
VILE: Acoustic Study of Inter and Intra Speaker Variation in Spanish

Elena Battaner, Juana Gil, Victoria Marrero
Departamento de Lengua Española y Lingüística General, Universidad Nacional de Educación a Distancia

Joaquim Llisterri, Carme Carbó, María Jesús Machuca, Carme de la Mota, Antonio Ríos
Departamento de Filología Española, Universidad Autónoma de Barcelona

vile@liceu.uab.es

ABSTRACT

The aim of the VILE project is the acoustic phonetic analysis of inter and intra-speaker variation in Spanish with particular emphasis in obtaining useful results for automatic speaker identification and for forensic phonetic practices. The first phase of the project –presented in this paper– has consisted in a review of the literature dealing with the acoustic parameters which are relevant for speaker identification, and in an analysis of existing spoken corpora in Spanish from which the data analysed in the project will be taken.

1. THE VILE PROJECT

VILE (Estudio acústico de la variación inter e intra en español - Acoustic Study of Inter and Intra Speaker Variation in Spanish) is a research project whose main purpose is the acoustic analysis of phonetic variation in Spanish in order to apply its results to the automatic recognition of speakers and to the practice of forensic phonetics. VILE seeks three aims: (i) to characterise from an acoustic point of view segmental and suprasegmental elements that contribute to a speaker’s individuality, as opposed to those that are common to a speech style, a geographic or social variety, or to a language; (ii) to obtain the phonetic knowledge needed to improve automatic systems for speaker recognition, identification, or verification; and (iii) to provide forensic phoneticians with new acoustic data to allow the comparison of known and unknown speakers with a higher degree of certainty.

The first step of this project has consisted on the delimitation of targeted phonetic phenomena, as described in section 2. In order to achieve it, a review of the publications in three main areas has been carried out, considering segmental and suprasegmental aspects: acoustic phonetic studies of Spanish, speaker recognition works based on phonetic parameters, and forensic phonetics literature. Taking into account the need for reusing resources, an analysis of existing Spanish speech corpora has been also accomplished, assessing their usefulness related to the three objectives of this project. Concluding remarks of this evaluation are listed in section 3.

The next steps of the VILE project will be centred on the acoustic analysis of phonetic phenomena selected from samples extracted from the chosen corpora. The results obtained will be studied from a variability point of view -both inter-speaker variation and intra-speaker variation-, taking into account the relation between both types of variability and their pertinence in automatic speaker recognition and forensic phonetics.

1 The project is financially supported by the Ministry of Science and Technology (BFF2001-2551, 2001-2004). For more information, see: http://liceu.uab.es/~joaquim/VILE.html
2. SELECTION OF PHONETIC PHENOMENA

2.1. Relevant parameters in the study of voice individuality

Research on speaker identification centres around three types of study: (a) those based on the visual recognition of spectrograms (i.e. Tosi et alii, 1972); (b) those dealing with the perceptual recognition of the speaker (i.e. Pollak et alii, 1954; Compton, 1963; Stevens et alii, 1968; Hollien et alii, 1982; Kuwabara and Takagi, 1991, Kreiman and Papcun, 1991; Pisoni 1993); and (c) those concerned with automatic speaker recognition, which in fact constitute the larger part of the works consulted for this project (i.e. Atal 1972; Wolf, 1972; et cetera).

It seems convenient to start with a critical reflection -often repeated in the literature- on the works included under (a). That is to say, on the use of a presumable objective technique employed in the visual interpretation of wide bandwidth spectrograms. It is the so-called “voiceprint” technique that, as pointed out by Künzel (1995), it is still applied in Spain. Outstanding features employed when applying this technique are –albeit they are not listed by any author– formant bandwidths, their central frequencies, or the spectral composition of fricatives and plosives. Visual similarities of these characteristics are judged based on the assumption that inter-speaker differences are larger than intra-speaker ones. However, this fact is not always confirmed by the spectrograms: it has been demonstrated that this technique’s margin for error is high, as it displaces the subjectivity involved in auditory or perceptual judgement to the visual one. These are the reasons for not considering in our project research belonging to (a).

This project mainly deals with works included in (b) and (c), that is to say, with the analysis of procedures where the influence of a human factor, nevertheless complementary, is limited; and where objective and specific parameters are considered. Doddington (1985) suggests a first division between high-level information parameters –such as dialect, style, et cetera-, and low level information ones: spectral amplitude, pitch, formant frequencies, and other acoustic features. First group corresponds to the so-called social and psychological dimension (Kuwabara and Sagisaka 1995), that is to say, all features that depend on social, economical, geographical, educative, psychological, sexual, or linguistic factors. The second group, on the other hand, is the one related to the physiological dimension, the one approached in this research. K.N. Stevens (1971) pointed out, amongst those features especially susceptible of being employed in the identification and discrimination between speakers, fundamental frequency (F0), and the glottal waveform -which is very different from speaker to speaker- as aspects referred to the source, and a set of characteristics related to the filter:

Formant frequencies. If mean values of formant frequencies in a wide enough number of vowels are considered, evidence of the speaker’s mean length of the vocal tract can be obtained. For this purpose, the F3 mean value is particularly useful, as it does not change considerably from vowel to vowel and gives more accurate information than F1 and F2 do. As vocal tract length increases, F3 frequency decreases.

Formant bandwidths. It is especially interesting to compare results of vowel [i], as it does not differ much in the case of a single speaker but presents clear differences between a group of speakers.

Turbulent noises. In the case of [s], there are certain intra-speaker differences, but they are clearly less marked than those inter-speaker. High frequency resonance of the vocal tract excited in the production of a turbulent noise depends on the shape of the cavities before the constriction and on the shape of the palate and tongue after it.
Nasal consonants. As in the former case, these present marked spectral inter-speaker differences, although Stevens himself acknowledged that analysed samples were very near in time. Probably, these characteristics could be altered with temporarily distant samples.

Hollien (1990, 1991) supported some of these characteristics and pointed out other features that, in his opinion, present high probability of being decisive when discriminating between speakers:

**Long Term Spectrum**: Especially useful with normalized data obtained in laboratory, it is also very resistant to the effect of stress on speech. Within his system, it is the result of the analysis of forty parameters extracted from the signal.

The vocalic formant Vector. It is an important parameter for the identification of speakers, as the individual vocal tract presents stability and because these features are resistant to distortion and interferences. Early work by Ladefoged and Broadbent (1957) is still interesting, since they also defend this parameter. Hollien, after reviewing the literature on this topic, chose two parameters to build up this vector: central frequencies of the first three formants (that seem especially revealing if at least three vowels [a i u] and syllable [na] are studied), and the distances between such formants (F1/F2, F2/F3) as they cannot be altered willingly (also Tosi et alii, 1972).

The temporal Vector. This vector has not been analysed in detail so far (Johnson et alii, 1984), but logically, it can be an important factor for speaker identification. Measures employed to build up this vector include: (i) total time of speech, defined as time in milliseconds to produce the emission of a given set of syllables; (ii) the proportion of this speech time, considered as the total time that contains acoustic energy within the emission; (iii) the proportion of silence intervals; (iv) speaking rate, taken from the measure of the completed syllables during a fixed time; and (v) the consonant/vowel duration ratio, that is to say, the relation between the time employed in the production of a consonant and the one employed in the production of a vowel within a given emission of CV.

\( F_0 \) Vector. In Hollien et alii’s work this vector implies the measure of thirty different parameters; in doing so, these authors hold that results are more reliable than those offered in former works.

All these vectors together give a complex profile of the speaker, based on natural data, i.e. extracted from the speech signal.

Other works as Atal (1972), Eskenazi et alii (1990), Kuwabara and Takagi (1991), and Kuwabara and Sagisaka (1995) approximately mention the same parameters as responsible for voice individuality. The summing up of these studies, and those quoted within, can be the following:

1. Parameters referred to the source: Mean value of \( F_0 \), shape of glottal wave, and \( F_0 \) fluctuation.
2. Parameters referred to the resonant cavity: Formant frequencies; formant bandwidth; formant trajectories; distances and ratios between formants; LTAS (Long Term Average Spectrum); turbulent noises; nasal consonants; coarticulatory effects (vowels, nasals, and liquids).
3. Temporal variables: total time of speech; silence and speech intervals ratio; speaking rate.

When selecting the parameters for the study, Wolf’s (1972) reflections on valid decision criteria can be taken into account:


(a) They must be parameters naturally and frequently present in normal speech. Coarticulatory effects—as they are, to a certain extent, “learned”—should not be accounted (on a different point of view, see Su et alii, 1974).

(b) They must be easy to measure.

(c) They should have the highest possible inter-speaker variability and the minimal intra-speaker variability. The shape of the glottal wave and turbulent noises seem to fulfill those conditions.

(d) They should not vary much in time nor to be affected by the speaker’s psychophysiological conditions. For this reason nasals could be eliminated (however, see Glenn and Kleiner (1968) and Wolf (1972), where nasals are considered very relevant). In spontaneous speech, a feature especially sensitive to stress is $F_0$, on the other hand, LTAS seems very resistant (Pittman, 1987; Hollien, 1990).

(e) These parameters must be resistant to ambient noise, and they should not result affected by conditions of transmission. Mean values of vocalic formants ($F_3$—that gives an indication of the vocal tract length—, $F_2$, and $F_1$) and their bandwidth seem very valuable in this sense, as they resist distortion and interferences. The study of extreme vowels [a i u] is advised. On the other hand, it is pointed out that $F_1$ and $F_2$ values of such vowels are the more stable and the less sensitive to context (Stevens and House, 1963).

(f) Finally, the speaker should not be able to modify these parameters willingly; that is to say, they must be resistant to voice disguise. The distance between the three first formants cannot be altered in this way.

2.2. On the different parameter’s relative importance

A review of the literature shows that there does not seem to be an agreement on the relative importance of different parameters. Conclusions on the priority of a parameter to study speaker individuality vary, although consulted authors try to assign them a hierarchical order. Final result can be the following:


(c) Priority of the spectral structure: absolute formant frequencies (Shearme and Holmes, 1959; Miller, 1964; Itoh and Saito, 1982; Carrell, 1984; Kuwabara and Ohgushi, 1987; Kuwabara and Takagi, 1991), or formant trajectories (Ingram et alii, 1996). On what formant or what distance between formants are more relevant, there is a lack of coincidence:

(d) Hollien, 1990: $F_1$, $F_2$, and $F_3$, and distances between $F_1$-$F_2$ and between $F_2$-$F_3$.

(e) Furui, 1986 and Ramón et alii, 2000: The more representative information on the speaker can be located between 2.5 KHz. and 3.5 KHz.

(f) There is the same priority for $F_0$ and formant structure: LaRiviere, 1975.

(h) It is not feasible to establish a priority: the importance of each parameter varies from speaker to speaker, and it also depends on the nature of the samples: Gobl, 1989; Kuwabara and Sagisaka, 1995.

2.3. Conclusions

After this first analysis of the literature, the following preliminary conclusions arise:

1. Within the revised material, there are not many recent titles: those published during the seventies and eighties are abundant, and they are considered as the starting point – without questioning them- in recent works.

2. The lack of coincident and concluding results about the more decisive parameter for speaker recognition seems to be due to the methodological approaches, usually very different between them. However, it is also difficult to establish an absolute hierarchy amongst parameters. Certain authors point out the interdependence of parameters, and that their relative priority would depend on the speaker. For example, listeners can use “low tone” as a primary key to identify speaker A, despite the fact that they apply “formant structure” to recognise speaker B (van Dommelen, 1987). That is to say, all mentioned features carry out a certain degree of information about the characteristics of a speaker, and so they seem potentially valid in the recognition task.

3. Recent works are usually centred on the methodological aspect: they seem to be especially concerned with different systems (parametric or non-parametric) to manipulate and control voice quality. There are many works written by specialists in the field of telecommunications, but studies of a strictly phonetic nature seem scarce, especially those that deal with articulatory phonetics (to study the differences between speakers in controlling and coordinating articulatory variables and their acoustic correlates).

4. It seems evident that speaker variability has not been as much studied as the invariant aspects of speech production. Thus, this research will often work with literature dealing with contrary purposes to those defined in the project.

5. The number of speakers handled to approach the different studies is also variable: from a minimum of eight to a maximum of forty different voices.

3. CORPORAS SELECTION

To accomplish the aim of the project VILE, it is essential to be able to access the speech databases that already exist for Spanish. Nowadays, without taking into account private databases, there are a lot of spoken databases with voice files that were recorded in controlled conditions. Most of them are the result of public funding. Therefore, it seems quite reasonable and convenient to take advantage of them. Only if such corpora do not take into account the aspects that the literature survey (see 2) has pointed out as primordial, a specific database for VILE will be created.

3.1. Available corpora in Spanish

As we have already pointed out, there is a good deal of speech databases in Spanish. They have been designed for specific or for general purposes, and some are quite easy to access, while others are not. Taking into account the VILE interests, the following have been selected.

ALBAYZÍN: it is one of the biggest speech database, developed in Spain (1992-1998) for speech recognition and synthesis, financed by the CICYT. The project, in which the main
Spanish research groups in speech technology participated, was coordinated by the Universitat Politècnica de Catalunya (Catalonia Polytechnic University).

| Speakers | 152 males, 152 females |
| Channel | Microphonic recording in an isolated room |
| Task | Read speech: |
| | - Phonetic corpus: 700 phonetically balanced sentences, in 2 subcorpora: a training one with four speakers and a testing one with 40 speakers. |
| | - Application corpus: 3900 sentences with geographic data. |
| | - Lombard corpus: 50 sentences from the phonetic corpus. They were read with high intensity meanwhile the speaker heard loud noises that came from the headphones. |
| Voice files | 15600 recordings of phonetically balanced sentences |
| References | Moreno et al. (1993); Casacuberta et al. (1992); Díaz Verdejo et al. (1998) |

EUROM1. It is usually described as the first really multilingual European speech database. Spanish was incorporated to this large EU project thanks to the Esprit project SAM-A. The corpora, equivalent in the eleven languages of the project, were recorded under the same conditions and with the same number of speakers.

| Speakers | 30 males, 30 females |
| Channel | Recording in an anechoic chamber |
| Task | Read speech: |
| | - Concatenated and isolated digits |
| | - 82 logatomes Ci/a/uCa, CCaIa, isolated and in context |
| | - 10 isolated words (part of the sentence with logatomes): pon, siempe, lejos, pones, aquel, quieto, di, igual, orando, dijo. |
| | - 40 paragraphs, each one of 5 sentences, with suprasegmental variation (questions, exclamations, enumerations…) |
| | - 50 sentences specific for each language: in Spanish there was a special emphasis on palatal phonemes, trills, fricatives, vocalic and consonantal sequences |
| Voice files | Not taking into account the tasks with digits, 770 recordings, which included pseudo-words (logatomes), sentences and paragraphs |
| References | Llisterri et al. (1993) |

MULTEXT. It is a part of EUROM. In this corpus, special attention has been paid to the prosodic features. It includes the original F0 and the stylised one, the transcription and coding of 15 selected paragraphs from the 40 that compose EUROM.

| Speakers | 5 males, 5 females |
| Channel | Recording in an anechoic chamber |
| Task | Read speech: |
| | -15 EUROM paragraphs by speaker; total duration: 52:21’ |
| | The F0 was stylised and resynthesised. It has the orthographic transcription, waveform, symbolic F0 coding (7 categories), original F0 and superimposed stylised F0. |
| Voice files | 150 read paragraphs |
| References | Campione and Véronis (1998); Estruch and Garrido (1996) |

GAUDÍ. According to (Ortega et al. 1998), it is “un gran corpus en español para identificación y verificación de hablantes”. It has been recently developed by the Escuela


Speakers | 224 males - 104 of them form the subcorpus AHUMADA  
|---|---
| 231 females |
Channel | Recording *in situ* (with several microphones) in a quiet room  
|---|---
| Telephonic recording |
Task | Read speech:  
|---|---
| - Isolated and concatenated digits  
| - Balanced sentences and text (the text was read with three different rhythms: normal, fast and slow)  
| - There was an specific text for each subject  
| - Free description of more than a minute. |
Spontaneous speech:  
|---|---
| - Free description of more than a minute. |
Voice files | Without taking into account the digits reading, 6,825 recordings |

SpeechDat. It is the result of a large European project, which takes into account 4000 speakers, grouped by age and speaking style. Nowadays more speakers are being incorporated to the project. Because its aim is to improve the tele-services, the data are recorded exclusively using the telephone, with the acoustical problems and limitations that such recording mode has.

<table>
<thead>
<tr>
<th>Speakers</th>
<th>2061 males, 1939 females, grouped by age (most of them are between 15-29) and speaking style (5 Spanish areas)</th>
</tr>
</thead>
</table>
Channel | Telephonic |
Task | Read speech:  
|---|---
| - Isolated and concatenated digits, amounts of money  
| - Letters (spelling), isolated words, dates, proper names, time sentences, questions and 9-10 balanced or phonetically rich sentences  
| Spontaneous speech:  
| - Telephone number, week day, date and a place  
| Lexicon, phonetic dictionary, orthographic transcription with comments |
Voice files | 18,036 recorded sentences |

As it can be observed in the above tables, a great deal of data can be obtained from the five mentioned databases.

For the segmental analysis, thousands of tokens of every Spanish phoneme can be studied under different conditions: created *ad hoc* stimuli such the ones of the EUROM list of pseudo-words (the most artificial, but also the one which there is more control over the variables), spontaneous speech in the GAUDÍ database, and the samples that can be obtained from the reading of balanced or phonetically rich sentences, paragraphs or texts, the most frequent task in speech databases.

The large amount of speakers from SpeechDat will allow, even though the recording was made telephonically, the analysis of features such as segmental duration.

The influence of variables such as intensity or articulatory effort can be studied from the Lombard Corpus of ALBAYZÍN, while the balanced AHUMADA-GAUDÍ texts are useful for the study of the effects of rhythm and speech rate.
Nevertheless, to study the suprasegmental features is much more difficult, since there are only 20 paragraphs with interrogative and exclamative sentences in EUROM1.

3.2. Intra-speaker variation

To characterise the changing elements and the invariable ones in different tokens of the same speaker, he or she has to be recorded producing the same utterance in different moments. That was how the pseudo-words of EUROM were recorded: each of the twelve speakers said the same word five times.

In GAUDÍ, the phonetically balanced sentences as well as the texts were read in three different recording sessions (apart from three sessions more, which were recorded telephonically); each recording sessions took place between 20 and 40 days after the previous one. Moreover, the effects of rhythm in a same speaker can be analysed through the reading of texts, since in each session the speaker was asked to do a normal reading, a slow reading and a fast one.

ALBAYCÍN allows the study of changes in the phonetic parameters when 20 speakers read sentences without environmental noise and then repeated the same sentences with it.

The GAUDÍ one minute spontaneous speech can be used to analyse the intra-speaker style variation—that is to say, the differences between a speaker’s spontaneous speech and the same speaker’s reading—if we select read fragments comparable to the spontaneous ones. However neither this database nor any other of the previous ones were specifically created to compare the inter and intra-speaker variation in different speaking styles and, therefore, it will be difficult to establish generalisations.

3.3. Inter-speaker variation

The aim of this kind of analysis is to establish the segmental and suprasegmental differences among speakers with a common geographic or social background when they produce the same sequences in the same speaking style. In other words, the study of inter-speaker variation will try to determine a speaker’s particular phonetic features as opposed to the common features of a group.

To investigate this, the recordings of all the speakers from the five databases can be used: more than 2000 speakers from SpeechDat, reading phonetically rich sentences, grouped by age and speech style; 445 from GAUDÍ, reading balanced sentences and a text of almost 180 words and also speaking spontaneously (description); 304 from ALBAYCÍN, all of them Central Castilian Spanish speakers, who read balanced sentences and 60 from EUROM1, who read sentences and paragraphs.

In conclusion, the existing corpora in Spanish do allow the study of segmental features in read speech; however, research on suprasegmental features and spontaneous speech will require the development of further resources in a second stage of the project.

4. CONCLUSIONS

The review of existing literature carried out with the aim of determining the relevant acoustic parameters for the study of inter-speaker and intra-speaker variation shows the difficulty to find acoustic phonetic studies on Spanish that approach in depth speaker recognition issues. The analysis of the literature available in other languages reveals that a certain unanimity exists among the diverse authors as for the acoustic parameters which define a voice profile, although it doesn’t seem easy to establish a clear priority among them, since the results of each investigation point out to one or another indistinctly. Yet, $F_0$, LTAS and formant
structure appear to be the most influential parameters when trying to recognise a speaker. Besides, it is observed that, in general introductions to forensic phonetics, more attention is devoted to the process of obtaining and processing the samples and to its legal implications than to the analysis of the relevant phonetic features of the samples. As a consequence, these works are not, in principle, especially useful for the purposes of this project.

Considering the need to reuse existing resources, an analysis of the spoken corpora developed for Spanish has also been carried out, taking into account their utility in relation to the project purposes. From this review follows that, in general, the data available for the segmental level are widely sufficient for a study like the proposed one. Nevertheless, as far as the suprasegmental elements are concerned, the representation is smaller, although EUROM1 and MULTEXT contain enough intonational variation at the sentence level, limited, nevertheless, to reading tasks. Only AHUMADA offers spontaneous conversation, as well as samples of temporal variation, since it includes texts read with three different reading rhythms. A corpus of spontaneous oral dialogues, with sufficient acoustic quality is still needed. This resource would be created within a second project, continuation of the current one.

REFERENCES


