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

Vincent Aubanel[a*], Maria Luisa García Lecumberri[b], Martin Cooke[a]

[a] Ikerbasque (Basque Foundation for Science), 48011, Bilbao, Spain
[b] Language and Speech Laboratory, Universidad del País Vasco, Paseo de la Universidad 5, 01006 Vitoria, Spain
[*] corresponding author. E-mail: v.aubanel@laslab.org

Keywords: Speech perception in noise, Phonetic balance, Spanish, Open speech resource.

Abbreviations:

SNR	Signal-to-noise ratio
SRT	Speech reception threshold

Submitted to the International Journal of Audiology, July 2013

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 Spanish talker. Sentences each contain 5 keywords for scoring and are placed into lists of 10 using an automatic phoneme-balancing procedure. *Study Sample*: 23 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.

1 Introduction

Although Spanish – after Mandarin – is the language spoken by the largest number of native speakers (Lewis, Simons and Fennig 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 approximately-equivalent to the Harvard sentence material (Rothauser et al. 1969) which is widely-used in speech perception tests (e.g., Bradlow, Torretta and Pisoni 1996; Hawley, Litovsky and Culling 2004; Hu and Loizou 2010; Cooke et al. 2013). The Harvard Corpus consists of phonemically-balanced lists of 10 sentences, where each sentence contains 5 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 (i) the Castillian Spanish hearing in noise test (Huarte 2008) based on a list of 240 Spanish sentences adapted from the English HINT test (Nilsson, Soli and Sullivan 1994); (ii) 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); and (iii) 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. All of these corpora are currently subject to certain limitations which constrain their wider usage, e.g., licensing restrictions or lack of published lists and speech recordings.

In addition to the sentence lists themselves, 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 maximises phonemic balance. The outcomes of a speech-in-noise test using the Spanish material is also reported.

2 The Sharvard Corpus

2.1 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. During this process part of the original English sentences were adapted for Spanish use 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 with the tip of the boot and with strength”). New

sentences were created where no reasonable Spanish translation was felt appropriate (e.g., “Mesh wire keeps chicks inside”).

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 reduce any predictability advantage which would arise from longer words. 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 and Mariño (1993). The pronunciation dictionary makes use of 31 symbols.

A phonemic level of analysis was employed in order to accommodate for pronunciation variations in any speech recordings of the sentence lists. To acheve this, the 7 common allophones were merged with their main phoneme category, as shown in Table 1. Phonemic balancing was therefore subsequently conducted on 24 phonemes.

2.2 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 from the phoneme frequency distribution in the language are evident, largely as a consequence of the omission of segments contained in high-frequency function words such as articles and pronouns.

[1] http://www.talp.upc.edu/index.php/technology/tools/signal-processing-tools/81-saga. Last viewed July 10, 2013

2.3 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 optimisation 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:

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

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 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.
Figure 2 shows the result of the optimisation procedure. Data are shown both pre- and post-optimisation as means over 10 independent runs with different randomised initial sentence groupings. The balancing procedure reduces the initial imbalance by a factor of five. Results are also shown for 5-sentence lists to illustrate the difficulty of achieving good phonemic balance with smaller subsets.

Similar data for the original Harvard Corpus is given for comparison, 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

[2]URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Last viewed July 10, 2013.

significantly with the current balancing algorithm. Only phonemes occurring in keywords were considered, but similar results were obtained based on all phonemes in the sentences.

Note that while the balancing method uses a linear combination of individual phonemes, it can be straightforwardly extended to provide a balancing of the sets with regard to other sets of segments, other attributes such as keyword lexical frequency, or their combination.

The sentence lists are provided as supplementary material and are available in the online version of the journal (see section 4).

3 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 a range of SNRs.

3.1 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 Sharvard Corpus. Talkers were asked to read each sentence at a normal speaking rate and to pause between utterances. Talkers were able to repeat any utterance where necessary. The sentence order was randomised – 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. Sentences were normalised by dividing the signal by the maximum amplitude observed for that talker in the entire corpus, and saved as individual WAV format files.

3.2 Talker intelligibility

A listening experiment was carried out to assess the overall intelligibility in noise of the Sharvard speech material. Psychometric functions were computed by measuring the proportion of keywords identified correctly in noise by native talkers as a function of signal-to-noise ratio (SNR). The two talkers were assessed independently using speech-shaped noise maskers whose long-term average speech spectrum matched each individual talker, presented at 11 SNR values linearly spaced from -11 dB to -1 dB, values chosen in pilots to produce keyword scores of 10 to 90 %.

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 (mean age=22.3 years, s.d.=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 11 SNR levels. Eleven 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 min, including a short practice session.

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). Model-free fits (Zychaluk and 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. Another possibility is that the individual talkers chosen to produce the two corpora differ in intrinsic intelligibility.

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 sentence material is available as a supplementary material in the online version of the journal, through the direct link to the article at http:///www.informaworld.com/(DOInumber). Phonemic transcriptions and spoken recordings of the entire corpus from one male and one female native Spanish talker are available online for unrestricted usage at http://laslab.org/resources/sharvard.

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.

References

Bradlow, A. R., Torretta, G. M. and Pisoni, D. B. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Commun*, 20(3), 255–272.

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 Commun*, 55, 572–585.

Hawley, M. L., Litovsky, R. Y. and 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(2), 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(7), 536–544.

Hu, Y. and 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(1), 427–434.

Huarte, A. 2008. The Castilian Spanish hearing in noise test. *Int J Audiol*, 47(6), 369–370.

Lewis, M., Simons, G. F. and Fennig, C. D. 2013. *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International. URL: http://www.ethnologue.com.

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

Moren Sandoval, A., Toledano, D. T., de la Torre, R., Garrote, M. and Guirao, J. M. 2008. Developing a phonemic and syllabic frequency inventory for spontaneous spoken Castilian Spanish and their comparison to text-based inventories. *Proceedings of the 6th Language Resources Evaluation Conference (LREC)*. Marrakech, Morocco.

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

Nilsson, M., Soli, S. D. and 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(2), 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 Acoust*, 225–246.

Zychaluk, K. and Foster, D. H. 2009. Model-free estimation of the psychometric function. *Atten Percept Psychophys*, 71(6), 1414–1425.

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 16333. The merging of the 7 allophones with their main phoneme categories is detailed in the rightmost column.

sound class	IPA	frequency (%)	mergers
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)
	g	1.87	(g: 0.88, ɣ: 0.99)
fricative	f	1.16	
	θ	1.65	
	s	6.48	(s: 6.41, z: 0.07)
	ʃ	0.42	
	x	1.44	
affricate	ʧ	0.94	
nasal	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	

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

List of Figures

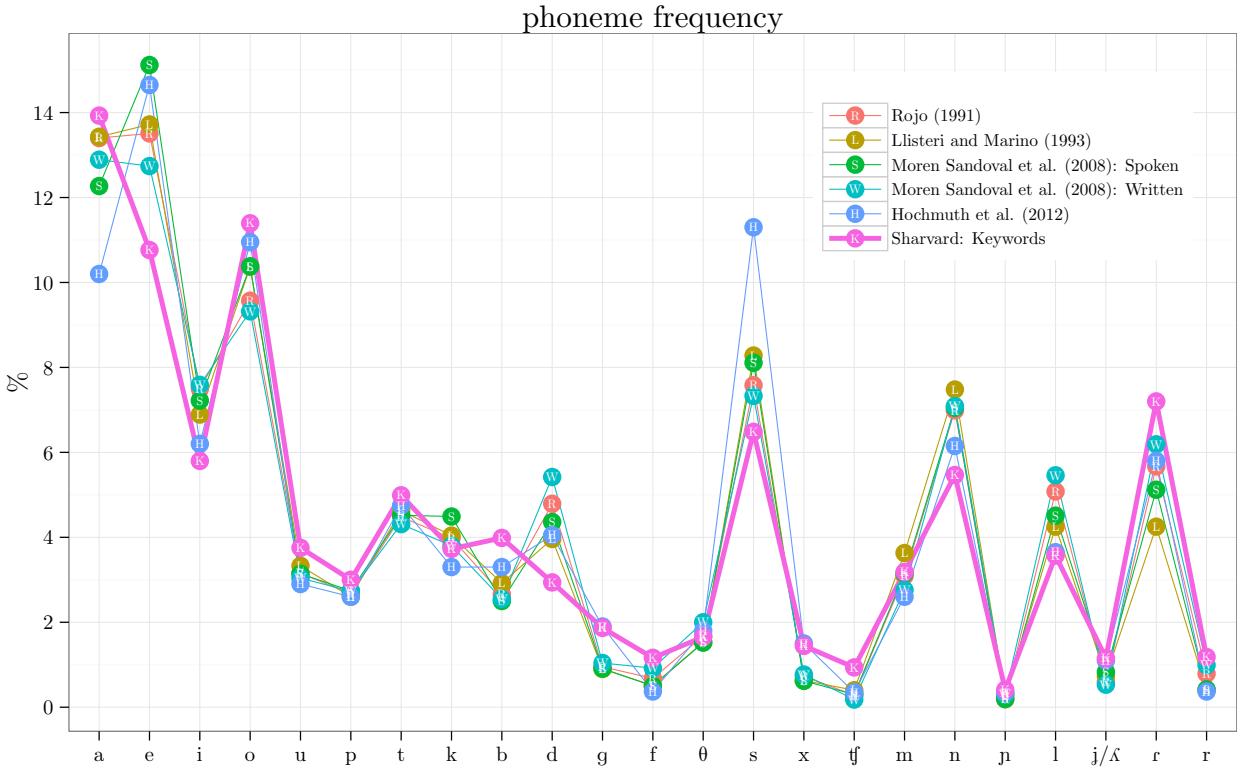

Figure 1: (Color online) Phoneme frequency distribution of keywords in the Sharvard Corpus. To accomodate for inventory disrepancies across corpora, counts for /ʎ/ and /ʝ/ are aggregated in this figure. Other sets are plotted for comparison: **Rojo (1991)**: 3.8 million words corpus with a variety of Castillan and Latin American Spanish written texts. Presented frequencies are the ones reported by Llisterri and Mariño (1993), adjusted to redistribute archiphoneme frequencies to map with their phonetic inventory of Spanish; **Llisterri and Mariño (1993)**: 100k phonetic segments automatically derived from orthographic transcription of three hours of semi-spontaneous speech provided by 3 native Spanish speakers; **Moren Sandoval et al. (2008) (Spoken)**: Spanish C-ORAL-ROM corpus, consisting of 42 hours of recorded speech by 429 speakers covering 3 styles of speech (informal, formal, media), containing 348k orthographically transcribed words with automatically-produced phonetic transcriptions; **Moren Sandoval et al. (2008) (Written)**: 480k 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 Fig. 1 of Hochmuth et al. (2012).

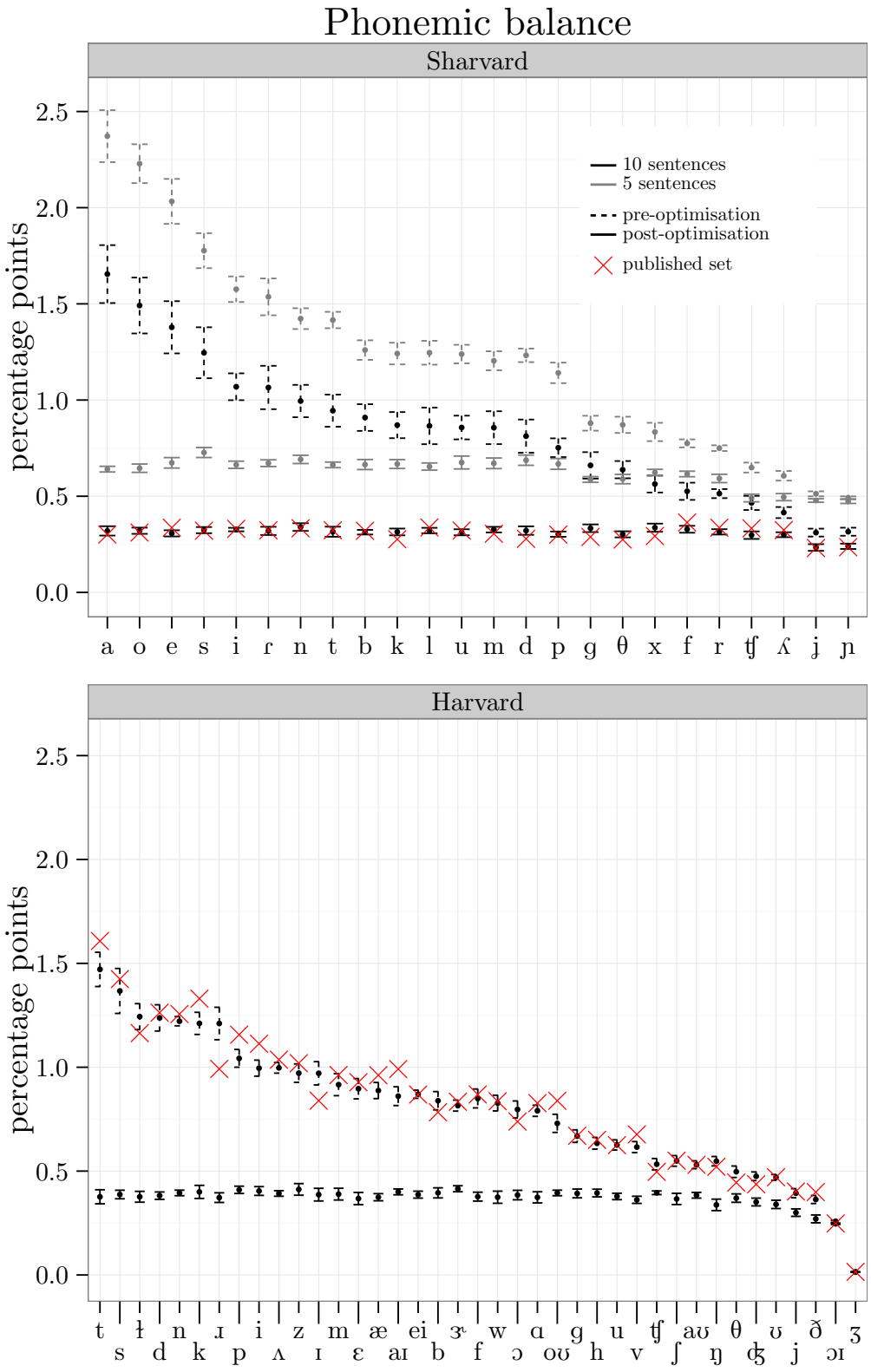

Figure 2: (Color online) Phonemic-balance for the Sharvard (upper) and Harvard (lower) corpora. The measure of phonemic balance for each phoneme is the mean over subsets of the magnitude of the difference between the frequency of that phoneme in the corpus as a whole and in the subset. Phoneme order is based on imbalance pre-optimisation. Error bars depict 95 % confidence intervals computed over 10 runs with different random initialisations. Red crosses are across list means for the published 10-sentence sets.

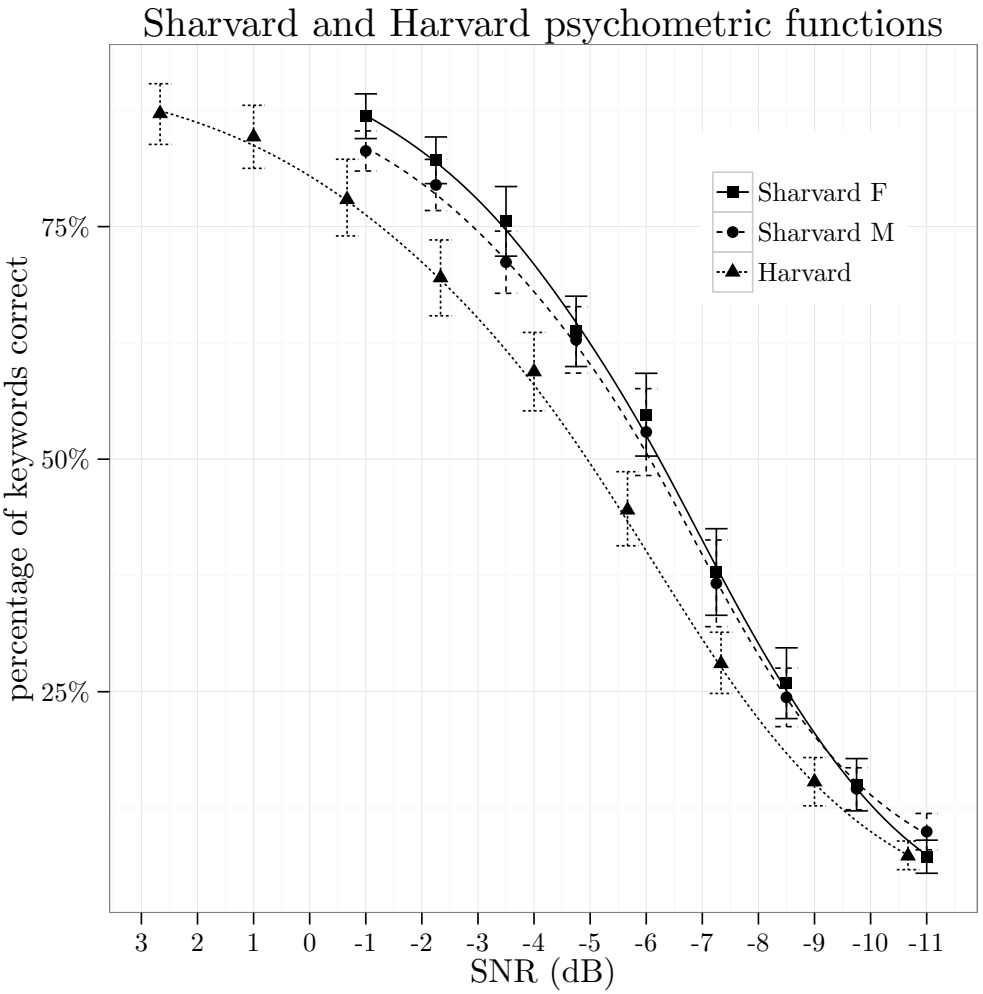

Figure 3: Psychometric functions for the male (M) and female (F) talkers of the Sharvard Corpus, along with those from a male talker of the original Harvard Corpus.