A Castilian Spanish digit triplet identification test for assessing speech intelligibility in quiet and in noise*

Patricia Pérez-González[1, 2], José M. Gorospe[2, 4], Enrique A. Lopez-Poveda[1, 2, 3]

[1] Unidad de Audición Computacional y Psicoacústica, Instituto de Neurociencias de Castilla y León, Universidad de Salamanca (Spain)

[2] Grupo de Audiología, Instituto de Investigación Biomédica de Salamanca, Universidad de Salamanca (Spain)

[3] Departamento de Cirugía, Facultad de Medicina, Universidad de Salamanca (Spain)

[4] Unidad de Foniatría, Logopedia y Audiología, Servicio de Otorrinolaringología, Hospital Universitario de Salamanca (Spain)

ealopezpoveda@usal.es

PACS: 43.71.Gv

Resumen

En castellano, a diferencia de otros idiomas, existen muy pocas pruebas validadas para evaluar la capacidad de una persona para percibir el habla. Aquí presentamos una prueba de identificación de tripletes de dígitos en castellano y proporcionamos resultados de referencia para personas normoyentes. La prueba consiste en identificar 100 tripletes de dígitos que se presentan a través de auriculares de inserción. Para considerar que un triplete se ha percibido correctamente, es necesario identificar correctamente cada uno de los tres dígitos en el orden en que se presentaron. Se incluyen dígitos desde el 0 al 9 pronunciados por cuatro locutores, tres hombres y una mujer. El 25% de los 100 tripletes se pronuncian por un locutor diferente. Además, se proporcionan curvas psicométricas (que ilustran el porcentaje de tripletes identificados correctamente frente al nivel sonoro de los tripletes) de referencia medidas en diez sujetos jóvenes normoyentes obtenidas en silencio y para relaciones de señal-ruido (SNRs) de 10, 0 y –10 dB. Los tripletes se presentaron con niveles sonoros entre 3 y 54 decibelios de sensación sonora (dB SL), es decir, decibelios sobre el promedio de los umbrales tonales de cada sujeto a 500 Hz, 1 y 2 kHz. Como ruido, se empleó un murmullo de 32 hablantes. Los umbrales de recepción verbal resultantes fueron 5.8, 6.1, 7.2 y 32.2 dB SL en silencio y para SNRs de 10, 0 y –10 dB, respectivamente. Se discuten, además, diferentes opciones para optimizar el tiempo de realización del test reduciendo el número de tripletes. El test puede ejecutarse de forma automática con el software AudioSpeech, desarrollado a medida en entorno Matlab™ específicamente para este fin. Tanto el software como los archivos de sonido con los dígitos están disponibles para su uso previa solicitud a los autores.

Abstract

In Castilian Spanish, unlike in other languages, there exist very few validated tests for assessing speech perception. Here, we present a digit triplet identification test in Castilian Spanish and provide reference results for listeners with normal hearing. The test consists of identifying 100 digit triplets delivered through insert earphones. A triplet is considered to be correctly perceived when each digit is correctly identified and reproduced in the correct position in the triplet. The test includes digits from 0 to 9 uttered by three male and one female speaker. 25% of the triplets were uttered by each of the speakers. Reference psychometric functions (representing percent correct triplet identification against speech level) were obtained for ten young normal-hearing listeners in quiet and for speech-to-noise ratios (SNRs) of 10, 0 and –10 dB. Speech level was varied between 3 and 54 decibels above the listener's individual mean tonal threshold level at 500 Hz, 1 and 2 kHz (dB SL). The noise was a 32-talker babble. Speech reception thresholds were 5.8, 6.1, 7.2, and 32.2 dB SL in quiet and for 10, 0, and –10 dB SNR, respectively. Ways are discussed to optimize testing time by adjusting the number of digit triplets per condition. The test may be administered automatically using AudioSpeech, a purpose-specific, custom-made application developed in Matlab™. This software and the digit recordings are available from the authors upon request.

* Nota del Editor: Por su actualidad e interés, publicamos este artículo en su versión inglesa.

1 Introduction

There is a wide variety of sound corpuses and tests for assessing speech perception in various languages. In English, for example, some popular speech tests are the CID W1 test [1], the Californian consonant test [2], or the hearing-in-noise test (HINT)[3]. There also exist speech tests in Dutch [4,5], French [6], or Polish [7]. In Spanish, there are sound corpuses for obtaining speech discrimination thresholds [8], speech reception thresholds [9,10], word discrimination scores [10,11], or speech intelligibility scores [12,13]. As of recently, there also exist Latin American Spanish [14] and Castilian Spanish [15] versions of the HINT. A widely used Castilian Spanish speech corpus is that of Cárdenas and Marrero [16]. An important distinctive feature of the latter is that it contains phonetically-balanced disyllabic words lists; that is, word lists whose phonemes are statistically representative of the Castilian Spanish language. Here, we describe the development of a Castilian Spanish speech test based on the identification of digit triplets.

Many speech corpuses consist of a limited number of short word (or sentence) lists whose items are often presented in the same order. This can be a disadvantage because listeners may memorize the lists after repeated applications of the test. Another disadvantage of these corpuses is that listeners may be unfamiliar with some of the words included in the test, which makes identification difficult. These two disadvantages may bias the results. The phonemes involved in the identification of digits are not statistically representative of a language. Nevertheless, digit identification is a convenient speech perception test in other respects [5]. First, digits are familiar to most listeners, regardless of their cultural level. Second, digits may be chosen and presented randomly. Therefore, digit identification scores may not be affected by the above mentioned biasing effects. Third, digit identification tests may be applied in a fully automatic manner, using a computer to present the sounds and to record the listener's response via a numerical keyboard. That is, the test may be applied without the need for an experimenter that controls the correctness of the response. Finally, the use of digit triplets allows obtaining precise identification estimates since the probability of correctly identifying a digit triplet by chance is 1/1000 (assuming 10 digits). (Note that the corresponding probability for a single-digit identification test would be only 1/10.)

Digit triplet identification has been employed to assess hearing impairment [4,7,17,18] as well as the performance and robustness of automatic speech recognition systems [19,20]. There exists single digit, digit pairs or digit triplet identification tests in English[21], in Polish [7], in Dutch [4,5], or in French [6]. A Spanish version of this test has been published only very recently, during the course of the present work [22].

The aim of this work was to develop a Castilian Spanish digit triplet identification test and to provide reference test results for young, normal-hearing listeners. The test is similar to those already available in Spanish or in other languages but differs from them in two respects. First, digits were recorded here with an acoustic manikin so that the test may be applied using insert earphones. Second, reference results are provided as psychometric functions showing percent correct identification versus speech level rather than speech-to-noise ratio (SNR). Reference functions are provided for a quiet condition and for SNRs of 10, 0, and –10 dB.

2 Material and Methods

2.1 Criteria for digit and speaker inclusion

Digits. Digit identification may be based primarily on the perceived number of syllables and then on the identification of the perceived phonemes [7]. In some languages, most digits are monosyllabic, which facilitates the identification of the rare multisyllabic digits and hence bias the results. To prevent this bias, digit identification tests in English typically omit the digit '7' because it is the only disyllabic digit [18,23]. Likewise, tests in Dutch omit digits '7' and '9' for the same reason [5]. In other languages, however, the number of syllables is not a reliable cue for digit identification. Ozimek et al. [7], for example, included all digits from '0' to '9' in their Polish digit triplet identification test because six of the ten digits are disyllabic in that language. Something similar applies to Spanish, where seven of the ten digits are disyllabic. Therefore, for the present test it was decided to include all digits from '0' to '9'.

Speakers. There are no consented criteria about the characteristics to be met by the speakers included in a sound corpus. Some studies have included a single male [1,2,4,7,18,21] or female [5,6,16,23,24] speaker. Others have included five male speakers [25], or five male and five female speakers [26]. Here, three male and one female speaker were included. It was decided to use volunteer amateur speakers because their utterances would sound more natural and representative of everyday speech than that of professional speakers. The native language of the four speakers was Castilian Spanish. Speakers were not paid for their services. Author PPG participated as the only female speaker.

2.2 Digit recording and processing

Before each recording session, speakers practiced to pronounce the digits at a natural speed and level. During

the recording session, speakers were asked to pronounce each of the ten digits three times in a row. The first and the third recorded utterances of each digit were discarded to minimize prosodic effects. The quality and naturalness of each recording was subjectively and independently judged by three native Castilian Spanish-speaking listeners. If necessary, recordings were repeated until they sounded natural to all three listeners.

Recordings were made in a low-reverberation, double-wall sound booth with dimensions of 1.75 m (width) by 2.67 m (length) by 1.97 m (height). Speakers were placed 115 cm in front of a Knowles Electronics Manikin for Acoustics Research (KEMAR) [27]. The KEMAR was equipped in its right ear with a silicon pinna (Knowles DB65), a Zwislocki coupler (Knowles DB100), and a half-inch microphone (B&K 4192). The microphone was connected to an sound digitizing card (RME Fireface 400) placed outside the booth via a 90º-adaptor (B&K UA0122) and a conditioner amplifier (B&K Nexus 2669). The sensitivity of the conditioner amplifier was set to 3.16 V/Pa. Recordings were digitized at a sampling rate of 44100 Hz with 32-bit analogue-to-digital resolution and were stored as mono WAV-format sound files in a computer.

Recordings were controlled and edited with Adobe™ Audition 3.0 (Adobe Systems Inc.). To attenuate the low-frequency background noise that could be perceived at high levels, each recording was filtered through a 10th-order high-pass digital Butterworth filter with a cut-off frequency of 75 Hz. The silence gaps between the digits were visually identified in the spectrogram and zeroed manually. The zeroed digit stream files were then automatically cut using a custom-made Matlab™ (The Mathworks Inc.) script to obtain a single sound file per digit. A total of 40 files (10 digits × 4 speakers) were obtained. These files are available from the authors upon request.

2.3 Collection and analysis of reference data

Listeners. Reference test results were collected for eight male and two female listeners. Their ages ranged from 24 to 31 years, with a mean age of 27.2 years. All listeners had a full clinical audiological evaluation prior to their inclusion in the study. They had normal tympanometry and their audiometric thresholds were less than 20 dB hearing level (HL) in both their ears at the audiometric frequencies from 125 to 8000 Hz [28]. The test was applied monaurally to the best ear of each listener (i.e., the ear with the lowest audiometric thresholds), which resulted in six right and four left ears. Listeners were not paid for their services. Author PPG participated as a listener.

Absolute detection thresholds (in dB SPL) for pure tones were then measured monaurally in the best ear of each participant using the same insert earphones (Etymotic ER2) that would be later used for the digit triplet identification test. Thresholds were obtained for pure tones at octave frequencies between 125 and 8000 Hz in addition to 3000 and 6000 Hz. The duration of the tones was 100 ms in total, including 5-ms onset and offset raised-cosine ramps. A two-alternative forced-choice adaptive procedure with feedback was employed. The initial level of the tones was set high enough so that the tones could be easily heard. A two-down, one-up adaptive rule was used to estimate threshold at the 70.7 % percent correct point in the psychometric function [29]. Three threshold estimates were obtained in this way for each frequency and the mean and the standard deviation (SD) were calculated. When the SD exceeded 6 dB, a fourth threshold estimate was obtained and included in the mean.

A "three-frequency average" absolute threshold (in dB SPL) was obtained for each listener as the arithmetic mean of the absolute thresholds at 500, 1000 and 2000 Hz. These frequencies were chosen for two reasons: first, because it is common in clinical practice to use the average threshold across these frequencies as a predictor of the loss of sensitivity for speech [30]; and second, to allow a direct comparison of the present reference data with corresponding data for the disyllabic phonetically-balanced word identification test of Cárdenas and Marrero [16], a very common clinical test in Castilian Spanish. The three-frequency average threshold ranged from 8.1 to 18.1 dB SPL across the 10 listeners, with a mean of 12.9 dB SPL and a SD of 2.96 dB.

Stimulus. The percentage of correctly identified digit triplets was measured as a function of speech level (L_S), in quiet and for different SNRs (Table 1). Speech level was expressed in dB re the individualized three-frequency average threshold (hereon referred to as dB sensation level or dB SL). This differs from the typical approach of other studies, where percent correct identification has been measured as a function of SNR for a fixed speech level of ~65 dB SPL. The present approach was deemed advantageous because it allows assessing speech perception for conditions where the speech level fluctuates, which are more representative of natural listening.

Table 1. Listening conditions for which reference data were obtained.

SNR (dB)	Speech level (dB SL)						
Quiet	3	6	9	12	15	27	
10	3	6	9	12	15	27	
0	3	6	9	12	15	21	27
–10	3	9	15	27	36	45	54

In the noise conditions, the noise was presented ipsilaterally to the speech signal. A different noise segment was used for each digit triplet. A 32-talker English babble was used as the noise. This type of noise was employed for convenience and because it is very common [21,31]. It is unlikely that using a multi-talker babble rather speech-shaped noise had a significant effect on the results because 8- and 128-talker babble have comparable masking effects as speech-shaped noise [25]. It is also unlikely that using English rather than a Castilian Spanish babble had a significant effect on the results because the average long-term spectrum of speech is comparable across languages [32].

The time course of the stimulus is illustrated in Fig. 1. Typically, it consisted of a 500 ms silence period, followed by a brief (10-ms) start warning sound, followed by a digit triplet, followed by a brief (10-ms) end warning sound. The time interval between the first warning sound and the first digit (or between the end of the last digit and the end warning sound) was 240 ms. The inter-digit time interval was 200 ms. When background noise was used, the noise was uninterruptedly presented during the time interval between the two warning sounds (Fig. 1). The duration of the preceding silence was set so as to give the listener sufficient reaction time after pressing the 'start test' button. The start and end warning sounds were broadband noises with a level of 50 dB SL. They acted as cues to focus the listener's attention in all conditions, particularly in the most difficult ones (i.e., lowest L_S and SNRs).

During testing, listeners sat in a double-wall sound booth. Stimuli were played digitally through an RME Fireface 400 sound card configured with a sampling frequency of 44100 Hz and a digital-to-analogue resolution of 24 bits. Stimuli were presented through Etymotic ER2 insert earphones designed to give a flat frequency response at the listener's eardrum.

The system was calibrated by connecting the earphones to a sound level meter (B&K 2238) with a Zwislocki coupler (Knowles DB100). Sound calibration was performed at 1 kHz and the measured sensitivity was applied to other frequencies. The speech level was defined as the RMS level from the onset of the first digit to the offset of the third digit; that is, the speech RMS level was defined including the zero-volt silence interval between digits. It is noteworthy that individualized digits were not equalized for RMS level. That is, the three digits within a triplet could have had slightly different RMS levels (given in Table 2 in arbitrary dB units). The influence of this on the reference results will be discussed below.

Table 2. RMS levels (dB re. 1) of the individual digits uttered by each speaker or the mean across speakers.

digit	FS	MS1	MS2	MS3	Mean
0	–32.0	–32.8	–32.6	–33.3	–32.9
1	–33.7	–36.3	–33.6	–33.6	–33.6
2	–32.5	–35.9	–33.4	–33.3	–33.4
3	–30.7	–33.2	–32.6	–30.5	–31.6
4	–32.2	–35.4	–33.3	–34.5	–33.9
5	–33.8	–35.2	–34.3	–34.9	–34.6
6	–29.0	–33.4	–30.7	–33.6	–32.2
7	–28.7	–33.6	–32.5	–32.9	–32.7
8	–35.3	–37.8	–34.9	–36.7	–35.8
9	–28.0	–34.4	–30.1	–29.7	–29.9
Min	–28.0	–32.8	–30.1	–29.7	–35.8
Max	–35.3	–37.8	–34.9	–36.7	–29.9
Mean	–31.6	–34.8	–32.8	–33.3	–33.0

Test procedure. Each listening condition was defined by the speech level and the SNR. Listening conditions are shown in Table 1. For each condition, listeners were presented with 100 digit triplets. Each individual digit was randomly chosen according to a uniform distribution of integer numbers from 0 to 9. Digit repetitions were allowed within a triplet. Each of the four speakers pronounced 25 triplets selected at random. The three digits in a triplet were always pronounced by the same speaker.

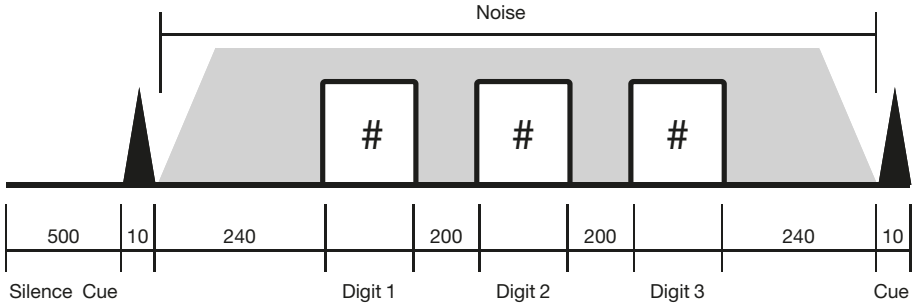

Figure 1. Stimulus time course. The duration of each segment is in ms.

Listeners were asked to identify each of the three digits after a triplet was presented and input their response via a computer numerical keyboard. Responses were recorded and analyzed to obtain the number of digit triplets identified correctly. No feedback was given on the correctness of their responses. In a few instances (< 1%), listeners unintentionally pressed a non-numerical key in the keyboard and the corresponding response did not contain three digits. Those responses were omitted from the confusion matrix analysis but were included in the other analyses for convenience and because they were so rare that they did not bias the results.

Listeners were trained in the task before actual data collection began. Training consisted of identifying 10 digit triplets for each listening condition and was structured in two 20-minute sessions. Data collection progressed from the easiest to the most difficult condition. The test took approximately 6 hours per listener (including resting time), distributed in several sessions. Individual sessions lasted from 40 minutes to 2 hours depending on the listeners' availability.

The test was run automatically using AudioSpeech, a custom-made software application developed in Matlab™. This software is available from the authors upon request.

Statistical analysis of reference test results. A Kruskal–Wallis one-way analysis of variance by ranks was applied to analyze percent identification errors across speakers or digits. Dunn's post-hoc test was applied to identify the different speaker(s) or digit(s). A difference was regarded as statistically significant when $p < 0.05$.

3 Results

3.1 Across-speaker comparisons

The test involved identifying digits pronounced by four speakers. Therefore, it is possible that identification scores differed across speakers. Indeed, Fig. 2 shows that the female speaker (FS) provoked fewer errors than any of the three male speakers (MS1, MS2 and MS3). This difference was statistically significant and was more pronounced in the most adverse conditions; that is, for SNRs of 0 and –10 dB and for low speech levels (results not shown).

It is uncertain why the female speaker provoked proportionally fewer identification errors than any of the three male speakers. One possibility is that the masking power of the noise was greater for the male than for the female voices. This could have happened, for instance, if the noise spectrum overlapped more with the female than with the male voice spectra. This, however, was unlikely because a detailed analysis of the spectra did not reveal significant differences that could explain the result in question. For example, the fundamental frequencies of speakers FS and MS1 were within the typical frequency range of female voices (140-400 Hz according to [33,34]) while that of speakers MS2 and MS3 were within the male voice range (70-200 Hz according to the same authors). Furthermore, there is no evidence (to our knowledge) that normal hearing listeners understand female speech better than male speech [35]. Another possibility is that the digits uttered by the FS were presented at a slightly higher level than for the other speakers (Table 2) and so the effective SNR could have been slightly higher for the digits uttered by the FS than for MS1, MS2 or MS3. Indeed, identification errors (Fig. 2) were negatively correlated with the mean RMS levels of the digits uttered by the individual speakers (Table 2), indicating that the higher the mean digit level the fewer the number of errors. The correlation, however, was not statistically significant (Pearson's $R = -0.74$; $N = 4$; single-tailed $p = 0.13$).

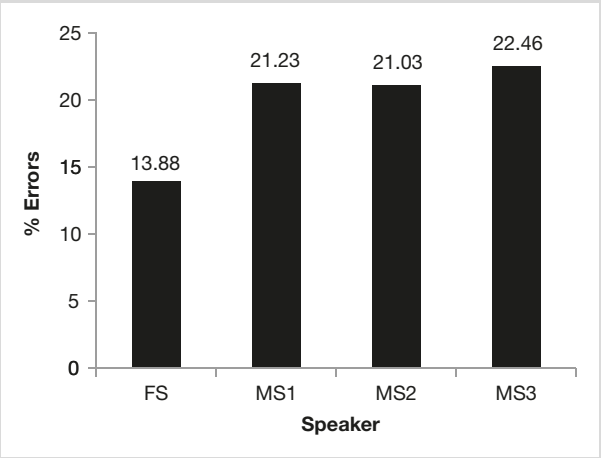

Figure 2. A comparison of global digit triplet percent identification errors across the four speakers. FS and MS# indicate female and male speaker, respectively. Errors were computed by combining results across all the listening conditions.

The results for the female speaker provide an upper limit for the performance that may be obtained with the present sound digit corpus. Even though the standard version of test involves 100 digit triplets pronounced by the four speakers in equal proportion, it is conceivable to apply a version of this test using only 25 digit triplets pronounced by the female speaker. For this reason, the following sections show separate reference results for these two versions of the test. These will be referred to as AS (for all speakers) and FS (for female speaker) versions, respectively.

3.2 Confusion matrices

Confusion matrices can be helpful for interpreting the identification errors made by hearing impaired listeners in terms of the frequency range of their hearing loss. They can also be helpful for the design of artificial digit recognition systems because they can be used to anticipate and/or correct identification errors made by such systems [36]. Table 3 shows confusion matrices for the AS and FS test versions in the quiet condition. Rows and columns correspond to presented and responded digits, respectively. Each table cell shows the number of times that the digit in the corresponding column was responded when the digit in the corresponding row was presented.

The matrix for the AS test version is based on the analysis of responses to 18000 presented digits (10 listeners × 6 speech levels × 100 triplets/speech level × 3 digits/triplet). 171 responses were invalid (i.e., they were unintentionally typed as alphanumeric rather than numeric triplets), hence omitted from the analysis. The matrix for the FS test version is based on the analysis of responses to 4500 digits pronounced by speaker FS only (10 listeners × 6 speech levels × 25 triplets/ speech levels × 3 digits/triplet). 63 invalid responses were omitted. Note that the two matrices combine responses across all six speech levels between 3 and 27 dB SL (Table 1).

Table 3. Confusion matrices for the quiet condition. Rows and columns indicate the digits that were presented and responded, respectively. The 'T' and 'Err' columns inform of the total number of times that a digit was presented and incorrectly identified, respectively. **A.** For the AS test version that includes 100 digit triplets pronounced by four speakers in equal proportion. **B.** For the FS test version that includes 25 digit triplets pronounced by the female speaker.

A) AS test version

		Digit responded											
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**T**	**Err**
Digit sent	**0**	**1261**	19	9	223	10	26	111	85	6	57	1807	546
	1	16	**1584**	42	22	41	53	15	18	40	34	1865	281
	2	1	66	**1577**	9	13	31	9	5	54	13	1778	201
	3	31	10	5	**1545**	7	10	110	36	6	15	1775	230
	4	2	15	13	3	**1658**	8	4	2	39	12	1756	98
	5	8	48	12	11	15	**1513**	21	42	40	18	1728	215
	6	21	9	7	85	8	33	**1499**	42	5	29	1738	239
	7	22	5	2	24	7	30	23	**1685**	4	48	1850	165
	8	10	79	41	8	33	46	2	7	**1606**	21	1853	247
	9	11	29	12	13	12	7	10	23	11	**1551**	1679	128
		1383	1864	1720	1943	1804	1757	1804	1945	1811	1798	**17829**	**2350**

B) FS test version

		Digit responded											
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**T**	**Err**
Digit sent	**0**	**366**	7	0	30	1	8	13	17	0	16	458	92
	1	0	**409**	8	2	5	11	2	0	10	7	454	45
	2	0	3	**404**	0	1	6	0	1	25	5	445	41
	3	10	0	0	**366**	4	1	23	3	2	7	416	50
	4	0	5	5	0	**403**	3	1	0	8	2	427	24
	5	4	9	3	1	4	**376**	3	7	6	2	415	39
	6	1	0	0	9	1	3	**406**	2	0	0	422	16
	7	2	0	0	1	2	4	2	**491**	0	2	504	13
	8	2	15	4	1	9	5	0	0	**446**	1	483	37
	9	0	3	0	1	1	0	2	7	0	**399**	413	14
		385	451	424	411	431	417	452	528	497	441	**4437**	**371**

The AS matrix shows that '0' was correctly identified on 70% of the occasions it was presented and so it was the hardest-to-identify digit. It was most frequently confused with '3'. By contrast, '4' was correctly identified on 94% of the occasions it was presented and so it was the easiest-to-identify digit. Nevertheless, identification errors across digits were statistically significant only between '0' and '4'. Of course, errors were most frequent at low speech levels (results not shown). Results for the FS test version were broadly qualitatively similar to those for the AS test version. In this case, however, the hardest-to-identify digit was '0' (80% correct identification), while '7' and '9' were the easiest-to-identify digits (97% correct identification each).

The uneven accuracy for identifying each digit is likely due to a combination of factors, including the number of syllables in the digit word (i.e., it is more easy to identify multisyllabic digits like '4' than monosyllabic digits like '2'), the frequency content (i.e., '0' was the most-difficult-to-identify digit possibly because the high-frequency content in its first syllable was more easily masked than for other digits), or the uneven RMS level across digits. Regarding the latter, a negative correlation was observed between the RMS level (Table 2) and the number of errors for the FS digits (Table 3) that just missed statistical significance (Pearson's R = -0.45, N=10, singled-tailed p=0.096). Note, however, that the correlation in question would explain only 20% of the variance of the identification error (R^2=0.2) and so the abovementioned factors and possibly others were also at play. Note also that RMS level played a much less significant role in the AS version of test as revealed by the low and non-significant correlation between the across-speaker mean digit RMS level (Table 2) and the AS errors (Table 3, AS test version) (Pearson's R = -0.12, N=10, singled-tailed p=0.37).

3.3 Psychometric functions

Figure 3 shows mean psychometric functions, which illustrate digit triplet correct identification (in %), against speech level (in dB SL). Different symbols illustrate results for different SNRs, as indicated by the legend. Figures 3A and 3B show results for the AS and FS test versions, respectively. Lines in Fig. 3 are sigmoidal functions least-squares fitted to the experimental data. Sigmoids had the form:

$$C = \frac{C_{max}}{1 + e\left(\frac{L_0 - L_s}{\tau}\right)}, \qquad (1)$$

where C is the percentage of correctly identified digit triplets, L_S is the speech level (in dB SL), C_{max} is the maximum percentage of correctly identified triplets across levels, τ is the slope of the function, and L_0 is the

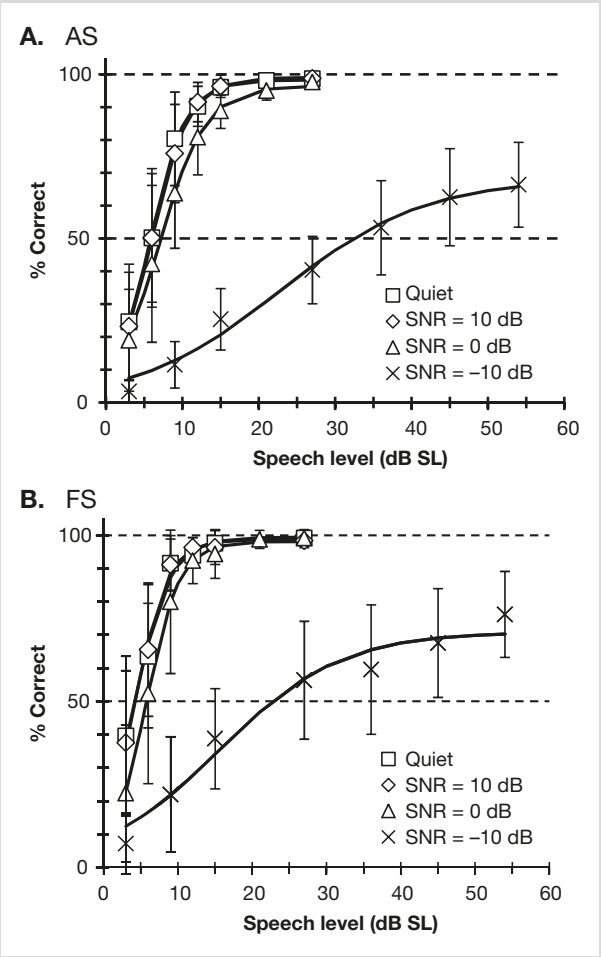

Figure 3. Mean psychometric functions for different SNRs. Symbols and error bars illustrate mean experimental scores ±1 SD. Lines illustrate sigmoidal function fits to the experimental scores. **A.** Results for the AS test version that includes 100 digit triplets per condition pronounced by four speakers in equal proportions. **B.** Results for the FS test version that includes 25 digit triplets per condition pronounced by the only female speaker.

speech level at which C becomes equal to $C_{max}/2$. Table 4 shows the resulting values of the fitting parameters (C_{max}, τ, and L_0) for each SNR and for the AS and FS versions of the test.

Comparable psychometric functions were obtained for the quiet condition and for the condition with an SNR of 10 dB. In these two conditions, identification reached nearly 100% correct for signal levels ≥ 20 dB SL [N.B.: C_{max} did not reach 100% because, for convenience, this analysis included very few (< 1%) invalid responses that were unintentionally typed as alphanumeric triplets; see Methods]. For the SNR of 0 dB and for speech levels below 20 dB SL, however, correct identification scores were lower than those observed in quiet for the same

speech levels. The psychometric functions for the SNR of –10 dB were strikingly different from the functions for other SNRs. First, maximum scores hardly reached 68% and 71% for the AS and FS test versions, respectively, even for speech levels above 55 dB SL. Second, these functions were clearly shallower than the functions for higher SNRs (e.g., for the AS test version, τ increased from 3 to 9.4 dB with decreasing SNR from 0 to –10 dB). Finally, these functions saturated at much higher levels than functions for higher SNRs (e.g., for the AS test version, the saturation speech threshold level increased from ~20 to > 55 dB SL with decreasing SNR from 0 to –10 dB).

Table 4. Parameters of the sigmoidal functions fitted to the experimental data (Eq. 1). Also shown are two data (RMS error and R^2) that inform of the goodness of fits. The last row of the table shows the speech reception thresholds (SRT) (in dB SL) obtained from the fitted functions.

		SNR (dB)			
	Test version	Quiet	10	0	–10
C_{max} (%)	FS	99.4	98.7	98.1	70.9
	AS	98.3	99.0	96.5	68.2
L_0 (dB)	FS	4.2	4.2	5.7	15.6
	AS	5.7	6.0	7.0	22.8
τ (dB)	FS	2.5	2.2	2.3	8.1
	AS	2.4	2.5	3.0	9.4
RMS err (%)	FS	6.2	4.4	2.6	11.0
	AS	3.1	0.9	2.9	6.6
R^2	FS	0.99	1.00	1.00	0.98
	AS	1.00	1.00	1.00	0.99
SRT (dB SL)	FS	4.2	4.3	5.8	22.7
	AS	5.8	6.1	7.2	32.3

3.4 Speech reception thresholds

The speech reception threshold (SRT) was defined here as the minimum speech level (in dB SL) at which normal hearing listeners correctly identified at least 50% of the digit triplets. SRTs for the different SNRs were estimated from the sigmoids fitted to the psychometric functions (Fig. 3). Resulting values are given in the bottom row of Table 4. Not surprisingly, SRTs increased with decreasing SNR. Figure 4 shows that it is reasonable to mathematically describe the observed trend as:

$$SRT\ (SNR) = SRT_{Quiet} + \exp(-SNR/ß) \qquad (2)$$

where SRT_{Quiet} is the SRT in the quiet condition and ß is a fitting parameter. When Eq. (2) was least-squares fitted to the estimated SRTs, β acquired values of 3.42 and 3.05 dB for the FS and AS test versions, respectively, and the fit was very good (the RMS error was 0.60 and 0.47 dB for the FS and AS test versions, respectively, and R^2 was 1 in the two cases). It would have been desirable to have SRTs for intermediate SNRs between –10 and 0 dB to corroborate Eq. (2). Nevertheless, the good quality of the fit to the available data suggests that it is reasonable to use Eq. (2) to interpolate the SRTs for SNRs between –10 and 100 dB.

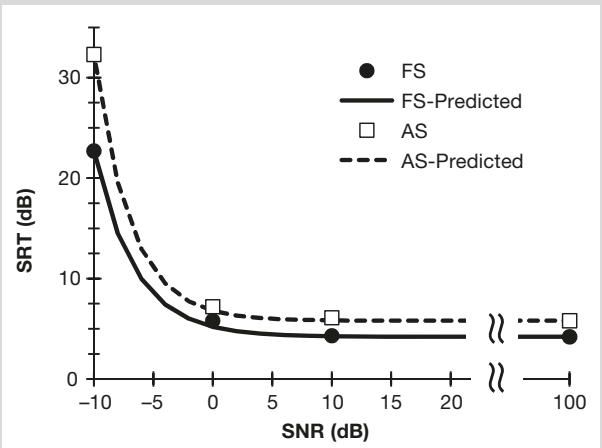

Figure 4. SRTs as a function of SNR. Symbols illustrate values inferred from sigmoids fitted to the psychometric functions. Lines illustrate trends predicted by Eq. (2).

4 Discussion

4.1 Comparison with digit triplet identification tests in other languages

Ozimek et al. [7] compared normative results of digit triplet identification tests in Polish, Dutch, German, and English. Adding the present results to that comparison is difficult for various reasons. First, tests in other languages used fixed level speech (or noise) and varied the SNR. Here, by contrast, the SNR was kept constant and the speech level was varied. Second, in other studies, the level of the reference sound (speech or noise) was expressed in dB SPL. Here, by contrast, the speech level was expressed in dB SL (i.e., decibels relative to the three-frequency average tonal threshold). That is, unlike other studies, the present approach granted the same sensation level across listeners for mid-frequency sounds, something that seems reasonable considering that the SRT is typically around 0 to 15 dB above the three-frequency tonal threshold and not at a fixed sound pressure level [16].

Our results may be compared with those of other studies only in a reduced number of conditions;

specifically, for –10 dB SNR and for a speech level of 47.50 dB SL (which corresponds to 60 dB SPL considering that the mean three-frequency average tonal threshold was 12.95 dB SPL). In these conditions, ~63% of the digit triplets were identified correctly in the AS version of the present Castilian Spanish test (Fig. 3A), compared to 80% for English, 69% for Dutch, 39% for Polish, and 38% for German. Therefore, the present scores are in line with scores of similar tests in other languages. The reason for the large variability of scores across languages is uncertain. It may be due to the fact that all other studies employed a single speaker and, as has been shown above (Fig. 2), identification scores may vary significantly across speakers. It may also be due to differences in the number of digit triplets per condition employed across studies. Most studies employed fewer than 30 digit triplets and, as will be discussed below, the fewer the number of triplets the greater variability of the results in some conditions.

4.2 Comparison with Castilian Spanish word identification tests

An interesting question is how representative of Castilian Spanish speech intelligibility are the results of the present digit triplet identification test? Of course, the answer is not straightforward but useful information in this regard may be obtained by comparing the present psychometric functions with those reported by Cárdenas and Marrero [16] for the identification of Castilian Spanish words. The comparison is informative because Cárdenas and Marrero used phonetically-balanced disyllabic words as well as polysyllabic words. A direct comparison is possible because in the two studies speech levels were expressed as dB SL.

Figure 5 facilitates the comparison in question for the quiet condition. Note that the present data (open squares, thick line) are compared with two word-identification functions: one for phonetically-balanced disyllabic words (gray, dashed-dotted line) and one for polysyllabic words (gray, dashed line). For any given speech level, performance was clearly better for the present digit triplet identification test than for any of the two word tests. This probably reflects that it is easier to identify elements from a known reduced closed set (like digits) than from an unknown open set (like words). In any case, the observed differences show that both tests (namely, digit triplet identification and word identification) are not equivalent. This indicates that it would be misleading to infer speech intelligibility in natural listening conditions from the results of the present test. Nevertheless, the present (and other) digit triplet identification tests may be advantageous over word identification tests in some respects, as described in the Introduction.

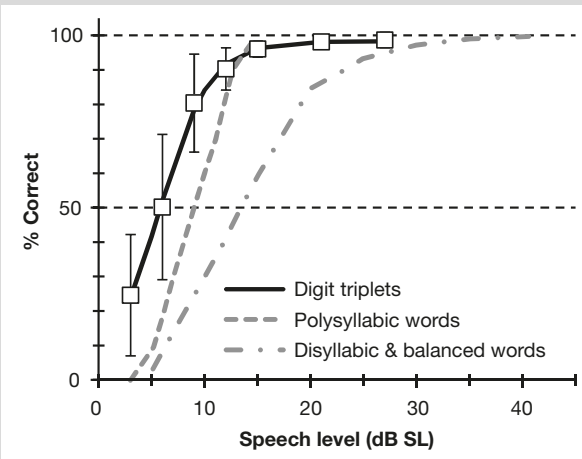

Figure 5. Comparison of the present psychometric function (AS test version) in the quiet condition with those obtained by Cárdenas and Marrero [16] using polysyllabic and phonetically-balanced disyllabic word lists.

4.3 Test time optimization

The time required to apply the present test is proportional to the number of triplets employed. For 100 triplets, the test takes approximately eight minutes per condition. This time may be reduced by using fewer triplets, which may be advantageous for practical (e.g., clinical) applications of the test. To our knowledge, there is no consensus on the optimum number of items to be employed in a speech identification test. Wilson et al. used lists of 28 digit pairs or 21 digit triplets [18]. Ozimek et al. carefully selected 100 digit triplets distributed in four lists of 25 triplets each [7]. Smits and colleagues used five lists of 23 digit triplets [4,5]. Ramkissoon et al. used 56 digit pairs [23]. Finally, Jansen et al. created 10 lists, each with 27 digit triplets [6]. None of these studies justified their chosen number of list elements, something that also applies to other consonant [26], word [1,2,10,16], or sentence [24] identification tests.

The listener's response to each individual digit triplet is, however, a binomial variable (i.e., individual responses may only be correct or incorrect). Therefore, it is possible to estimate the number of triplets in each condition, n, as [37]:

$$n \geq \left(\frac{z}{\varepsilon}\right)^2 \times p(1-p). \qquad (4)$$

where z is the chosen confidence interval (i.e., z would be equal to 1.96 or 2.58 for confidence intervals of 95 and 98%, respectively), p is the proportion (not percentage) of correctly identified triplets, and ε is the

sample error of the proportion. Since the sample proportion of correctly identified triplets, p, depends on the speech level and the SNR (Fig. 3), n also depends on these variables. Using Eq. (4) it is easy to show that n is largest when $p = 0.5$, which by definition corresponds to the SRT. In this case, for a confidence interval of 95% ($z = 1.96$) and accepting a 10% error ($\varepsilon = 0.1$), n would be equal to 96. Of course, n would be smaller in other listening conditions where the proportion of correctly identified digits is lower or higher than 0.5. Indeed, having set a confidence interval and a sample error of the proportion, the optimal n for each listening condition may be estimated using Eq. (4) together with the psychometric functions of Fig. 3. Equation (4) may also be used to show that for 25 digit triplets (a typical number in speech perception tests), the maximum sample error of the proportion would be less than 20% ($\varepsilon = 0.2$) in the worst possible case (i.e., in the SRT).

Reference results have been provided here for two test versions: the FS version, which uses only 25 digit triplets per condition, all of them uttered by the female speaker FS; and the AS version, which uses 100 digit triplets per condition uttered by the four speakers in equal proportions (25 triplets each). The FS test version takes a quarter of the time required by the AS test version (approximately 2 min *vs*. 8 min per condition). However, it uses digits uttered by a single speaker and so it is less representative of natural listening in this regard. Furthermore, it uses fewer triplets than the AS version (25 *vs*. 100) and so the error in the sample estimate of the proportion is greater than that of the AS test version (20% *vs*. 10% in the worst possible case–the SRT).

4.4 Final remarks

Great care has been exercised when developing the present test. In retrospect, however, the test could be improved in at least two respects. First, by employing digits recorded in each of the three positions where they can appear within a triplet. This would bring in prosodic effects representative of natural language. Second, the test could be improved by equalizing the RMS amplitude of *all* the digits employed in the test. While equal RMS amplitude does not guarantee equal loudness or audibility, amplitude equalization might nonetheless reduce the potential contribution of uneven RMS amplitude to across-digit errors (cf. Sections 3.1 and 3.2). It should be stressed, however, that while improvements may be made, the test is still valid for its purpose so long as it is administered using identical procedures and conditions as were used here to obtain the reference data.

5 Conclusions

1. A new Castilian Spanish digit sound corpus and software has been developed to assess speech perception using digit triplet identification.
2. The corpus includes digits uttered by four speakers. The use of multiple speakers improves the representativeness of the test but increases the results variability.
3. Reference results have been provided for a standard (multi-speaker) and a faster (one female speaker only) version of the test that require presenting 100 and 25 digit triplets per condition, respectively. There is 95% probability that the maximum error of the results be less than 10 and 20%, respectively, in the worst possible case (i.e., in the SRT).
4. Reference test results have been provided for listeners with normal hearing, at different speech sensation levels, in quiet and for SNRs of 10, 0 and –10 dB. Any change in the application conditions of the test may invalidate the comparison of the results with the present reference data.

Acknowledgements

We thank Almudena Eustaquio-Martín, Peter T. Johannesen, Enzo Aguilar and Victoria Marrero for their comments and suggestions on an earlier version of this paper. Author PPG was funded by a predoctoral scholarship of the Education Council of the Junta de Castilla y León. Work funded by the Spanish Ministry of Science and Innovation (ref. BFU2009-07909 and BFU2012-39544-C02).

References

[1] I. J. Hirsh, H. Davis, S. R. Silverman, E. G. Reynolds, E. Eldert, and R. W. Benson, "Development of materials for speech audiometry," J. Speech Hear. Disord. **17**, 321-337 (1952).

[2] E. Owens and E. D. Schubert, "Development of the California Consonant Test," J. Speech Hear. Res. **20**, 463-474 (1977).

[3] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085-1099 (1994).

[4] C. Smits and T. Houtgast, "Recognition of digits in different types of noise by normal-hearing and hearing-impaired listeners," Int. J. Audiol. **46**, 134-144 (2007).

[5] C. Smits, T. S. Kapteyn, and T. Houtgast, "Development and validation of an automatic speech-in-noise screening test by telephone," Int. J. Audiol. **43**, 15-28 (2004).

[6] S. Jansen, H. Luts, K. C. Wagener, B. Frachet, and J. Wouters, "The French digit triplet test: a hearing screening tool for speech intelligibility in noise," Int. J. Audiol. **49**, 378-387 (2010).

[7] E. Ozimek, D. Kutzner, A. Sek, and A. Wicher, "Development and evaluation of Polish digit triplet test for auditory screening," Speech Comm. **51**, 307-316 (2009).

[8] J. Tato, *Lecciones de audiometría*, (El Ateneo, Buenos Aires, 1949).

[9] C. Cancel-Ferrer, "Pruebas auditivas para pueblos de habla española," Otoryngologica **3**, 40-74 (1952).

[10] H. H. Zubick, L. M. Irizarry, L. Rosen, P. Feudo, Jr., J. H. Kelly, and M. Strome, "Development of speech-audiometric materials for native Spanish-speaking adults," Audiology **22**, 88-102 (1983).

[11] M. Pollack, *Development of a test of auditory Word discrimination in Spanish*, (University of Southern California, Los Angeles, 1971).

[12] O. Ferrer, "Speech audiometry: a discrimination test for Spanish language," Laryngoscope **70**, 1541-1551 (1960).

[13] T. Berruecos and J. Rodriguez, "Determination of the phonetic percent in the Spanish language spoken in Mexico City, and the formation of PB lists of trochaic words.," Int. Audiol **6**, 211-216 (1967).

[14] C. Baron de Otero, G. Brik, L. Flores, S. Ortiz, and C. Abdala, "The Latin American Spanish hearing in noise test," Int. J. Audiol. **47**, 362-363 (2008).

[15] A. Huarte, "The Castilian Spanish hearing in noise test," Int. J. Audiol. **47**, 369-370 (2008).

[16] M. R. Cárdenas and V. Marrero, *Cuaderno de logoaudiometría*, (Universidad Nacional de Educación a Distancia, Madrid, 1994).

[17] F. E. Musiek, "Assessment of central auditory dysfunction: the dichotic digit test revisited," Ear Hear. **4**, 79-83 (1983).

[18] R. H. Wilson, C. A. Burks, and D. G. Weakley, "A comparison of word-recognition abilities assessed with digit pairs and digit triplets in multitalker babble," J. Rehabil. Res. Dev. **42**, 499-510 (2005).

[19] A. Narayanan and D. Wang, "Robust speech recognition from binary masks," J. Acoust. Soc. Am. **128**, EL217-222 (2010).

[20] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, M. Alsulaiman, and M. Bukhari, "Formant analysis in dysphonic patients and automatic Arabic digit speech recognition," Biomed. Eng Online. **10**, 41 (2011).

[21] R. H. Wilson and D. G. Weakley, "The use of digit triplets to evaluate word-recognition abilities in multitalker babble," Seminars in Hearing **25**, 93-111 (2004).

[22] M. A. Zokoll, S. Hochmuth, A. Warzybok, K. C. Wagener, M. Buschermohle, and B. Kollmeier, "Speech-in-Noise Tests for Multilingual Hearing Screening and Diagnostics1," Am. J. Audiol. **22**, 175-178 (2013).

[23] I. Ramkissoon, A. Proctor, C. R. Lansing, and R. C. Bilger, "Digit speech recognition thresholds (SRT) for non-native speakers of English," Am. J. Audiol. **11**, 23-28 (2002).

[24] T. Cervera and J. Gonzalez-Alvarez, "Test of Spanish sentences to measure speech intelligibility in noise conditions," Behav. Res. Methods **43**, 459-467 (2011).

[25] S. A. Simpson and M. Cooke, "Consonant identification in N-talker babble is a nonmonotonic function of N," J. Acoust. Soc. Am. **118**, 2775-2778 (2005).

[26] R. V. Shannon, A. Jensvold, M. Padilla, M. E. Robert, and X. Wang, "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71-L74 (1999).

[27] M. D. Burkhard and R. M. Sachs, "Anthropometric manikin for acoustic research," J. Acoust. Soc. Am. **58**, 214-222 (1975).

[28] ANSI, *S3.6 Specification for Audiometers*, (American National Standards Institute, New York, 1996).

[29] H. Levitt, "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 466-477 (1971).

[30] R. S. Schlauch and P. Nelson, "Puretone evaluation," in *Handbook of Clinical Audiology,* J. Katz, R. Medwetsky, R. Burkhard, and L. Hood, eds., (Lippincott Williams & Wilkins, Baltimore, 2009), pp. 30-49.

[31] C. Cherry, *On human Communication*, (MA: MIT Press, Cambridge, 1996).

[32] D. Byrne, H. Dillon, K. Tran, S. Arlinger, K. Wilbraham, R. Cox, B. Hagerman, R. Hetu, H. Kei, C. Lui, J. Kiessling, M. Nasser Kotby, N. Nasser, W. A. E. El

Kholy, Y. Nakanishi, H. Oyer, R. Powell, D. Stephens, R. Meredith, T. Sirimanna, G. Tavartkiladze, G. I. Frolenkov, S. Westerman, and C. Ludvigsen, "An international comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**, 2108-2120 (1994).

[33] A. Boothroyd, *Speech acoustics and perception*, (Pro-Ed, Austin (Tex), 1986).

[34] B. Yang, "A comparative study of American English and Korean vowels produced by male and female speakers," J. Phonetics **24**, 245-261 (1996).

[35] M. A. Kiliç and F. Ogüt, "The effect of the speaker gender on speech intelligibility in normal-hearing subjects with simulated high frequency hearing loss," Rev. Laryngol. Otol. Rhinol. (Bord.) **125**, 35-38 (2004).

[36] O. C. Morales and S. Cox, "On the Estimation and the use of confusion-matrices for improving ASR accuracy," in *10th Annual Conference of the International Speech Communication Association,* (2009), pp. 1579-1582.

[37] E. A. Lopez-Poveda, *Fundamentos de estadística*, (La Popular, Albacete, 2002).

Paremias, aforismos, adagios y otros relatos acústicos

¡Plop!

Y el Maestro dijo:

«Hoy, a partir de este momento, solamente voy a emitir sonidos con mi cuerpo o con objetos que se transporten usualmente. Debéis reconocerlos e indicar una sugerencia de persona, situación o cosa que lo produzca».

Los alumnos se miraron extrañados. Algunos incluso se burlaban.

El Maestro empezó por colocarse el dedo en la boca, apretó los labios y abrió la cavidad como para emitir la vocal O, y sacó el dedo emitiendo un sonido.

¡PLOP!, sonó.

Un alumno dijo que era igual al descorche de una botella de vino. Otro dijo que le recordaba a la máquina de expedir los billetes en los ferrocarriles.

Y como que nadie más decía otra cosa, la alumna aventajada dijo: «Es el mismo sonido que produce Harrison Ford en dos películas «Armas de mujer» y «Seis días, siete noches».

El Maestro, realmente asombrado no sólo de la memoria de la alumna, sino también de su capacidad de relacionar los sonidos de situaciones muy diversas, le preguntó: «¿Crees que este sonido podría considerarse como una huella o símbolo sonoro de dicho actor?».

La alumna lo meditó unos instantes y respondió: «¿Algo así como la imagen visual de Hitchcock, que siempre aparece en alguna escena de las películas que dirige?».

«Así es, en efecto», le contestó el Maestro.

«Yo creo que cada uno de nosotros tiene o produce unos sonidos determinados, que son distintos a los de los demás. Sí, creo que sería parte de su personalidad sonora», concluyó la alumna.

Y el Maestro continuó: «De la misma forma que si somos altos, bajos, delgados, regordetes, cuatro ojos, etc. que es una de las formas de ponernos apodos, ¿pensáis que podemos tener apodos por nuestra personalidad sonora?».

Algunos alumnos afirmaron que sí con la cabeza.

«A un alumno de la clase de al lado lo llamamos «El Chapas» porque siempre lleva cosas metálicas que suenan», dijo uno.

«¿Te refieres al que lleva cadenas colgando y botas con herrajes?», preguntó el Maestro.

Todos reconocieron a qué alumno se estaba refiriendo por los sonidos que emitía.

«Y la chica del cascabel que le cuelga de su bolso», dijo otro alumno.

El Maestro pensó para si en cómo reconocía a la Directora por su andar majestuoso en los pasillos del centro.

«Y el del calzado deportivo especial que hace nyc-nyc, en cualquier pavimento y no sólo en el polideportivo», añadió otro.

Casi todos los alumnos se iban animando a participar.

«¿Y de mí?», cortó el Maestro.

El silencio se impuso en el aula. Nadie se atrevía a decir nada.

El alumno burlesco se levantó y dijo alto y fuerte: «El apodo es Maestro roncador».

El Maestro sabía que cuando se quedaba en el salón de profesores para tomar un café después de las comidas, los restantes profesores desaparecían en breves instantes.

Asintió con la cabeza e iba a hacer la conclusión cuando la alumna aventajada añadió:

«Aunque algunos te reconocen como el Maestro Roncador, e incluso algunos creen falsamente que ese es tu apellido, otros te reconocemos sólo como el Maestro».

Francesc Daumal Domènech
Maestro Roncador