The phonetic nature of the Northern Italian allophones [s] and [z] in words with variable realization: electroglottographic and acoustic evidence[1]

MARCO BARONI, UCLA

1. Introduction

In Northern Italian,[2] the alveolar fricatives [s] and [z] are in complementary distribution. In particular, only the allophone [z] can occur in intervocalic position (as in the examples in (1.a)); only the allophone [s] can occur word-initially before vowels (as in the examples in (1.b)).

(1)	a.	['ka**z**a]	“home”	b.	['**s**anto]	“saint”
		['vi**z**o]	“face”		['**s**ot:o]	“under”
		[ri'**z**ata]	“laughter”		[**s**o'lare]	“solar”

I will refer to the fact that the alveolar fricative occurring in intervocalic position is always [z] with the descriptive label of “intervocalic voicing”, and I will adopt the symbol /S/ to refer to the alveolar fricative “archiphoneme”, not specified for [±voice]. As I showed in Baroni 1997, intervocalic /S/ voicing is an extremely productive phenomenon of contemporary Northern Italian (for example, it applies in the production of nonsense words and recent loanwords). There is, however, a systematic class of exceptions to it, exemplified by the forms in (2):

(2)	[a**s**i'm:ɛtriko]	“asymmetrical”
	[a**s**o'ʧale]	“anti-social”

As the English glosses suggest, in similar cases the alveolar fricative occurs in a special context: these words are formed by a prefix ending with a vowel followed by a stem beginning with /S/. Thus, the examples in (2) indicate that the distribution of [s] and [z] is sensitive to morphological structure: intervocalic voicing is blocked when the vowel preceding /S/ does not belong to the same morpheme. I will refer to this phenomenon as Intervocalic Voicing Blocking (IVB). As shown in Baroni 1997, IVB is also a productive phenomenon of Northern Italian (for example, it takes place in prefixed nonce forms). Consider now the forms in (3):

(3)	[pre'**z**unto]	“presumed”
	[re**z**is'tente]	“resistant”

As the glosses suggest, these could also be considered prefixed words in which the alveolar fricative is stem-initial. However, in these cases intervocalic voicing is not blocked. Intuitively, the reason for this is that the forms in (3) are not morphologically complex from a synchronic point of view. In contemporary Italian, their morphological structure is opaque, and speakers treat them as monomorphemic forms.[3]

[1] Thanks to Bruce Hayes, Pat Keating, Ian Maddieson, Motoko Ueyama, Carson Schütze and, especially, Sun-Ah Jun for advice and help with this project.

[2] With the term Northern Italian, I refer to the variety of Standard Italian spoken in Northern Italy. Northern Italian differs from (Central) Standard Italian only in terms of phonology (intervocalic /S/ voicing is not a systematic property of Standard Italian).

[3] Here and below, when discussing potentially prefixed forms, I often refer to non-prefixed words as monomorphemic. Notice, however, that usually these non-prefixed words are not truly monomorphemic, since they bear, at least, an inflectional suffix. For example, when I claim that

Interestingly, the speakers oscillate between [s] and [z] realizations of a number of words with potentially stem-initial /S/, such as the ones in (4):[4]

(4)	[bi'**s**ɛsto] / [bi'**z**ɛsto]	“referring to leap-year”
	[bi**s**e't:riʧe] / [bi**z**e't:riʧe]	“bisecting (line)”
	[ko'**s**eno] / [ko'**z**eno]	“cosine”

The reason why in these cases speakers tend to oscillate between [s] and [z] is that the morphological status of similar forms is ambiguous: they are not as transparent as the forms in (2), but not as opaque as the forms in (3) (for a more explicit characterization of morphological opaqueness/transparency, see Baroni 1997). Thus, speakers sometimes treat forms such as the ones in (4) as prefixed, and sometimes as monomorphemic. When they treat these forms as prefixed, /S/ is stem initial, and IVB applies, i.e., /S/ is realized as [s]. When they treat them as non-prefixed, /S/ is intervocalic within the same morpheme, and intervocalic voicing applies, i.e., /S/ is realized as [z].

More explicitly, in the model presented in Baroni 1997 words with variable /S/ realization are characterized by double lexical representations. Speakers set up two lexical entries for a morphologically ambiguous word such as *coseno* “cosine”: a prefixed and a non-prefixed representation. In the prefixed representation /S/ is stem-initial, and hence it is specified as [s]. In the non-prefixed representation, /S/ is intervocalic within a morpheme, and hence it is represented as [z].

Borrowing a standard idea from the literature on lexical access (see Massaro 1994 for a review), I assumed that each of the two representations of a morphologically ambiguous word is associated with a certain activation threshold. When the speaker wants to produce such a word, both representations are activated, and the first one that reaches its activation threshold will be the one chosen for production. In the case of words, such as *coseno*, for which the [s] and [z] realizations are equally likely, the activation thresholds of the prefixed and non-prefixed representations are similar, so that the chances of winning the lexical decision race are similar for both forms (again, see Baroni 1997 for further details).

An assumption behind this model is that the distinction between [s] and [z] is categorical even in the case of words with variable /S/ realization. Words with variable /S/ realizations have two representations, one in a morphologically complex format with stem-initial [s], and one in a monomorphemic format with intervocalic [z]. However, there is no reason to believe that the complex and simple representations of a morphologically ambiguous word differ from the representations of unambiguously complex and simple words, respectively.

In this model, while the decision to retrieve the word in the complex vs. simple format may depend on gradient factors, once the choice is made, there is nothing gradient about the retrieved form. The complex representation of *coseno,* for example, is identical (in morphological terms) to the representation of an unambiguously prefixed form, such as *asimmetrico* “asymmetrical”, and thus the stem-initial /S/ of both items is specified as [s]. The simple representation of *coseno* is identical to the representation of an unambiguously non-prefixed form, such as *presunto* “presumed”, and thus the intervocalic /S/ is specified as [z] in both items.

One could conceive alternative models in which morphological representations *per se* are gradient and, consequently, the /S/ in ambiguous forms is realized as an intermediate sound between [s] and [z], or, at least, in ambiguous forms with variable /S/, the boundary between the [s] and [z] realizations is blurry.

In this paper, I present electroglottographic and acoustic data supporting the claim that the distinction between [s] and [z] is always categorical, i.e. that even the least voiced tokens of [z] are

speakers treat *presunto* as a monomorphemic word, I mean that they do not treat it as a prefixed form -- but still it is likely that they are aware of the fact that the final *-o* is the masculine singular suffix. Thus, *monomorphemic* is used here as a synonym of *non-prefixed*.

[4] In Baroni 1997 I present and discuss the empirical evidence supporting the claim that forms such as the ones in (4) are produced with large within and between speakers variation.

still significantly more voiced than the most voiced tokens of [s], even in the case of variable /S/ realization of the same word.

2. Preparation of the experiment

In order to support the claim that the distinction between [s] and [z] is always categorical, I designed a corpus that could allow me to compare instances of the following classes:

- word-initial [s];
- morpheme-internal intervocalic [z];
- intervocalic [s] of a morphologically complex nonce formation, where IVB always occurs;
- [s] and [z] of a word with variable realization;
- intervocalic [z] of a morphologically complex word with prefix-final /S/.[5]

Since it was crucial to compare [s] and [z] of a word with variable /S/ realization, I needed to elicit from the subjects a certain number of [s] and [z] tokens of the same word. Thus, I selected five words for which, in an earlier survey, I recorded a considerable amount of within speaker variation between [s] and [z]:

(5)	[bi+/S/es'tile]	“adjective referring to leap-years”
	[bi+'/S/ɛsto]	“adjective referring to leap-years”
	[ri+'/S/ak:a]	“undertow”
	[ri+'/S/alta]	“she/he stands out”
	[ri+/S/ar'ʧibile]	“that can be indemnified”

These words contain two prefixes (*bi-* and *ri-*) that are productive in contemporary Italian, and thus it was possible to match them with prefixed nonce formations.

In order to allow the comparison among the classes listed above, each of the words in (5) was matched with: a clitic + [s]-initial word sequence; a word with morpheme-internal intervocalic [z]; a nonce word formed by a prefix and a stem beginning with [s]; a nonce word formed by a prefix ending with [z] followed by a stem beginning with a vowel.

Within each set, each word (or clitic-word sequence) had the same number of syllables, stress fell on the same syllable, /S/ occurred in the same location, and it was surrounded by the same vowels.

The set of forms associated to [bi+'/S/ɛsto] follows (this is the only set that was actually used, as we will see):

(6)[6] *word with variable /S/:*
[bi+'/S/ɛsto]
nonce word formed by prefix + stem beginning with [s]:
[bi++'sɛrʤo] “double Sergio”[7]
(Sergio = proper name)
nonce word formed by prefix ending in [z] + stem:
[diz++ɛt:sja] “she/he de-Ezio-ifies”
(Ezio = proper name)
clitic + word beginning with [s]:

[5] Prefix-final intervocalic /S/, unlike stem-initial /S/, is always realized as /z/, i.e. it is always subject to intervocalic voicing.

[6] The boundary symbols used in these transcriptions are explained in (7) below.

[7] The speakers were invited to think of a couple of very close friends, both named Sergio, who were collectively referred to, by other friends, as “il Bisergio”. Both speakers found the nonce formation weird but acceptable.

	[di##'sɛl:a]	“of saddle”[8]
morpheme internal intervocalic [z]:		
	[mi'zɛrja]	“misery”

Six lists were prepared. The first list included the five words in (5) and eight fillers. Each of the other lists included the four forms associated with one of the words in (5) and eight fillers (for example, one of the lists contained the four forms in (6) and eight fillers). Subjects had to read the first list, and subsequently the list corresponding to the stimulus word of the first list that they produced with the most variation.

This two-stage procedure could in principle cause ordering effects, but the alternative would have been to present the subjects with 14 repetitions of a single list of 25 stimuli and 48 fillers (1022 tokens in total), which would have been too long for EGG data collection (we need many repetitions in order to get enough tokens of both [s] and [z] in the variable /S/ cases).

One fourth of the fillers consisted of lexicalized prefixed words, one fourth of prefixed nonce formations, one fourth of non-prefixed simple words, one fourth of clitic-word sequences. No filler contained /S/.

The stimuli and the fillers were embedded in the carrier sentence “Dico ____ di nuovo” (“I say ____ again”).

The sentences were presented on a computer screen.[9] Each sentence stayed on the screen for 2500 msec. The sentences were separated by 500 msec intervals. Subjects were presented with 2 repetitions of a training set consisting of 10 items, then with 14 repetitions of the first list and (after a pause) with 12 repetitions of the second list (each time, the sentences were presented in a different random order). Before the first presentation of each list, 5 extra fillers appeared on the screen, in order to ensure that the stimuli would never occur at the beginning of the block.

The subjects were made familiar with the list of the words that they had to read before each reading session, and it was made sure that they understood and found the nonce forms acceptable.

3. Administration of the experiment and data analysis

Two subjects took part in the experiment. They wore an electroglottograph (EGG)[10] and a head-mounted microphone. The subjects were recorded in the sound booth of the UCLA Phonetics Laboratory. The EGG and acoustic signals were recorded to an audio tape, and subsequently digitized at a 16 kHz sampling rate and analyzed using Kay CSL software.

Since the first subject did not show any variation between [s] and [z] in the reading of the first list (all words were consistently produced with [z]), he was not asked to read the second list and his data were not analyzed.

The second subject produced the word *bi/S/esto* 4 times with [s] and 8 times with [z], and he did not show any variation in the production of the other forms in the first list.[11] Consequently, the second list presented to this subject contained the words matched with *bi/S/esto*. The statistics

[8] While *of* is the literal translation of the preposition *di*, in this case *from* may be a more appropriate English equivalent, since the phrase *di sella* is typically used as part of the expression *cadere di sella*, which means ‘to fall from saddle’.

[9] The computerized stimuli for this experiment were prepared and presented using PsyScope (Cohen, MacWhinney, Flatt & Provost 1993).

[10] The EGG allows the investigator to detect whether the glottis is open or closed: when the vocal folds are in contact, the EGG will show low impedance; when they are not in contact, the EGG will show high impedance. Thus, EGG data can be very useful to compare a voiced sound such as [z] (characterized by the presence of vocal fold vibration, i.e., periodic vocal fold contact) with a voiceless one, such as [s] (characterized by the absence of vocal fold vibration).

[11] Probably, the reason why subjects hardly produced any variable forms is that the list they had to read was relatively short, and thus they were able to consistently produce each target word with the same voicing value across repetitions.

presented below are computed on the basis of all 4 [s] tokens of *bi/S/esto* and 6 randomly selected tokens of each other category.

I measured the following three voicing-related properties:

- Proportion of Voicing: ratio of the voiced portion of the fricative (the portion of the fricative corresponding to some vibration on the EGG display) to the overall duration of the fricative.
- Energy of Voicing: ratio of the maximum amplitude value of a glottal cycle at the center of the fricative[12] to the maximum amplitude value of a glottal cycle at the center of the following vowel (the central portions of the fricative and vowel were identified on synchronized acoustic waveform and wide-band spectrogram displays).
- Duration: measured on the acoustic waveform and on a synchronized wide-band spectrogram (the point corresponding to a sudden weakening of F3 on the spectrogram was considered the onset of the fricative; the onset of the first periodic wave after the noise on the acoustic waveform display was considered the offset of the fricative)

4. Results

In this section, I will refer to the various [s] and [z] classes using the following symbols:

(7) +s = [s] realization of variable /S/: [bi+'**s**ɛsto]
+z = [z] realization of variable /S/: [bi+'**z**ɛsto]
++s = stem-initial [s] in nonce prefixed form: [bi++'**s**ɛrʤo]
z++ = prefix final [z] in nonce prefixed form: [di**z**++ɛt:sja]
##s = word initial [s]: [di##'**s**ɛl:a]
z = morpheme internal intervocalic [z]: [mi'**z**ɛrja]

4.1 Proportion of Voicing

Table 1 reports the mean, standard deviation and minimum and maximum values of Proportion of Voicing for each category:[13]

Category	Word/Phrase	Mean	SD	Min	Max
+s	[bi+'**s**ɛsto]	.2633	.0758	.1972	.3613
+z	[bi+'**z**ɛsto]	1	0	1	1
++s	[bi++'**s**ɛrʤo]	.1693	.0422	.1009	.2246
z++	[di**z**++ɛt:sja]	1	0	1	1
##s	[di##'**s**ɛl:a]	.1608	.0163	.1380	.1837
z	[mi'**z**ɛrja]	1	0	1	1

Table 1: proportion of voicing

The data support the claim that the distinction between voiced and voiceless categories is always categorical: all the tokens of all the voiced categories (+z, z++, z) have a voiced portion / overall duration ratio of 1, i.e. there is no trace of devoicing, whereas even the most voiced voiceless token (the maximal value of the category +s) has a voiced portion/overall duration ratio of approximately 1/3 (the voiced portion of the voiceless tokens always occurs at the beginning).

The following figures display synchronized acoustic and EGG waveforms of sample +s and +z tokens, respectively.

[12] In the cases in which there was no sign of vibration at the center of the fricative, the value entered was 0.

[13] The statistical analyses reported in this paper were conducted using the SPSS 6.1 package.

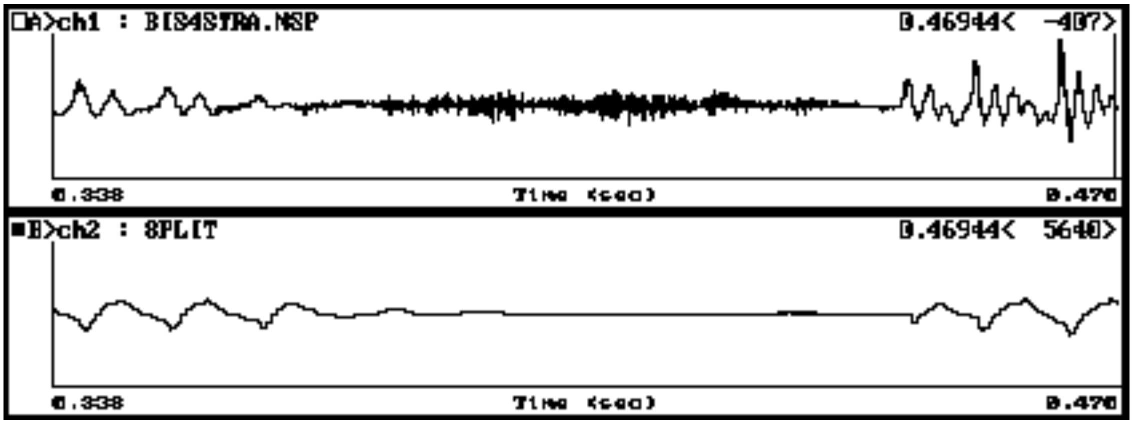

Figure 1: Acoustic waveform (first window) and synchronized EGG signal (second window) corresponding to a +s token.

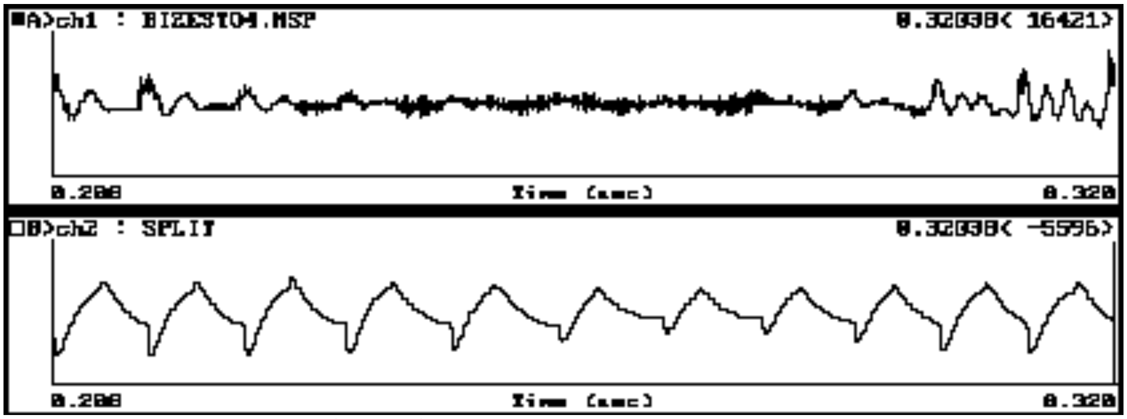

Figure 2: Acoustic waveform (first window) and synchronized EGG signal (second window) corresponding to a +z token.

The results in Table 1 also suggest that +s is more voiced than the other voiceless categories. However, this is a consequence of the fact that the +s tokens are on average shorter than the ++s and ##s tokens (see 4.3 below), while the amount of progressive (forward) voicing affecting the tokens of each class is constant.

Table 2 shows that the difference in the absolute duration of the voiced portion of the three voiceless classes is rather small (duration values are expressed in msec):

Category	Word/Phrase	Mean	SD	Min	Max
+s	[bi+'**s**ɛsto]	21.9	4.7	15.4	26.3
++s	[bi++'**s**ɛrʤo]	17.8	5.1	10.1	24.8
##s	[di##'**s**ɛl:a]	18.7	3	15.9	23.4

Table 2: duration of the voiced portion of the voiceless classes (in msec)

I ran an ANOVA comparing the groups of Table 2. As expected, the difference among groups is not statistically significant (F (2, 13) = 1.17, p = .3417).

4.2 Energy of voicing

Table 3 reports the mean, standard deviation, minimum and maximum value of energy of voicing for each category (these are proportional values, since the peak amplitude of the central cycle of each fricative is divided by the peak amplitude of a cycle in the middle of the following vowel):

Category	Word/Phrase	Mean	SD	Min	Max
+s	[bi+'**s**ɛsto]	.0	.0	.0	.0
+z	[bi+'**z**ɛsto]	1.1771	.1912	.8535	1.3954
++s	[bi++'**s**ɛrʤo]	.0	.0	.0	.0
z++	[di**z**++ɛt:sja]	1.3313	.2418	1.0324	1.7711
##s	[di##'**s**ɛl:a]	.0	.0	.0	.0
z	[mi'**z**ɛrja]	1.0058	.1028	.8985	1.1334

Table 3: energy of voicing

Obviously, since no voiceless token is voiced for more than one third of its duration, no voicing energy in the center of the fricative is recorded for any voiceless class. Again, this contrasts sharply with the characteristics of the voiced classes: the minimum value amongst the voiced categories (an instance of +z) is .8535, which means that even in this case the voicing energy of the fricative is almost as high as that of the following vowel.

The mean voicing energy values of the voiced classes are quite similar, although the mean of the z++ class is rather high, and that of z is rather low. An ANOVA comparing the voiced classes revealed the presence of significant differences (F (2, 15) = 4.52, p = .0291). The Scheffé post hoc test indicated that the difference between z++ and z is statistically significant (α =.05). This fact is quite surprising: why should the energy of a prefix final [z] be higher than the energy of a morpheme internal intervocalic [z]? Possibly, this is simply due to the fact that our measure of energy of voicing, while useful to distinguish broad categories, such as voiced vs. voiceless, is not very reliable in detecting finer distinctions.

4.3 Duration

Table 4 reports the mean, standard deviation and minimum and maximum duration values of each category:

Category	Word/Phrase	Mean	SD	Min	Max
+s	[bi+'**s**ɛsto]	84.9	13.5	72.8	103.7
+z	[bi+'**z**ɛsto]	60.9	7.1	48.9	68.3
++s	[bi++'**s**ɛrʤo]	108	11.3	93.4	122.3
z++	[di**z**++ɛt:sja]	70.3	8.1	62	85.2
##s	[di##'**s**ɛl:a]	116	10.4	99.7	127.4
z	[mi'**z**ɛrja]	63.1	9.3	54.1	78.6

Table 4: duration

The mean duration of each voiceless category is higher than that of any voiced category. Notice however that the difference between the voiced and voiceless class along this parameter is not as sharp as the difference in terms of the proportion and energy of voicing. This is probably due to the fact that, while proportion and energy of voicing are direct measures of voicing, duration is only an indirect cue. While it is usually the case that, since a short duration makes voicing easier, voiced consonants are shorter than voiceless ones (see Lisker 1978, among others), voicing (i.e. vocal cord vibration) and duration are distinct properties. Thus, it is not surprising that the duration distinction between [s] and [z] is not as sharp as the distinction in proportion and energy of voicing. When we measure duration as a cue of voicing, we are not measuring voicing *per se*: the duration contrast is rather a “side effect” of the voicing distinction.

It is interesting that the mean duration of +s is considerably lower than the mean duration of the other two voiceless categories. I ran an ANOVA that detected the presence of significant differences (F (5, 28) = 34.01; p < .001). The Scheffé post-hoc test indicated that +s is only significantly different from the shortest voiced category (+z), whereas ++s and ##s significantly differ from all the voiced categories and from +s.

Since +s does not differ from ++s and ##s in terms of voicing (as shown by the data on the duration and energy of voicing), it is not plausible that the difference in duration between +s and the other voiceless categories depends on different degrees of voicing. Rather, this difference could follow from the fact that forms with a lexicalized meaning, such as *bisesto*, must be stored in the lexicon as single units, while prepositional phrases and prefixed nonce formations are likely to be assembled on line. Thus, the [s]’s in *Bisergio* and *di sella* are word-initial, in the sense that these forms are created by juxtaposing the preposition *di* and the prefix *bi-* to independent words, whereas in *bisesto* the [s] is word internal. I conjecture that the longer duration of [s] in *Bisergio* and *di sella* reflects the commonly observed phenomenon of “domain initial strengthening” (see Jun 1993, Fougeron & Keating 1997 and the references quoted there).

5. Conclusion

The claim that the distinction between [s] and [z] is always categorical is strongly supported by the EGG data on the proportion and energy of voicing. All the instances of [z] have a voiced portion to overall duration ratio of 1, whereas the [s] tokens only show some (weak) trace of voicing assimilation to the preceding vowel. Similarly, the energy of the central glottal pulses of all the [z] tokens is close to the energy of the glottal pulses of the following vowel (and often higher), whereas there is not even a trace of vocal fold vibration in the central portion of any [s] token.

In average, the [s] tokens are also longer than the [z] tokens, but the most interesting fact emerging from the duration measurements is that the +s category is significantly shorter than the ++s and ##s categories. We attributed this difference to the fact that the +s tokens are instances of word internal [s], whereas ++s and ##s tokens are instances of word initial [s].

The results show that, even in the ambiguous cases, morphological uncertainty is not reflected in the phonetic production of [s] and [z]: once the speaker decides to produce one of the two phones, the chosen phone is produced as a fully voiceless or voiced sound.

This in accordance with the model presented in Baroni 1997, in which variability in /S/ realization is not the consequence of gradient morphophonological representations, but it follows from the fact that morphologically ambiguous words have two lexical entries: one in which the word is represented as prefixed, with a stem-initial [s], and one in which the word is monomorphemic, with intervocalic [z].

REFERENCES

Baroni, M. 1997. The representation of prefixes in the Italian lexicon: Evidence from the distribution of [s] and [z], MA thesis, UCLA.

Cohen, J., B. MacWhinney, M. Flatt & J. Provost 1993. PsyScope: an interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers, Behavior Research Methods, Instruments and Computers 25: 257-271.

Fougeron, C & P. Keating 1997. Articulatory strengthening in prosodic domain-initial position, Journal of the Acoustical Society of America 106: 3728-3740.

Jun, S. 1993. The phonetics and phonology of Korean prosody, Ph.D. dissertation, OSU.

Lisker, L. 1978. Rapid vs. Rabid: A catalogue of acoustic features that may cue the distinction, Status Report on Speech Research 54: 127-132.

Massaro, D. 1994. Psychological aspects of speech perception, in M. A. Gernsbacher (ed.) Handbook of psycholinguistics, New York: Academic Press: 219-263.