Marrero, V., Battaner, E., Gil, J., Llisterri, J., Machuca, M. J., Marquina, M., . . . Ríos, A. (2008). Identifying speaker-dependent acoustic parameters in Spanish vowels. In *Proceedings of Acoustics'08*. (pp. 5673-7). Paris, France, June 29 - July 5, 2008. Acoustical Society of America - European Acoustics Association - Société Française d'Acoustique.

http://liceu.uab.cat/~joaquim/phonetics/VILE/VILE_Acoustics08.pdf

Identifying speaker-dependent acoustic parameters in Spanish vowels

V. Marrero[a], E. Battaner[b], J. Gil[a], J. Llisterri[c], M. Machuca[c], M. Marquina[c], C. De La Mota[c] and A. Rios[c]

[a]Universidad Nacional Educación a Distancia, Desp. 707A. C/ Senda del Rey, 7, 28040 Madrid, Spain
[b]URJC, C/Tulipan s/n, 28345 Móstoles (Madrid), Spain
[c]UAB, Filologia. Edifici B. Campus Bellaterra, 08193 Barcelona, Spain
vmarrero@flog.uned.es

In the frame of VILE Projects (Inter-and-Intra-Speaker-Variation-in-Spanish) we try to identify which acoustic parameters of the vowels are more dependent on the individual characteristics of the speaker and less on the linguistic variables. Variations of the standard deviation (SD) are analysed taking into account different groupings of the variables considered.
30 speakers from the AHUMADA database read the same text in three sessions. Mean value of four formants (F1-F2-F3-F4) and fundamental frequency (F0) are analysed in 1850 Spanish vowels, surrounded by unvoiced stops or /s/.
Speaker-dependent parameters (F3-F4) are expected to show a lower SD when grouped by speaker/session, Vowel quality parameters (F1-F2) are supposed to have a lower SD when grouped by phoneme. F0 would be related both to the speaker and the vowel quality.
Results show that F2 is the parameter with the highest SD. F4 the one with lowest SD. F0 is highly variable between vowels. No significant differentes are found in any of the parameters when grouping by session or by speaker.
Speaker/session clustering (all vowels together) compared with clustering by vowel (all speakers together) shows SD 50% higher in F1-F2, lower in F4 (75%) and F0 (66%). F3 shows no significant differences between both groupings.

1 Introduction

Our aim has been to identify which acoustic parameters of the vowels (four first formants and fundamental frequency) depend more on the individual characteristics of the speaker and less of the linguistic variables (vowel quality).

According to literature, high formants (F3 and F4) convey individual information, while F1 and F2 are dependent on vowel quality [1, 2, 3, 4, 5]. Fundamental frequency (F0) should be the most complex acoustic cue, being related in many languages to vowel quality (intrinsic F0), and suprasegmental variations (intonation, tone, stress), but playing also an important role in speaker identification [6, 7, 8].

The sample (obtained from AHUMADA Database [9]) allows inter-session comparison: 30 male speakers read the same text in three sessions, separated approximately by an interval of a month.

Parameters have been have been classified in three different ways: by session (first, second and third), by speaker (from 1 to 30) and by phoneme (/i, e, a , o/, /ˆi, `e, `a, `o/).

We analyse variations on Standard Deviation (SD) -the root mean square deviation of values from their arithmetic mean (σ)- the most common, simple and well-known measure of statistical dispersion, as an index of how widely spread are the values when grouping together parameters by session, by speaker or by phoneme.

If data are very similar, close to the mean, then the Standard Deviation will be small; if data are very variable, far from the mean, then the standard deviation will be large. When acoustic cues are clustered by phoneme, less SD should correspond, then, to parameters more dependent on vowel quality. On the contrary, when they are grouped by speaker, less SD should indicate more dependency on individual parameters.

2 Hypotheses

1 When grouping data by speaker or session, individual parameters (F3 and F4) will show less SD than vowel quality parameters (F1 and F2).
2 When grouping data by phoneme, individual parameters (F3 and F4) will show more SD than vowel quality parameters (F1 and F2).
3 F0 will show some characteristics of both individual parameters and vowel quality parameters.

2 Method

Thirteen male speakers of Spanish selected from the AHUMADA database read a phonetically balanced text in three different recording sessions. The values of 5 acoustic parameters (F0, F1, F2, F3 and F4) for the Spanish vowels /i, e, a, o/ in lexically stressed and lexically unstressed syllables have been automatically extracted with Praat [10]; then, the results were manually supervised

All vowels analysed were preceded or followed by unvoiced stops or by /s/, due the to low coarticulatory influence exerced by this context. This explains the irregular distribution of the phonemes shown in Table 1

The vowel /u/ has been excluded because of its low frequency of occurrence in Spanish [11]

	/i/		/e/		/a/		/o/		Total
Syllable	S	U	S	U	S	U	S	U	
Contexts number	1	1	6	1	3	1	5	3	21
Segments	90	89	84	536	270	83	450	265	1867

Table 1 Sample by phoneme. S = Stressed, U = Unstressed

The number of segments is the results of multiplying the number of phonemes per 3 sessions per 30 speakers, excluding some measures that have been rejected or lost. The total number of data values obtained for each of the acoustic parameters mesured is summarised in Table 2.

	F0	F1	F2	F3	F4	Total
Items	1848	1849	1847	1849	1847	9240

Table 2 Sample by acoustic parameter

In order to assess the behaviour of the parameters under consideration, data have been grouped using five criteria:

a) *By sound*. This is the smallest cluster, with just three elements: three sessions of one speaker in one session. The sample is then divided in almost 1850 groups (30 speakers * 3 sessions * 21 contexts). It represents the minimal variation (same speaker, same sound, same context and same stress). The obtained SD represents the baseline for the comparisons with the rest of the combinations.

b) *By session*. The three sessions of each speaker are assembled. The sample is divided in 90 groups (three sessions of 30 speakers). Differences between vowels are ignored, as well as syllable or context differences. Only speaker characteristics during the session are considered: one voice in one moment.

c) *By speaker*. All the sessions of each speaker are collapsed, the sample is fractionated in 30 groups, the same number of speakers. Again, differences between sounds, syllables or contexts are neglected, but also between sessions: one voice in whatever moment.

d) *By phoneme*. The sample is divided in just eight groups: -i-, -´i-, -e-, -´e-, -a-, -´a-, -o-, -´o- of variable size (cfr. Table 1). In this way, differences between speakers or sessions are not taken into account, since they remain inside the group. Differences in linguistic factors (voice quality and stress) are enhanced.

Even if in Spanish stressed and unstressed vowels (/i/ and /´i/, /e/ and /´e/, and so on) are the same phoneme, we choose this label for convenience; in any case involving a phonological opposition between them.

e) *All together (total)*. SD in 1847-1849 samples obtained by acoustic cue (see Table 2). It represents the maximum range of variation to compare with.

3 Results

Results of Standard Deviation taking into account the different groups (sound, session, speaker, phoneme and all together) are shown in Table 3. Data grouped by sound exhibit the minimum degree of variation.

	F0	F1	F2	F3	F4
a. By sound	7.22	26.00	97.75	151.71	157.88
b.By phoneme	12.44	82.13	411.79	228.44	225.18
c. By session	13.46	86.07	407.45	239.96	235.61
d. By speaker	19.58	39.37	183.39	263.62	320.15
e. Total	21.06	84.03	410.14	288.19	332.73

Table 3 Mean of SD in Hz

The level of significance obtained by comparing different groups can be observed in Table 4, which shows the statistical relevance of differences between clusters. Results have been obtained applying a Student's T-test, with distribution of two tails in two samples of equal variances (homoscedastic). Pairings with p-value above the usual threshold chosen for statistical significance (0.05) are shadowed in dark grey. In light grey, $p > 0.009$.

	S/P	S/Ss	S/Sp	P/Ss	P/Sp	Ss/Sp
F0	0.0000	0.0000	0.0000	0.0000	0.0008	0.3775
F1	0.0431	0.0000	0.0000	0.0000	0.0000	0.8116
F2	0.0262	0.0000	0.0000	0.0000	0.0000	0.7102
F3	0.0333	0.0000	0.0014	0.3484	0.5013	0.5948
F4	0.0015	0.0000	0.0014	0.0036	0.0078	0.5721

Table 4 Significance of differences (Student's T)
S=Sound, P=Phoneme, Ss=Session, Sp=Speaker

The difference between grouping the data by session or grouping by speaker (henceforth unified) was not significant for any acoustic parameter: the little increase of variability that appears when clustering three sessions of each speaker is not relevant, as expected. This fact reinforces the validity of SD as an adequate measure for our aims.

But data in Hz are hard to compare, because dimensions vary very much between F0, a harmonic with 120 Hz as mean value, and the fourth formant (about 3600 Hz as mean value). Another way to get the same results is to divide absolute data by the smallest element in each column (*by sound*, the first one); The proportion of differences in Standard Deviation taking into account all mentioned groups can be observed in Table 5. The data by sound (the group with least variation) is compared with the other groups to quantify the proportion.

	F0	F1	F2	F3	F4
a. By sound	0.00	0.00	0.00	0.00	0.00
b. By phoneme	2.71	1.51	1.88	1.74	2.03
c. By session	1.71	3.15	4.43	1.50	1.42
d. By speaker	1.86	3.31	4.38	1.58	1.49
e. Total	3.00	3.23	4.41	1.89	2.11

Table 5 Proportion of differences

The biggest range of variation is found for the two low formants (F2, F1), the proportion of SD appears to be three or four times bigger with respect to the baseline in groupings by session/speaker (4.43/4.38 and 3.15/3.31 respectively). Also F0 trebles the proportion of SD in the total of the sample. High formants (F3, F4), on the other side, show less fluctuations, the double at maximum (1.89 for F3, 2.11 for F4).

In parallel with this high global variability, F1 and F2 show the most dramatic reduction of SD when clustering data by phoneme (i.e. language dependent units): less than a half with respect to the values found in groupings by session or

by speaker. F2, in fact, is the only cue with a SD mean higher by session than by speaker, even if group size is threefold in the first case.

As for F0, the highest variation is found when studying data grouped by phoneme. Fundamental frequency seems to be the most sensible cue to the grouping by session and by speaker: SD is approximately a 30% lower than by phoneme in that case. Also F4 shows the same tendency, even if slightly less accentuated (27%). In F3 the lowest SD corresponds likewise to speaker/session group, but the difference with phoneme cluster is only 10%.

In consequence, we can consider that variations in SD in F0, F4 and F3 can be explained by the factor 'speaker/session', whereas F1 and F2 variations can be explained better by the factor 'phoneme'.

4 Discussion

The acoustic parameters analysed in this study are among to the most widely employed in forensic speaker identification and speaker verification.

F0 mean value and Standard Deviation are "among the most frequently-used parameters' for voice identification, although they are not able to provide 'a robust index index...for identification purposes" [7]. In forensic studies, long-term F0 mean and SD are the basis of measures such as likelihood ratio (LR) [12], the base-value factor (Traunmüller [13]), or the alternative baseline (Lindh [14, 15]).

Formants or formant trajectories ared avocated by Hollien [4], Kuwabara and Tagaki [16], Kreiman and Papcun [17] as main clues to identify speaker's voice

SD is also used in speaker verification, as a part of the score normalization techniques applied to the text-independent speaker verification systems ([18], [19]).

However, the comparison between different SD depending on data grouping, specially the contrast between linguistic and individual variables, can cast, in our opinion, new light on the identification of the acoustic parameters which appear to be more in the characterization of a voice.

In general, the first two formants present the largest range of variation: SD proportion can be three or four times more if data are grouped by session or by speaker, or pulled all together. On the contrary, if if data are grouped by phoneme, the variation decreases to less than the half with respect to the grouping by session/speaker. High formants (F3, F4) show less fluctuation. The highest variation is found in the *by phoneme* group, whilst the lowest one is found in the group *by speaker/session*.

Fundamental frequency seems to be the most sensible parameter to the individual grouping (by session/speaker): SD is approximately a 30% lower than by phoneme.

F3 shows a particular behavior, in the sense of a high -unexpected- dependency on vowel quality. Former results [20] show a systematic relation between the value of F3 and certain linguistic phenomena such as vocalic retroflexion and nasality; since F3 is also related to the size of the vocal tract -specially the area behind the lower teeth [21, 22]- the highest values of F3 are found in the vowel /i/ and the lowest ones in vowel /a/.

5 Conclusions

When grouping data by speaker or session, individual parameters (F3 and F4) will show less SD than vowel quality parameters (F1 and F2).

The mean SD values obtained in the grouping by speaker are arranged in the following order:

$$F4 < F3 < F0 < F1 < F2$$

Our results show that the high formants, especially F4, are more dependent on the speaker's voice than on vowel quality; on the contrary, F2 and F1 are close related to vowel quality differences

If data are gruped data by phoneme, individual parameters (F3 and F4) will show more variation than vowel quality parameters (F1 and F2).

The mean SD values obtained in the grouping by phonene can be ordered as follows:

$$F1 < F3 < F2 < F4 < F0$$

This ranging reveals the fact that F1 is the acoustic parameter which exhibits a strongest relationship with vowel quality. Furthermore, a dependency beteween F3 and vowel quality, even more marked than in the case of F2, has been observed.

F0 shares of the characteristics of individual parameters and some of the features of vowel quality parameters.

Fundamental frequency appears to be the parameter with highest SD when data are grouped by phoneme, even if stressed vowels are computed separately from unstressed vowels, to avoid the influence of lexical stress.

Acknowledgments

The authors would like to extend their appreciation to Carme Carbó, Natalia Madrigal and Montserrat Riera for their previous work on the research project VILE (Variación acústica inter e intralocutor en español).

This research has been achieved with the financial support of the Ministerio de Educación y Ciencia. Project Nº HUM2005-6980FILO.

References

[1] K. Stevens: "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds", *Proc. 7th Intern. Congr. Phon. Sc*., Montreal, 206-227 (1971).

[2] B.S. Atal: "Automatic Speaker recognition based on pitch contours", J.A.S.A. 52, 1687-1697 (1972).

[3] F. Nolan: *The Phonetic Bases of Speaker Recognition*, Cambridge University Press, Cambridge (1983).

[4] H. Hollien: *The Acoustics of Crime. The New Science of Forensic Phonetics*, Plenum, Nueva York (1990).

[5] H. Kuwabara, Y. Sagisaka: "Acoustic characteristics of speaker individuality: Control and conversion", *Speech Communication* 16, 165-173 (1995).

[6] A. Braun: "Fundamental frequency - How speaker-specific is it?", A. Braun and J.P. Köster (eds.) *Studies in Forensic Phonetics* 9-23, Trier: Wissenschaftlicher Verlag (1995).

[7] M. Jiang: "Fundamental frequency vector for a speaker identification system", *Forensic Linguistics* 3, I, .95-107 (1996)

[8] K. Johnson: "The role of perceived speaker identity in F0 normalization of vowels", *J.A.S.A*. 88:2, 642-654, (1990)

[9] J. Ortega, J. González, V. Marrero: "AHUMADA: a large corpus in Spanish for speaker characterization and identification", *Speech Communication* 31 (2-3), 255-264 (2000).

[10] P. Boersma, D. Weenink. *Praat: doing phonetics by computer*. Retrieved from http://www.praat.org/.

[11] Rojo, G.: "Frecuencia de fonemas del español actual", in Brea, M.-Fernández Rei, F. (coord.) *Homenaxe ó profesor Constantino García*. Santiago de Compostela: Universidade de Santiago Compostela. Servicio de Publicación e Intercambio Científico. pp. 451-467 (1991)

[12] P. Rose: "How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency? *Speech Communication* 10, 229-247 (1991)

[13] H. Traunmüller and A. Eriksson: "The frequency range of the voice fundamental in the speech of male and female adults". Unpubished manuscript. http://www.ling.su.se/staff/hartmut/f0_m&f.pdf (1995)

[14] J. Lindh: "Fundamental Frequency and the Alternative Baseline in Forensic Speaker Identification". *Proceedings IAFPA 2007*, The College of St Mark & St John, Plymouth, UK. www.iafpa.net-abstracts07-Lindh_-_IAFPA_2007. (2007)

[15] J. Lindh: "Preliminary F0 statistics and forensic phonetics". *Proceedings IAFPA 2006*, Department of Linguistics, Göteborg University. www.ling.gu.se-konferenser-iafpa2006-Abstracts-Lindh__IAFPA2006. (2006).

[16] H. Kuwabara, T. Takagi: "Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method", *Speech Communication* 10, 491-495 (1991).

[17] J. Kreiman, G. Papcun: "Comparing discrimination and recognition of unifamiliar voices", *Speech Communication* 10, 265-275 (1991).

[18] J. Mariéthoz, S. Bengio: "A Unified Framework for Score Normalization Techniques Applied to Text Independent Speaker Verification" *IEEE Signal Processing Letters*, Volume 12. www.idiap.ch-ftp-papers-2005-mariethoz-ieee-letters-2005. (2005)

[19] J.R. Saeta, J. Hernando, "Automatic Estimation of A Priori Speaker Dependent Thresholds in Speaker Verification", in *Proceedings 4th International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA*), pp 70-77, Ed. Springer Verlag, Guilford (United Kingdom). http://www.certiver.com/filesfordownload/79.pdf (2003).

[20] M.ª J. Albalá, E. Battaner, M. Carranza, J. Gil, J. Llisterri, M.ª J. Machuca, N. Madrigal, M. Marquina, V. Marrero, C. de la Mota, M. Riera, A. Ríos. "VILE: nuevos datos acústicos sobre vocales del español", *Proceedings of the IV Congreso de Fonética Experimental*. Granada, Universidad de Granada, forthcoming.

[21] J. Sundberg: "Observations on a professional soprano singer", *STL-QPSR* 1, 14-24 (1973)

[22] J. Sundberg: "Articulatory interpretation of the 'singing formant'", *J.A.S.A*. 55, 838–844, 1974