informa
healthcare

International Journal of Audiology 2014; Early Online: 1–6

International
Journal of
Audiology

Technical Report

The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology

Vincent Aubanel*, Maria Luisa García Lecumberri† & Martin Cooke‡

*The MARCS Institute, University of Western Sydney, Penrith, NSW, Australia, †Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain and ‡Ikerbasque (Basque Foundation for Science), Bilbao, Spain

Abstract

Objective: The current study describes the collection of a new phonemically-balanced Spanish sentence resource, known as the Sharvard Corpus. *Design:* The resource contains 700 sentences inspired by the original English Harvard sentences along with speech recordings from a male and female native peninsular Spanish talker. Sentences each contain five keywords for scoring and are grouped into 70 lists of 10 sentences using an automatic phoneme-balancing procedure. *Study sample:* Twenty-three native Spanish listeners identified keywords in the Sharvard sentences in speech-shaped noise. *Results:* Psychometric functions for the Sharvard sentences indicate mean speech reception thresholds of − 6.07 and − 6.24 dB, and slopes of 10.53 and 11.03 percentage points per dB at the 50% keywords correct point for male and female talkers respectively. *Conclusions:* The resulting open source collection of Spanish sentence material for speech perception testing is available online.

Key Words: Speech perception in noise; phonemic balance; Spanish; open speech resource

Although Spanish—after Mandarin—is the language spoken by the largest number of native speakers (Lewis et al, 2013), there are very few open source Spanish speech resources (e.g. sentence lists or recordings) available for audiological use. The current article describes a new sentence corpus for Spanish motivated by the desire to provide a useful resource for speech perception studies with Spanish listeners, and in particular to enable cross-study comparisons based on the use of common, easily-available materials. The corpus is intended to be equivalent to the Harvard sentence material (Rothauser et al, 1969) which is widely-used in speech perception tests (e.g. Bradlow et al, 1996; Hawley et al, 2004; Hu & Loizou 2010; Cooke et al, 2013). The Harvard Corpus consists of phonemically-balanced lists of 10 sentences, where each sentence contains five keywords used for scoring. Given this ancestry, the new resource is known as the 'Sharvard Corpus'.

The Sharvard Corpus complements existing Spanish audiological materials such as (1) the Castilian Spanish hearing in noise test (Huarte, 2008) based on a list of 240 Spanish sentences adapted from the English HINT test (Nilsson et al, 1994); (2) the Spanish matrix-style (*name verb numeral object adjective*) sentence lists where any combination of the 10 alternatives for each word type yields a semantically valid sentence, e.g. '*Claudia busca tres zapatos enormes*' ('Claudia is looking for three enormous shoes') (Hochmuth et al, 2012); (3) the phonetic corpus of the Albayazin speech database (Moreno et al, 1993) which contains two sets of phonetically balanced sentences, a set of 200 sentences selected from spontaneous speech transcriptions and a set of 500 sentences selected from written texts; and (4) the test of Spanish sentences (Cervera & González-Alvarez, 2011), containing six lists of 50 sentences, balanced for phonetic content (with respect to five phoneme classes) and predictability of the final word, i.e. high- or low-predictability based on the initial part of the sentence. All of these corpora are currently subject to one or more limitations which constrain their wider usage, e.g. licensing restrictions or lack of published lists and speech recordings. Unlike existing corpora, sentence lists and recordings for the Sharvard Corpus are available for unrestricted usage.

Correspondence: Vincent Aubanel, The MARCS Institute, University of Western Sydney, Locked Bag 1797, Penrith, NSW 2751, Australia. E-mail: v.aubanel@uws.edu.au

(Received 29 July 2013; accepted 14 March 2014)

ISSN 1499-2027 print/ISSN 1708-8186 online
DOI: 10.3109/14992027.2014.907507

Abbreviations

RMS Root mean square
SNR Signal-to-noise ratio
SRT Speech reception threshold

In addition to the 70 lists of 10 sentences, the new corpus contains speech signals from recordings of the complete corpus by one male and one female Spanish native speaker, along with phonemic transcriptions. Subsequent sections describe the design of the sentence material and outline an automated list selection procedure which maximizes phonemic balance. The outcomes of a speech-in-noise test using the Spanish material is also reported.

The Sharvard Corpus

Sentence material

As a starting point, the Harvard sentences were translated into Spanish in order to obtain a corpus with a similar level of difficulty as that of the original English material.

Sentences were then verified and revised by two native Spanish speakers using the following criteria. First, the original English sentences were adapted to Spanish syntactic and pragmatic conventions as well as to the cultural context and to meet the constraint on number of syllables per word (see below). For example, 'Kick the ball straight and follow through' was transformed into 'Dale al balón con la punta de la bota y fuerte' ('Kick the ball strongly with the tip of the boot'). Second, new sentences were created where there was no reasonable Spanish translation (e.g. 'Mesh wire keeps chicks inside' or 'A gold ring will please most any girl').

Sentences were restricted to contain exactly five keywords. These are almost always content words, but occasionally—as in the original Harvard Corpus—pronouns and other function words could be marked as keywords, e.g. when the sentence context could call for an emphasis ('El té no se hace con agua fría', 'Tea can't be made with cold water') with 'no' tagged as a keyword. All words were restricted to have a maximum of two syllables to ensure a reasonable number of lexical competitors. In all, 700 sentences were generated in this way.

A pronunciation dictionary for keywords was constructed using the Saga toolkit[1] which uses the grapheme-to-phoneme conversion rules established in Llisterri & Mariño (1993). The pronunciation dictionary makes use of 31 symbols to describe the phonological system of Castilian Spanish.

A phonemic level of analysis was employed in order to accommodate for pronunciation variations in different recordings of the sentence lists. To achieve this, the seven allophones /j, w, β, ð, ɣ, z, ŋ/ were merged with their main phoneme category, as shown in Table 1. Phonemic balancing was therefore subsequently conducted on 24 phonemes.

Overall phoneme frequency distribution

The phoneme frequency distribution of the Sharvard Corpus is compared against other published corpora of Spanish in Figure 1. The distribution is generally consistent with the frequency distribution of previous corpora, both spoken and written. Some departures are apparent for certain phonemes, largely as a consequence of the omission—in aggregating counts for keywords only—of phonemes that occur in high-frequency function words such as articles and pronouns. For example, discarding 'en', 'el', and 'de' contributes to an under-representation of /e/, /d/, /n/ and /l/ in the Sharvard Corpus.

Table 1. Phoneme inventory used for transcription, along with their frequency of occurrence in keywords for the corpus, based on a total phonemes-in-keywords count of 16 333. Frequencies of the allophones given in the last column sum up to the frequency of their main phoneme category.

Sound class	*IPA*	*Frequency (%)*	*Allophone frequency (%)*
vowel	a	13.93	
	e	10.76	
	i	5.80	(i: 3.19, j: 2.61)
	o	11.39	
	u	3.76	(u: 1.65, w: 2.11)
plosive	p	3.00	
	t	4.99	
	k	3.73	
	b	3.99	(b: 2.33, β: 1.66)
	d	2.94	(d: 1.49, ð: 1.45)
	ɡ	1.87	(ɡ: 0.88, ɣ: 0.99)
fricative	f	1.16	
	θ	1.65	
	s	6.48	(s: 6.41, z: 0.07)
	ʝ	0.42	
affricate	x	1.44	
nasal	ʧ	0.94	
	m	3.18	
	n	5.47	(n: 5.14, ŋ: 0.33)
	ɲ	0.41	
lateral	l	3.57	
	ʎ	0.73	
rhotic	ɾ	1.18	
	r	7.20	

Phonemic balance

Phonemic balance is traditionally understood as the approximate equivalence of the phoneme frequencies in a given corpus with the phonemic distribution of the language from which the sample is drawn. For speech materials used to compare multiple conditions, it is useful to define subsets of the corpus in which the equivalence principle also applies. Phonemic balance across subsets was achieved here using an automated optimization procedure which is able to partition the complete sentence set into balanced subsets of arbitrary size. The algorithm makes use of the squared Euclidean distance $d_{s,c}$ between the distribution of phoneme frequencies f in a given subset of sentences s and that of the whole corpus c:

$$d_{s,c} = \sum_{p=1}^{P} (f_{p,s} - f_{p,c})^2 \quad (1)$$

where P is the number of phonemes (24). Sentences are initially partitioned randomly into subsets of size S (here, $S = 10$) and the distance d calculated for each subset. Then, a randomly-chosen sentence from the subset with the largest distance (the 'worst' subset) is interchanged with a sentence from another subset in such a way that the value of d decreases for both subsets. This process iterates until no further interchanges involving the worst subset are possible. Note that the worst subset is not necessarily the same subset

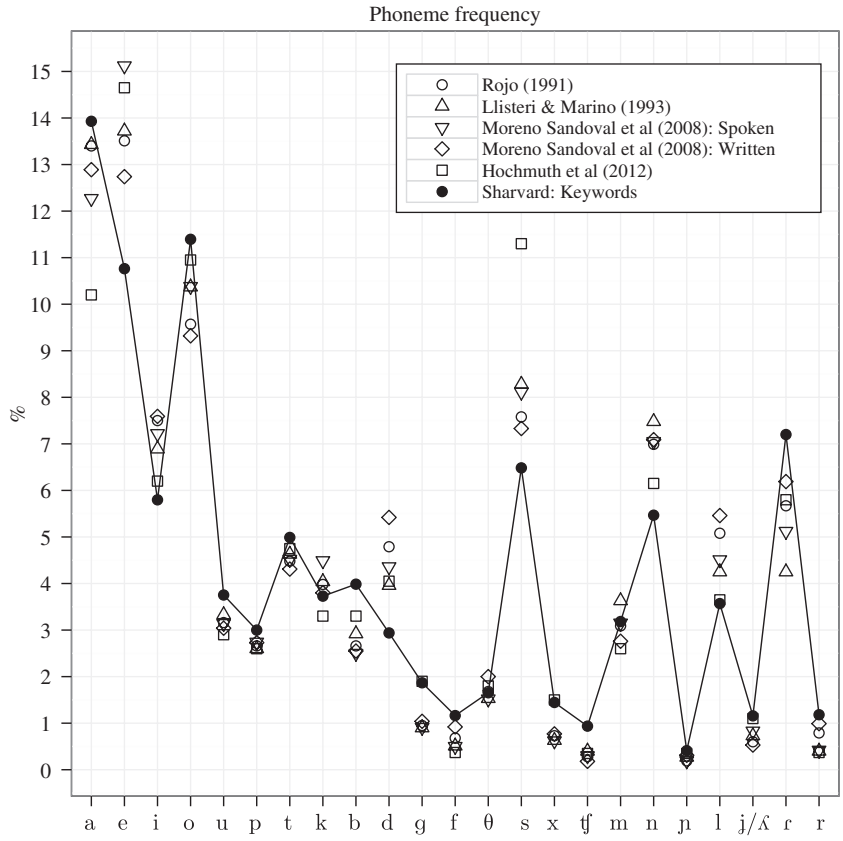

Figure 1. Phoneme frequency distribution of keywords in the Sharvard Corpus. To accomodate for inventory discrepancies across corpora, counts for /ʎ/ and /ʝ/ are aggregated in this figure. Other sets are plotted for comparison: **Rojo (1991)**: 3.8 million words corpus with a variety of Castilian and Latin American Spanish written texts. The frequencies presented are the ones reported by Llisterri & Mariño (1993), adjusted to redistribute archiphoneme frequencies to map with their phonetic inventory of Spanish; **Llisterri & Mariño (1993)**: 100 000 phonetic segments automatically derived from orthographic transcription of three hours of semi-spontaneous speech provided by three native Spanish speakers; **Moreno Sandoval et al (2008) (spoken)**: Spanish C-ORAL- ROM corpus, consisting of 42 hours of recorded speech by 429 speakers covering three styles of speech (informal, formal, media), containing 348 000 orthographically-transcribed words with automatically-produced phonetic transcriptions; **Moreno Sandoval et al (2008) (written)**: 480 000 word written corpus from a news agency, automatically transcribed; **Hochmuth et al (2012)**: Matrix-based material constructed to represent the phonemic distribution of Spanish, with frequency values estimated from Figure 1 of Hochmuth et al (2012).

on each iteration. At this point, the process is repeated for the second worst subset, and continues until no subset can be improved by interchanges.

The upper panel of Figure 2 shows the result of the optimization procedure for the Sharvard Corpus. Data are shown both pre- and post-optimization as means over 10 independent runs with different randomized initial sentence groupings. For each phoneme, phonemic balance is computed as the mean over subsets of the magnitude of the difference between the frequency of that phoneme in each subset and in all subsets. Global balance is quantified by the mean over phonemes of that quantity. The balancing procedure achieves a 2.70-fold reduction of mean phonemic balance and a 13-fold reduction of its standard deviation (unbalanced: $\mu = .840$, $\sigma = .357$, balanced: $\mu = .311$, $\sigma = .026$). Results are also shown for 5-sentence lists to illustrate the difficulty of achieving good phonemic balance with smaller subsets. In that case, a 1.95-fold reduction (7.14-fold reduction in standard deviation) is obtained (unbalanced: $\mu = 1.221$, $\sigma = .518$, balanced: $\mu = .625$, $\sigma = .073$).

Using the same balancing procedure, the lower panel of Figure 2 presents data for the original Harvard Corpus, using the CMUdict[2] pronunciation dictionary for American English. Intriguingly, the phonemic balance of the published Harvard Corpus is quite poor and can be improved significantly with the current balancing algorithm (published: $\mu = .815$, $\sigma = .339$, balanced: $\mu = .365$, $\sigma = .067$). Only phonemes occurring in keywords were considered for balancing, but similar results were obtained using all words (not included in Figure 2; published: $\mu = .685$, $\sigma = .283$, balanced: $\mu = .309$, $\sigma = .058$).

Speech material

In addition to the sentence lists and associated annotations, the Sharvard Corpus is distributed with spoken recordings of the entire corpus from one male and one female talker. This section documents the recording procedure and presents keyword intelligibility data for the two talkers at nine signal-to-noise ratios (SNRs).

Sentence elicitation

A female talker (age: 48) and a male talker (age: 28), both chosen for the clarity of their speech, were recruited to read the complete

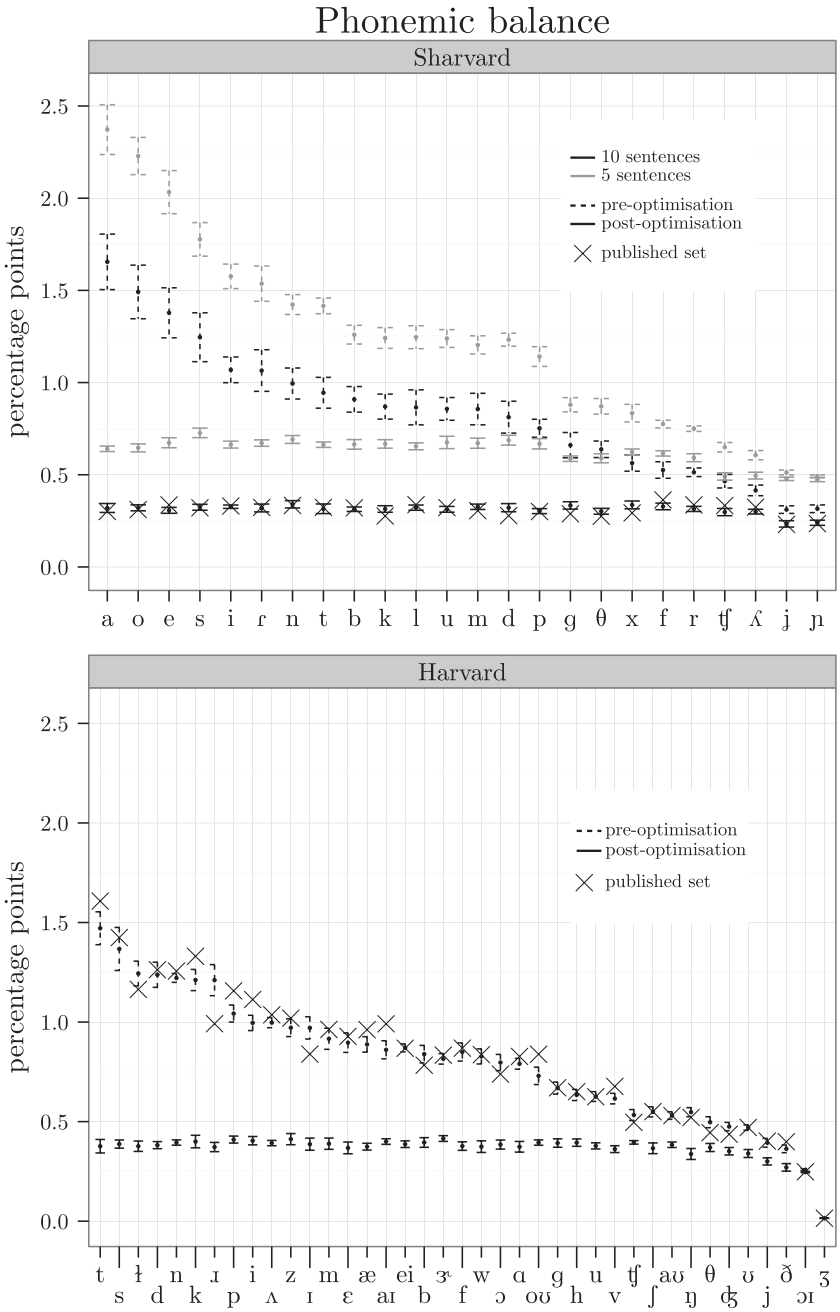

Figure 2. Phonemic balance for the Sharvard (upper) and Harvard (lower) corpora. Phoneme order is based on imbalance pre-optimization. Error bars depict 95% confidence intervals computed over 10 runs with different random initializations. Crosses are across-list means for the published 10-sentence sets.

Sharvard Corpus. Both had a peninsular Spanish accent as generally spoken in the northern half of Spain. Talkers were asked to read each sentence at a normal speaking rate and to pause between utterances. Talkers were able to repeat any utterance if they felt they had not produced it correctly. The sentence order was randomized—and therefore unrelated to the final ordering of sentence lists (see section 2.3) —in order to avoid potential sequencing effects, e.g. list intonation or fatigue.

Recordings were made in a sound studio in the Phonetics Laboratory of the University of the Basque Country using a table top AKG XL-II microphone and digitized at 48 kHz/16-bits with a RME Fireface 800 analogue-to-digital converter. Sentences were manually segmented and screened, and in the case of the male talker a small number of within-sentence pauses were shortened, a process that retained their naturalness. Sentences were normalized by dividing the signal by the maximum amplitude observed for that talker in the entire corpus, and saved as individual WAV format files.

Talker intelligibility

A listening experiment was carried out to assess the overall intelligibility in noise of the Sharvard speech material. Psychometric

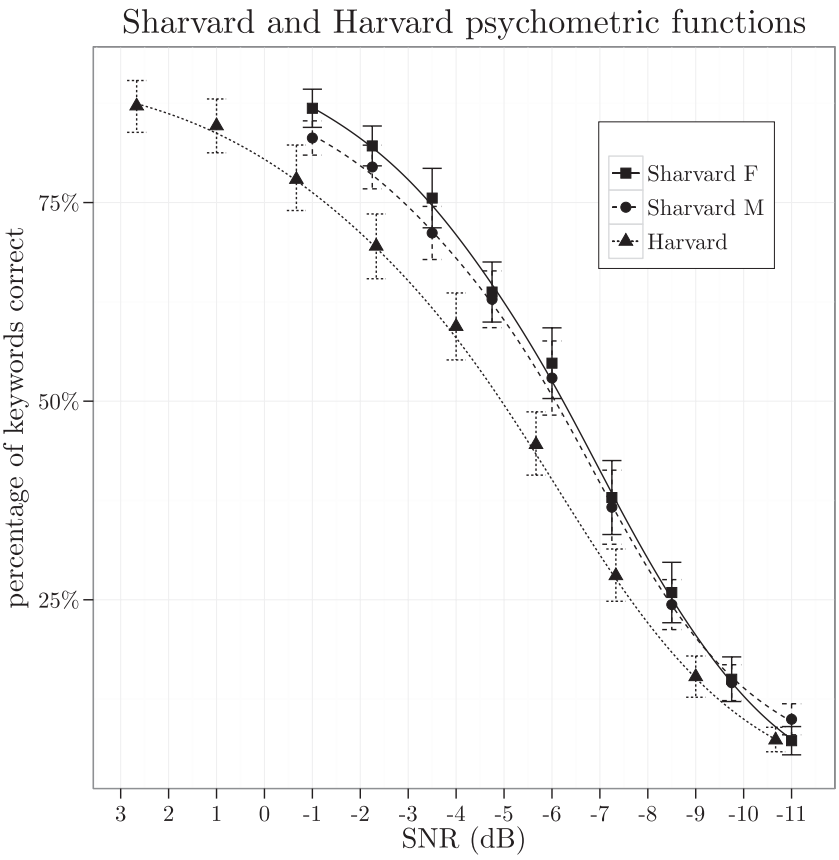

Figure 3. Psychometric functions for the male (M) and female (F) talkers of the Sharvard Corpus, along with those from a male talker of the original Harvard Corpus.

functions were computed by measuring the proportion of keywords identified correctly in noise by native talkers as a function of SNR. The two talkers were assessed independently using speech-shaped noise maskers whose long-term average speech spectrum matched each individual talker, presented at 9 SNR values linearly spaced from -11 dB to -1 dB. These values were chosen in pilots to produce keyword scores spanning the range 10 to 90%. Speech-plus-noise mixtures were constructed following the procedure described in Cooke et al, 2013, section 3.3: sentences were centrally-embedded in the masker, which started/ended 500 ms before/after the speech to prevent reductions in intelligibility due to co-gating (Cervera & Ainsworth 2005). Speech level was scaled to reach the required SNR for the region where it overlapped with the masker, and the mixtures were presented at a fixed level of 79 ± 0.6 dB (A), measured using a Bruel & Kjaer artificial ear (model 4153) coupled to a Bruel & Kjaer sound level meter (model 2250 light) over Sennheiser HD 380 pro closed headphones.

Twenty-three listeners were recruited from the undergraduate population at the University of the Basque Country. Listeners were paid for their participation. Following hearing screening, 22 subjects (Age: $\mu = 22.3$ years, $\sigma = 2.69$) with bilateral hearing better than 20 dB HL for the range $125 - 8000$ Hz were retained for the study. Listeners participated in two sessions on different days in which they heard either the female or male talker, and gender order assignment was balanced across listeners. Stimuli were randomly sampled from the entire corpus and presented in 20-sentence blocks for each of the nine SNR levels. Nine different blocks were generated at each SNR to ensure that, overall, each subset of 20 sentences was heard the same number of times at each SNR and that listeners heard each sentence only once. Blocks were assigned to listeners following a Latin square design. Stimuli were presented using a custom MATLAB programme. The experiment was self-paced: participants were asked to type what they heard, after which the next stimulus was presented following a short delay. Each session lasted around 45 minutes, including a short practice session. Over the two sessions each listener heard a total of 360 sentences, half of which spoken by the male talker, the other half by the female talker.

Responses were corrected automatically for common alternative word forms (e.g. digit input for numbers). The mean percentage of correctly identified keywords is plotted in Figure 3, along with similar data from a British English talker producing the original Harvard sentences (from Cooke et al, 2013. Note that the SNR range used in that study was -11 dB to $+3$ dB). Model-free fits (Zychaluk & Foster 2009) of the psychometric curves are shown. Estimated speech reception thresholds (SRT) at a range of correctness levels are provided in Table 2.

The male and female Sharvard talkers possess a similar intelligibility versus SNR relation while the male talker for the original Harvard Corpus exhibits a SRT at 50% correct which is 1.2 dB higher. Language factors such as vowel inventory size or stress patterns may underlie this difference. For example, it may be easier for Spanish listeners to identify a masked vowel sound since the inventory from which it is drawn has fewer elements than in English. The greater number of bisyllabic words in the Spanish corpus compared

Table 2. Speech reception thresholds (SRT) and psychometric function slopes at correctness levels of 25, 50, and 75%.

	Sharvard F		*Sharvard M*		*Harvard*	
	SRT	slope	SRT	slope	SRT	slope
25%	− 8.52	9.61	− 8.42	8.97	− 7.64	8.39
50%	− 6.24	11.03	− 6.07	10.53	− 4.94	8.96
75%	− 3.44	6.84	− 2.91	5.73	− 1.27	4.88

to the English counterpart may also provide additional segmental cues for recognition. Another possibility is that the individual talkers chosen to produce the two corpora differ in intrinsic intelligibility (i.e. talker-related intelligibility due to factors such as speech rate and unreduced phoneme targets). For example, large inter-individual differences in intrinsic intelligibility in noise have been observed amongst English speakers at similar SNR levels (Barker & Cooke, 2007, Figure 4).

Summary

An audiological resource for the Spanish language based on the Harvard sentence material is presented. The new 'Sharvard Corpus' contains 700 sentences partitioned into phonemically-balanced subsets of 10 sentences. The original Harvard sentences were translated and adapted to the syntactic, pragmatic, and phonetic structure of Castilian Spanish as well as to Spanish cultural conventions. With minor phonetic adjustments (such as the exclusion of the phoneme /θ/ and consequent rebalancing of /s/), and lexical modifications (such as the substitution of 'euro' to the local currency), the Sharvard Corpus is readily adaptable to other Spanish varieties including those of Latin America.

The sentence material, phonemic transcriptions and spoken recordings of the entire corpus from one male and one female native Spanish talker are available as a supplementary material in the online version of the journal, through the direct link to the article at http://informahealthcare.com/doi/abs/10.3109/14992027.2014.907507.

Notes

1. http://www.talp.upc.edu/index.php/technology/tools/signal-processing-tools/81-saga. Last viewed July 10, 2013.
2. http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Last viewed July 10, 2013.

Acknowledgements

This work was supported by the LISTA Project, funded from the Future and Emerging Technologies programme within the 7th Framework Programme for Research of the European Commission, FET-Open grant number 256230. We thank Ainara Imaz for initial screening of the Spanish translations and Letizia Marchegiani and Albino Nogueiras for making available the pronunciation dictionary.

Declaration of interest: The authors report no conflicts of interest.

References

Barker J. & Cooke M. 2007. Modelling speaker intelligibility in noise. *Speech Comm*, 49, 402–417.

Bradlow A.R., Torretta G.M. & Pisoni D.B. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Comm*, 20, 255–272.

Cervera T. & Ainsworth W.A. 2005. Effects of preceding noise on the perception of voiced plosives. *Acta Acust*, 91, 132–144.

Cervera T. & González-Alvarez J. 2011. Test of Spanish sentences to measure speech intelligibility in noise conditions. *Behav Res*, 43, 459–467.

Cooke M., Mayo C., Valentini-Botinhao C., Stylianou Y., Sauert B. et al. 2013. Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Comm*, 55, 572–585.

Hawley M.L., Litovsky R.Y. & Culling J.F. 2004. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *J Acoust Soc Am*, 115, 833–843.

Hochmuth S., Brand T., Zokoll M.A., Castro F.Z., Wardenga N. et al. 2012. A Spanish matrix sentence test for assessing speech reception thresholds in noise. *Int J Audiol*, 51, 536–544.

Hu Y. & Loizou P.C. 2010. On the importance of preserving the harmonics and neighboring partials prior to vocoder processing: Implications for cochlear implants. *J Acoust Soc Am*, 127, 427–434.

Huarte A. 2008. The Castilian Spanish hearing in noise test. *Int J Audiol*, 47, 369–370.

Lewis P.M., Simons G.F. & Fennig C.D. 2013. Ethnologue: Languages of the World. Seventeenth edition. Dallas, USA: SIL International.

Llisterri J. & Mariño J.B. 1993. Spanish adaptation of SAMPA and automatic phonetic transcription. Tech. rep. SAM-A/UPC/001/V1.

Moreno Sandoval A., Toledano D.T., de la Torre Á., Garrote M. & Guirao J.M. 2008. Developing a phonemic and syllabic frequency inventory for spontaneous spoken Castilian Spanish and their comparison to text-based inventories. Marrakech, Morocco: LREC.

Moreno A., Poch D., Bonafonte A., Lleida E., Llisterri J. et al. 1993. Albayzín speech database: Design of the phonetic corpus. Berlin: Eurospeech. 175–178.

Nilsson M., Soli S.D. & Sullivan J.A. 1994. 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.

Rojo G. 1991. Frecuencia de fonemas en español actual. Homenaxe ó profesor Constantino García. Santiago de Compostela: Universidade de Santiago de Compostela: Servicio de Publicación e Intercambio Científico, pp. 451–467.

Rothauser E.H., Chapman W.D., Guttman N., Hecker M.H.L., Nordby K.S. et al. 1969. IEEE Recommended practice for speech quality measurements. *IEEE Trans Audio Acoust*, 225–246.

Zychaluk K. & Foster D.H. 2009. Model-free estimation of the psychometric function. *Atten Percept Psycho*, 71, 1414–1425.