. 1981; 69: 1465-1475.

6. Di Benedetto M. G. Acoustic and perceptual evidence of a complex relation between F1 and F0 in determining vowel height. Journal of Phonetics 1994; 22.

7. Fant G. Non-uniform vowel normalization. STL-QPSR 1975; 2-3.

8. Verbrugge R.R, Strange W, Shankweiler DP, Edman TR. What information enables a listener to map a talker's vowel space? J. Acoust. Soc. Am 1976; 60: 198-212.

9. Macchi M. J. Identification of vowels spoken in isolation vs vowels spoken in consonantal context. J. Acoust. Soc. Am. 1980; 68: 1636-1642.

10. Syrdal A. K., Gopal H. S. A perceptual model of vowel recognition based on the auditory representation of American English vowels. J. Acoust. Soc. Am. 1986; 79: 1086-1100.

11. Gerstman, L. H. Classification of self-normalized vowels. IEEE Trans. Audio Electroac. 1968; AU-16: 78-80.

12. Stevens K. N., House A. Perturbation of vowel articulations by consonantal context: An acoustical study. J. Speech Hear. Res. 1963; 6: 111-128.

13. Lindblom B. On vowel reduction. R. Inst. Technol. Stockholm, Speech Transmission Lab. Rep. No.29, 1963.

14. Lisker L. On reconciling monophtongal vowel percepts and continuously varying F patterns. Haskins Lab., Stat. Rep. Speech Res. SR-79/80, 1984.

15. Di Benedetto M. G. Vowel representation: Some observations on temporal and spectral properties of the first formant frequency. J. Acoust. Soc. Am. 1989; 86: 55-66.

16. Di Benedetto M. G. Frequency and time variations of the first formant: Properties relevant to the perception of vowel height. J. Acoust. Soc. Am. 1989; 86: 67-77.

17. Muljacic Z. Fonologia della lingua italiana, Bologna, 1972.

18. Lehiste I., Peterson G.E. Transitions, glides and diphtongs. J. Acoust. Soc. Am. 1961; 33: 268-277.

19. Stevens K. N., Keyser S.J., Kawasaki H.: Toward a phonetic and phonological theory of redundant features. In: Perkell J. and.Klatt D.(eds) Symposium on Invariance and Variability of Speech Processes. Lawrence Erlbaum Assoc., Hillsdale, NJ, 1986, pp 426-449.

20. Huang C. B. Perceptual correlates of the tense/lax distinction in general American English. Master's thesis, Massachusetts Institute of Technology, Cambridge (MA) 1985.

21. Di Benedetto M. G. The influence of the first formant frequency trajectory shape on the perception of the tense/lax distinction of American English and Italian vowels. Studi italiani di linguistica teorica ed applicata 1987; anno XVII, no 2-3: 289-297.

22. Di Benedetto M. G., Flammia G. Vowel distinction along auditory dimensions: a comparison between a statistical and a neural classifier. Verba 90, Rome, 1990.

23. Flammia G. Classificazione statistica e neurale su base percettiva nel riconoscimento delle vocali italiane. Tesi di Laurea, Università degli Studi di Roma 'La Sapienza', 1988.

24. Ferrero F. Diagrammi di esistenza delle vocali italiane. Alta Frequenza 1968; 37: 54-58.

25. Disner S. F. Vowel quality: the relation between universal and language specific factors. Ph.D. Dissertation, University of California, Los Angeles, 1983.

26. Hult G. Some vowel recognition experiments using multilayer perceptrons. STL-QPSR 1989; 1: 125-130.

27. Klatt D. H. Software for cascade/parallel formant synthesizer. J. Acoust. Soc.Am. 1980; 67: 971-995.

28. Pickett J. M. The Sounds of Speech Communication, Baltimore: University Park Press, 1980

29. Esposito A. Feature extraction from speech: the acoustic attributes of vowel and stop consonants in Italian. (in Italian), Ph.D thesis, Naples University ``Federico II" 1995.

30. Vannucci A. Acoustic correlates of distinctive features of Italian stops. J. Acoust. Soc. Am., 1994; 95, 2pSP25: 2874.

31. Halle M., Hughes G. W., Radley J. P. A. Acoustic properties of stop consonants. J. Acoust. Soc. Am., 1957, 29:107-116.

32. Fant G. Stops in CV-syllables.In Fant G. (ed.), Speech Sounds and Features. MIT press, Cambridge, MA, 1973, pp 110-139.

33. Jacobson R., Fant G., Halle M. Preliminaries to speech Analysis. MIT press, Cambridge, MA, 1963.

34. Zue V. Acoustic characteristics of stop consonants: a controlled study. Technical Report n. 523, Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1976.

35. Stevens K. N., Blumstein S. E. Invariant cues for place of articulation in stop consonants. J. Acoust. Soc. Am., 1978; 64:1358-1368.

36. Searle C. L., Jacobson J. Z., Kimberly B. Speech as patterns in the 3-space of time and frequency. In Cole R. A. (ed.), Perception and production of fluent speech. Erlabaum, Hillsdale, NJ, 1979.

37. Blumstein S.E., Stevens K.N. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants, J. Acoust. Soc. Am, 1979; 64(4): 1001-1017.

38. Esposito A. The amplitude of the peaks in the spectrum as acoustic attributes of the place of articulation. In Elenius K. and Branderud P. (eds), Proceeding of ICPhS95. Arne Strombergs Grafiska, Stockholm, 1995, vol 1, pp 38-41.

39. Esposito A. The amplitude of the peaks in the spectrum: data from [a] context. In Kokkinakis G. (ed), Proceeding of EUROSPEECH97. University of Patras, 1997, vol 1, pp 1015-1018.

40. Klatt D.H. MIT SpeechVax User's Guide. Copyright 1984 by Dennis H. Klatt.

41. Blumstein S. E., Stevens K. N. Perceptual Invariance and onset spectra for stop consonants in different vowel environments. J. Acoust. Soc. Am., 1980, 67(2): 648-662.

42. Delogu C., Paolini A., Ridolfi P., Vagges K. Intelligibility of speech produced by text-to-speech systems:the amplitude in good and telephonic conditions. Acta Acustica 1995, 3:89-96.

43. Plauché M., Delogu C., Ohala J. J. Asymmetries in Consonant Confusion. In Kokkinakis G. (ed), Proceeding of EUROSPEECH97. University of Patras 1997, vol 4, pp 2187-2191.

44. Rochet L. B., Rochet A. P. The perception of the single-geminate consonant contrast by native speakers of Italian and Anglophones. In Elenius K. and Branderud P. (eds), Proceeding of ICPhS95. Arne Strombergs Grafiska, Stockholm, 1995, vol 3, pp 616-619.

45. Shrotriya N., Siva Sarma A. S., Verma R., Agrawal S. S. Acoustic and perceptual characteristics of geminate Hindi stop consonants. In Elenius K. and Branderud P. (eds), Proceeding of ICPhS95. Arne Strombergs Grafiska, Stockholm, 1995, vol 1, pp 132-135.

46. Rossetti R. Caratteristiche acustiche del fenomeno di geminazione nelle consonanti occlusive Italiane: applicazione all'adattamento automatico di pronunce straniere, Laurea dissertation for Electrical Engineer degree,Rome University "La Sapienza", INFOCOM dept., Via Eudossiana, Rome, Italy, 1993.

47. Rossetti R Gemination of Italian stops, J. Acoust. Soc. Am., 1994, 95, 2pSP26: 2874.

48. Esposito A., Di Benedetto M. G. Acoustical and perceptual study of gemination in Italian Stops. International Institute for Advanced Scientific Studies (IIASS), I9608, Salerno, Italy, 1996.

49. Pickett J. M., Decker L. R. Time factors in perception of a double consonant. Language and Speech, 1960; 3:11-17.

50. Huggins A. W. F. Just noticeable differences for segment duration in natural speech J. Acoust. Soc. Am., 1972; 51(4):1270-1278.