The Fharvard corpus: A phonemically-balanced French sentence resource for audiology and intelligibility research

Vincent Aubanel, Clémence Bayard, Antje Strauss, Jean-Luc Schwartz

The Fharvard Corpus: A phonemically-balanced French sentence resource for audiology and intelligibility research

Vincent Aubanel[a,1], C. Bayard[a], A. Strauß[b], J.-L. Schwartz[a]

[a] *University of Grenoble Alpes, CNRS, GIPSA-lab, Grenoble, France*
[b] *University of Konstanz, Konstanz, Germany*

Abstract

The current study describes the collection of a new phonemically-balanced sentence resource for French, known as the Fharvard corpus. The resource consists of 700 sentences inspired by the original English Harvard sentences, along with audio recordings from one female and one male native French talker. Each of the sentences contains five mono- or bisyllabic keywords and are grouped into 70 lists of 10 sentences using an automatic phoneme-balancing procedure. Twenty-three normal-hearing French listeners identified keywords in the Fharvard sentences in speech-shaped noise. Psychometric functions for the Fharvard sentences indicate mean speech reception thresholds of -4.48 and -3.87 dB and slopes of 10.55 and 12.52 percentage points per dB at the 50 % keywords correct point for the female and male talkers respectively. The complete list of Fharvard sentences and the associated audio recordings are available online for speech perception testing.

Keywords: Speech perception; Noise; Psychoacoustics/Hearing Science; Behavioural Measures; Speech Intelligibility; Speech database; French

Email address: vincent.aubanel@gipsa-lab.fr (Vincent Aubanel)
[1] Corresponding author.

With close to 300 million talkers, French is the fifth most spoken and the second most learned language in the world [40], with predictions for an increase in the proportion of French talkers in the near future[2]. In contrast, resources for administering audiology tests are few, with some now out of date or with limited availability to both practitioners and researchers.

While basic phonetic decoding abilities can be evaluated using word- or sub-word-level lists, for which there are a few commonly used resources available for French [e.g., 17, 15, 25], higher order speech comprehension requires longer and more complex sequences, that is, sentence-long material. Sentence-long materials were originally developed in the context of speech transmission technologies in the fifties for American English [e.g., 35] and since then, have been a material of choice for speech intelligibillity and hearing research [see, e.g., 33, 27, 2, 22, 24, 13, for a short list of example studies for both English and Spanish]. Their usefulness in clinical settings has been demonstrated in a range of applications, from the assessement of hearing aids to the evaluation of speech comprehension in cochlear-implanted users, and to the evaluation of functional hearing abilities in the workplace [see a detailed list in 38]. Linguistic material longer than the word is also ideally suited to study multisensory speech perception, such as the benefit of cued speech for hearing-impared populations [7]. Finally, sentence-long material has also recently received a renewed interest in cognitive neurosciences [18], as it enables researchers to focus on more naturalistic and ecologically valid material than was previously used and to address new theoretical issues relating to temporal and hierarchical aspects of speech perception [see, e.g., 36, 3, 29, and references therein].

Three sets of sentences lists are referenced in the speech audiometry catalogue edited by the French National Hearing Aid College (Collège National d'Audioprothèse),

[2] https://www.francophonie.org/La-langue-francaise-dans-le-monde-49280.html. Last viewed 8 Apr. 2019

one of the main references for current audiology practice in France[3]. The first set has been developed by [17] as part of his pioneering work in establishing speech audiometry in France. The Fournier lists consist of 10 sets of 10 relatively short sentences with a fixed grammatical structure (Subject, Verb, Object) and are, as such, relatively predictable. The Verb and Object are tagged as keywords and the sentences were not phonemically balanced, neither at the keyword nor the sentence level. The second set contains the Combescure sentences [11], a set of 200 sentences, phonemically balanced to match the distribution of phoneme frequency in French and assembled in subsets of 10 sentences. There is a broad range of variation in sentence length, degree of difficulty and language style across sentences (compare *Ce petit canard apprend à nager* 'This little duck is learning to swim' with *A la hâte, le métayer ensilait ses récoltes avant l'hiver* 'Hastily, the metayer was ensiling his harvest before winter'), with some terms having become outdated. Because of the high degree of care that went into the making of the sentences, this resource is probably the most often used to date and is also referenced in the Good Practice Guide published by the French Society of Audiology (Société Française d'Audiologie)[4]. The third set, recently proposed by [38], is an adaptation of the American English Hearing-In-Noise-Test [HINT, 32] to Canadian French, referred to hereafter as HINT-Fr. This corpus consists of 240 sentences balanced into sets of 20 sentences produced by one male speaker (a professional voice actor). As the initial HINT sentences were an adaptation of the BKB (Bamford-Kowal-Bench) sentences [8] which were designed to be used with British children, the HINT-Fr sentences have a relatively simple structure, a lexical field adapted to use with children and include lexical items specific to Canadian French (e.g., *Il mange de la crème glacée*, 'He's eating ice cream'). A version suited to European French talkers was subsequently developed by [26], by selecting sentences from an initial set of 589

[3] http://www.college-nat-audio.fr/listes-cd-audiometrie-vocale.html. Last viewed: 8 Apr. 2019

[4] https://sfaudiologie.fr/?q=node/47. Last viewed: 8 Apr. 2019

sentences read by one male speaker in a two-step process: first on the basis of equal intelligibility scores across a cohort of hearing-impaired listeners, then according to the homogeneity of their score across a second cohort of normal-hearing listeners. The resulting set of 140 sentences is not publicly available. In addition to pre-established sentence lists, a sentence generator, named FrMatrix, has been proposed by [21], where sentences can be created by combining 5 words in sequence, each from a distinct 10-element closed-set category (Name, Verb, Numeral, Object, Color, in that order).

With the exception of FrMatrix, where individual word scores can be combined to obtain a combined sentence score, sentences are usually evaluated using a single binary score, e.g., a sentence is scored as correct only if all words are correctly recognised. Apart from its low granularity (i.e., only one value per sentence), this method of scoring has a low sensitivity as a sentence receives a score of 0 whether 10% or 90% of its words have correctly been recognised by the listener. In order to improve the sensitivity of this scoring method, a list of acceptable variations for non-keywords was included in the HINT-Fr lists, introducing some degree of complexity to the material, while maintaining the same granularity, that is, a single binary value per sentence.

Motivation for a new linguistic and audiological material

As reviewed above, a number of resources have been developed over the years for testing sentence-level intelligibility, but no resource can concurrently meet the basic criteria required for quality evaluation of speech intelligibility for French talkers in modern day practice and research settings. These criteria combine:

- adequate size and variability
- difficulty homogeneity within the proposed material
- lexical diversity corresponding to contemporary French usage
- the possibility to scale intelligibility at the word level inside the sentence
- phonemically balanced material

- reference speech reception thresholds (SRTs)
- availability of the database and the associated linguistic and phonemic content

This is the gap that the current work aims to fill. To that end, we capitalise on a previous series of developments of such resources, first in English with the Harvard database [35], then in Spanish with its Sharvard adaptation [4]. This leads to two contributions. Firstly, we describe the French adaptation of the Harvard corpus, the Fharvard corpus, which consists of a list of 700 unique sentences, a size deemed large enough to allow to use the same material for long and/or multiple tests with the same participants. Three native French speakers translated and adapted English Harvard sentences, resulting in a contemporary usage of French. In creating the sentences, we aimed to maintain the variability in syntactic structure, difficulty homogeneity and lexical diversity already found in the base material, and a strict 5-keyword structure was respected to allow for a graded scoring of sentences at the word level uniformely throughout the corpus. The final corpus is made by grouping the sentences into 70 phonemically-balanced sets of 10 sentences, following an optimisation procedure that minimises the difference in phoneme distribution across lists. . Secondly, we provide audio recordings of the whole material by one female and one male talker and we evaluate the intelligibility of a subset of this material in noise at various signal-to-noise ratios (SNRs), providing reference speech reception thresholds (SRTs) with normal listeners for the two talkers. Both the sentence list material and the audio recordings are made freely available for download and use (see Section *Availability of the corpus*)

Contribution 1 - A phonemically balanced linguistic corpus for evaluation

This section provides the methodology used to construct the Fharvard linguistic database, and the evaluation of its distributional properties at the phonemic level.

Sentence material

Sentences were constructed in a similar way to Sharvard sentences [4]. As a starting point, English sentences were translated into French to maintain a similar level of difficulty. Then each sentence was modified freely, which sometimes resulted in an altogether new sentence being created so that it contained 5 keywords of at most 2 syllables. This syllabic constraint was used to limit predictability effects associated with words having three syllables or more. Note that for the purpose of syllable count, word-internal schwas, which can be supressed or produced in Northern and Southern varieties of French respectively, were discarded, e.g., the word *fièrement* /fjɛʁ(ə)mɑ̃/ 'proudly' was counted as having two syllables. Keywords were defined with reference to a list of non-keywords, which were established as a list of common function words (articles, pronouns, prepositions and conjunctions, e.g., *la* 'the' (fem.), *elle* 'she', *de* 'of', *avec*, 'with' respectively) and frequent occurrences of common verb forms, e.g., *a* 'has', *sont* 'are', *était* 'was'.

Phoneme frequency distribution

The phonemic annotation was obtained with the phonetiser module of the EasyAlign toolkit [19]. For the purpose of phonemic balancing, a phonological representation was obtained for each keyword by discarding word-final schwas and liaison consonants, as their phonetic realisation may vary from speaker to speaker. Table 1 shows the phoneme distribution for the keywords of the 700 sentences. Counts for infrequent phonemes /œ̃/ (5 occurrences) and /ŋ/ (4 occurrences) were added to the more frequent /ɛ̃/ and /n/ phoneme categories respectively.

[Table 1 about here.]

The phoneme frequency distribution of **Fharvard keywords** was compared with phoneme frequency distributions for French computed on corpora covering various speech styles. **[39]** is the first large-scale report of phoneme frequencies in French,

built from close to 12 hours of drama and spontaneous speech on television, totaling 200,000 occurrences of phonemes. **Lexique3** [31] is a widely used resource for psycholinguistic research offering word frequencies estimated from a 14.8 million word corpus of novels and essays and a 50 million word corpus of movie subtitles [30]. We use the latter to compute phoneme frequencies, as word frequencies estimated on movie subtitles were found to better represent language use than that computed on books [30, 10]. [1] compared phoneme frequencies of three corpora, with a view to characterise linguistic differences across speech styles, using automatic speech recognition techniques. Phoneme frequencies were estimated for a corpus of telephone conversations (**Convtel**, 120 hours), broadcast news (**Journ.**, 25 hours) and a 32 hour subset of the **PFC** corpus (Phonologie du Français Contemporain) [16], a corpus aimed at characterising phonological variation in French, mixing read and spontaneous speech. We also included data from the **CID** corpus (Corpus of Interactional Data) [9] for its focus on speech-in-interaction, providing semi-automatic annotation at the phonemic (totaling approx. 200,000 phonemes), prosodic, morphosyntactic, syntactic, discursive and mimo-gestual levels.

We also compared Fharvard keywords' phoneme frequency distribution with that of other published sentence material used in audiology and intelligibility research. Phoneme frequencies were computed on the **Combescure** sentences [11] after they were converted to their phonological form using the phonetiser module of EasyAlign [19]. We also included published phoneme distribution of **HINT-Fr** [38] and **FrMatrix** [21].

Figure 1 shows the comparison of the phoneme frequency distribution of Fharvard keywords with that of other resources, distinguishing *corpus*-based (upper panel A) and *material*-based (lower panel B) distributions. In both cases, we chose Lexique3 as the reference corpus, given its representativity of phoneme frequency in common French usage. A few differences are apparent between corpora (Figure 1 A), mainly

stemming from the frequency difference of specific lexical items characteristic of the corpora's speech styles. For example, conversational speech is characterised by increased occurrences of words such as *moi/mmmh, oui/ouais, ah!, et* 'me/hum, yes/yep, oh!, and' leading to a relatively higher proportion of /m, w, a, ɛ/ respectively in the Convtel and CID corpora. The high frequency of /ɛ/ in the CID corpus is a consequence of its merging with the other frequent phoneme /e/ in that corpus (see full list of mergers in Figure 1 caption). Conversely, phonemes such as /l/ and /d/ are slightly under-represented in Fharvard keywords since, by construction, frequent function words such as *le/la/les* 'the (fem.)/the (masc.)/the (plural)' and *de/du* 'of/of the' respectively were not included in the phoneme counts for that corpus. The high ranking of /ʁ/ in Fharvard keywords may also be a consequence of its greater representation in content words, as it also reaches a high value in written text-based corpora such as Lexique3 and Journ., and a lower value in the conversation-based corpora CID and Convtel. Finally, we note that the variability of /ə/ can have different origins: in the case of Fharvard, the low ranking is the result of discarding highly frequent function words containing schwas (e.g., *le/de*) and the deliberate suppression of final word schwas in phonological representations of keywords. The relative higher frequency for other corpora is likely the result of increased occurrences of schwas in a conversational setting (CID, Convtel), the result of the merging with other mid-vowels (CID, Journ., Convtel, PFC), or a combination of these factors.

As can be seen in Figure 1 B, there is an overall similar degree of agreement across materials and with the reference corpus Lexique3. However, individual phoneme departures from the reference distribution appear to be less systematic as in the case of corpora: while a relative increase of /ʁ/ in FrMatrix and Fharvard and /l/ in HINT-Fr are likely attributed to an increased proportion of content words and function words respectively, the departure of /k/ in HINT-Fr and of /z/ in FrMatrix is less easily explained. In the case of vowels, the discrepancy between materials seems to be the

result of a difference in the coding in mid-vowels: a greater relative proportion of /ɛ/ in the Combescure database is compensated by a relative lower proportion of /e/, with similar compensatory effects for /e/ vs. /ɛ/ in HINT-Fr and /o/ vs. /ɔ/ in FrMatrix. In sum, while this review of existing corpora shows that there is no gold standard for phoneme distribution for French, given its dependence on speech styles in particular, we show that the phonemic composition of the Fharvard corpus accurately represents the phoneme distribution for French, equating and sometimes exceeding in that respect other published materials.

[Figure 1 about here.]

Phonemic balancing of the Fharvard corpus

Phonemic balance is classically defined as the degree of agreement of the phoneme distribution of a subset of a corpus to a reference distribution, whether that distribution is representative of a reference language, or it is computed on the whole study corpus. In the current work, we distributed the Fharvard sentences into subsets of 10 sentences to match the distribution of the complete set of 700 sentences, as shown in Figure 1. The phonemic balance β is measured for a given list L of 10 sentences as the distance in the Euclidian P-space (here, P=33 phonemes) between that list and the corpus C. That is,

$$\beta = \sqrt{\sum_{p=1}^{P} (f_{p,L} - f_{p,C})^2}, \tag{1}$$

where $f_{p,L}$ and $f_{p,C}$ are the frequency of the phoneme p in the list and the corpus, respectively. We used the optimisation procedure for phonemic balance described in [4], which consisted of, starting from an initial ordering of the sentences, iteratively swapping two sentences from two lists if the swapping decreased the phonemic balance of both lists. For the current corpus, minimal phonemic balance was obtained after around $5,000$ iterations. Subsequent runs of the optimisation procedure using the balanced set as a starting point decreased marginally the average phonemic balance of the

lists and no further improvement was obtained after 4 runs of the procedure. Since the procedure minimises the unweighted sum of the phonemic balance over all phonemes, every phoneme contributes equally to the global minimisation. This contrasts with the standard approach for comparing two frequency distributions, the Kullback-Leibler divergence measure [23]. While this metric is well-suited when applied to large corpora, in particular as it introduces a correction for very low frequency values, this property is not desirable when applied to small samples, as differences between low-frequency values are then over-represented, and numerical corrections for zero-values become necessary. The Euclidian distance, which does not suffer from theses shortcomings, is the method used in the Sharvard corpus [4] and similar approaches [32]. Figure 2 shows the by-phoneme phonemic balance of the corpus pre- and post-optimisation. The procedure significantly reduced the phonemic balance from an average value of 0.85 (SD = 0.25) pre-optimisation to 0.36 (SD = 0.03) post-optimisation (paired t-test: $t = 11.12$, df = 32, $p < 0.001$). This represents a 2.3-fold reduction in mean phonemic balance and a 9.1-fold reduction in its standard deviation, values which are comparable to those obtained for the Sharvard and Harvard corpora [see 4].

To provide a further indication of the validity and quality of our method, we applied our procedure to an already phonemically balanced sentence material, the Combescure sentences [11]. The phonemic balancing of this material in the original publication was operated by manual reorganisation of the sentences using the chi-squared statistics as the minimisation parameter. We expected the pre-optimised β of that material to have an already low value, but we did not have any hypothesis as to whether our procedure would reduce further that value. We found that the phonemic balance of this corpus could only marginally (albeit significantly) be improved (reduction from 0.42 (SD = 0.22) pre-optimisation to 0.34 (SD = 0.16) post-optimisation, paired t-test: $t = 4.43$, df = 32, $p < 0.001$). The relatively low value of pre-optimised β shows the quality of phonemic balance in that material. The fact that the β value can

be reduced further by our procedure additionally suggests that our method is slightly superior to the one employed there, probably owing to the possibility of exploring a much larger combination of permutations than what can be practically achieved in a manual procedure. This further confirms the validity of our approach.

[Figure 2 about here.]

Contribution 2 - An evaluated acoustic database for audiology and intelligibility assessment

This section presents the acoustic database associated with the linguistic material together with the results of a listening experiment aiming to provide some basic speech audiometry characteristics for the Fharvard corpus, for use in both speech intelligibility research and audiology practice settings.

Recording of the corpus

Sentences were recorded by two talkers, a female and a male in their early thirties and sixties respectively at time of recording, both chosen for the clarity of their speech. While both talkers have a broad standard French accent, they both present mild features of Northern and Southern French varieties respectively, a difference mainly reflected by the greater number of schwas realised as full syllables for the male talker compared to the female talker.

Both talkers provided written consent and each recorded the entire set of sentences by reading at their preferred pace the printed lists of sentences. Recordings were made in a sound attenuated room of the lab and speech was recorded with an AKG C1000S microphone and digitised with a DPS Reality acquisition card. Sentences were subsequently manually verified and segmented into individual files. We checked that the entire set of 700 produced sentences was exempt of audible acoustic perturbations, mispronunciations or hesitations by either talker.

On average, the female talker produced the sentences with a duration of 2.47 s ($SD = .27$), and a speaking rate of 3.78 words/s ($SD = .44$). The male talker produced the sentences with an average duration of 2.53 s ($SD = .29$), corresponding to a speaking rate of 3.71 ($SD = .47$). A paired t-test on both parameters revealed that the female talker produced sentences with significantly shorter durations (mean difference: $-.056$ s, $t(699) = 8.05$, $p < .01$) and faster speaking rate (mean difference : .074 words/s, $t(699) = -7.19$, $p < .01$) than the male talker possibly in relation with the tendency for the male talker to produce more schwas.

Stimuli

A subset of 360 sentences was randomly selected from the initial 700-sentence corpus, and presented to listeners in a speech in noise recognition task. Half of the sentences of this selection were spoken by the female talker, and the other half by the male talker. Sentences were mixed with a stationary speech-shaped noise masker, which was computed independently over recordings of all sentences for each talker, so that the long-term average speech spectrum of the masker matched that of the respective talker. Speech-plus-noise mixtures were constructed for 9 different SNR values, linearly spaced from -11 dB to -1 dB (i.e., -11, -9.75, -8.5, -7.25, -6, -4.75, -3.5, -2.25, -1 dB) as these values were expected to span the range of 10 to 90 % of correctly identified keywords. Mixtures were constructed by mixing each sentence with a portion of masker of duration 1 s longer than the sentence, so that the sentence was preceded and followed by 500 ms of masker. Speech level was scaled to reach the required SNR for the time interval where the sentence overlaps with the masker. The mixtures were presented binaurally at a fixed level of 79 dB SPL, measured with a Bruel & Kjaer artificial ear (model 4153) equipped with a Bruel & Kjaer microphone (model 4134) coupled to a Bruel & Kjaer NEXUS system over Sennheiser HD 280 pro closed headphones.

Participants and procedure

Twenty-five subjects were recruited in the student and staff population of the Speech and Cognition Department of the lab and received a gift card for their participation. Following hearing screening, 23 subjects (12 females and 11 males, mean age: 23.8, SD=3.14) 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.

In each session, the 180 stimuli were presented in 20-sentence blocks for each of the 9 SNR levels. Sentence order was randomised across blocks for each participant, and blocks were assigned to participants 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 they pressed 'Enter' to launch the next stimulus. Each session lasted around 45 minutes, including a short practice session.

Results

Responses were corrected automatically for common alternative word forms, including homophones (around 6000 forms obtained from the lexique.org database). An additional dictionary including common spelling mistakes and digit input for numbers was compiled following manual review of listeners responses, and totalled 70 entries. The mean percentage of correctly identified keywords is plotted in Figure 3 alongside data collected in similar settings for British English [14] and Spanish [4]. Psychometrics functions were estimated for each talker indepentently using a nonparametric approach using local linear fitting [41] and are also shown in Figure 3. Speech reception thresholds (SRTs) for 50 % correct keywords and associated slopes were obtained by getting the SNR level and slope values corresponding to a performance level of 50 % on the psychometric function, fitted separately for each participant. Table 2 shows the

mean and standard deviation of these values over participants.

[Figure 3 about here.]

[Table 2 about here.]

There was a small but significant difference between the 50 %-SRT across talkers [$t(22) = 6.89,\ p < .01$], with the male talker's SRT being 0.69 dB higher than the female talker. Similarly, the slope for the male talker was slightly but significanlty steeper than that for the female talker [mean difference: 2.34 %/dB, $t(22) = 4.73$, $p < .01$].

The current design used a random selection of sentences per SNR level to provide an overall evaluation of intelligibility of the sentence material and as such precludes from directly assessing intelligibility variability within and across the lists of 10 phonemically-balanced sentences. This design, howevern, enabled us to compute an estimate of the effect of list size on the modelling of responses. To this end, we fitted a generalized linear mixed-effects model with a logit link function to the keyword responses [function `glmer()` in the R package `lme4` [6]], specifying a random intercept by *list* and by *participant* and taking *talker* and *SNR level* as fixed effects. We explored the arbitrary grouping of sentences into lists ranging from 5 to 20 sentences per list (i.e., 5, 10, 15, 20), and fitted a model for 20 random combinations of sentences for each grouping. As is shown on Table 3, we found that the size of the variance associated with the *list* factor decreased from 5 to 20 sentences per list and was in the same range as that for the *participant* factor for a grouping value of 10 sentences per list. This suggests that grouping Fharvard sentences in lists of 10 elements provides a good compromise between testing time and sensitivity.

[Table 3 about here.]

Discussion

The present study reports on two contributions presented at the beginning of the paper. The first one is a database of 700 sentences of similar syntactic and semantic complexity, each containing 5 keywords enabling intelligibility scoring in various experimental designs, and grouped in 70 sets of 10 sentences with a balanced distribution of their phonemic inventory. The second contribution consists of the complete recordings of the database by one male and one female speaker, together with a perceptual evaluation of the intelligibility of these two recordings with different level of speech-shaped noise. We now turn to a discussion of some aspects and potential limitations of these two contributions.

Regarding the specification of the linguistic material, we identified a range of criteria that make up a corpus that could be used in a wide variety of situations (see *Introduction*). However, satisfying all criteria at once is unpractical and some compromise has to be reached. For example, we controlled difficulty homogeneity by relying on the combined judgement of three native French speakers when creating the sentences, in lieu of an evaluation of the complex interactions between lexical, syntactic, semantic and psycholinguistic properties, which are difficult to quantify accurately [28]. Recent statistical approaches such as [37] may be considered to enrich the current corpus with an evaluation of individual sentence complexity.

Similarly, lexical diversity and homogeneity of the sentence material was determined partly by the Harvard sentences base material and partly by the creativity of the native speakers. Finer control could in principle be exerted by, for example, specifying limits for the lexical frequency of individual keywords during word selection and for the number of occurence of words in the corpus, adding some degree of complexity in the creation process. We note however that the initial size of the corpus allows for a refinement of lexical aspects, through the selection of subset of sentences, the addition of a lexical-related parameter to the balancing procedure, or both.

Regarding the acoustic database and its perceptual evaluation, we must first stress that individual sentence intelligibility ultimately depends on a range of factors beyond the sentence composition itself, including the particular recording conditions, talker intrinsic intelligibility and listeners' sociolinguistic background to name a few. Specific uses of the corpus requiring a finer calibration of individual sentence intelligibility (i.e., requiring a higher SRT slope, see below) could be addressed by operating subsequent sentence selection of the initial corpus, as was done in [26], resulting however in a drastic reduction in final corpus size.

Turning to intelligibility results, we find that, as shown in Figure 3, speech reception thresholds (average of -4.3 dB across the two talkers) and associated slopes (12.54 %/dB) for French span a similar range as those obtained for English (-4.94 dB; 8.96 %/dB) and Spanish (-6.16 dB; 10.78 %/dB) with this similar type of material [4]. Language factors such as differences in phoneme size inventory or lexical syllabic distribution in the three languages could explain the variation. For example, a lower proportion of bisyllabic than trisyllabic words in French as opposed to English and Spanish [respectively: 12, 20, 34] could suggest a reason for the tendency for French to require a higher SRT than the other languages in the task of identifying (comparatively less frequent) mono- or bisyllabic words in noise. However, it should be noted that across-language variation is of the same order of magnitude than across-talker variation, which could explain this tendency instead. Indeed, it is well known that individual talkers vary in their intrinsic intelligibility, owing to differences in various dimensions such as voice quality or speaking rate [5].

The average slope across talkers of 12.54 %/dB, which can be taken as a sensitivity measure is also comparable to that obtained with other French material such as FrMatrix (14.0 %/dB), but lower than that obtained in the FIST corpus (20.2 %/dB). This latter corpus was, however, specifically constructed to maximise sensitivity by an iterative sentence selection process based on equal intelligibility, and as a result

comprises a limited set of sentences, which are additionally variable in length and not phonemically balanced. Indeed, apart from offering a good level of sensitivity for assessing speech audiometry, the Fharvard corpus presents the additional advantage of phonemically balanced sets of sentences, a constant number of keywords per sentence for word-level scoring, and a greater corpus size allowing, for example, to test participants in several sessions without the need to reuse sentences.

Summary

A new resource for audiology and intelligibility research in French is presented, based on the Harvard sentence material. The Fharvard corpus is a collection of 700 sentences, grouped into 70 lists of 10 sentences, each list with a balanced phonemic content across keywords. The original English sentences were translated and freely adapted to French, maintaining a five mono- or bisyllabic keywords criterion and a homogeneous level of difficulty for testing with adults. The whole material was recorded by two speakers and a representative subset was evaluated for intelligibility in noise by normal-hearing listeners.

Availability of the corpus

The list of sentences is available in the online supplementary materials. The audio recordings of the sentences by a female and a male talker is available at http://dx.doi.org/10.5281/zenodo.1462854.

Acknowledgements

This work was supported by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s"). We thank Laura Machart for her help in collecting data.

References

[1] Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Proc. of Journées d'Études sur la Parole (JEP)*, pages 389–400, Dinard, France.

[2] Aubanel, V. and Cooke, M. (2013). Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. In *Interspeech*, pages 3592–3596, Lyon, France.

[3] Aubanel, V., Davis, C., and Kim, J. (2016). Exploring the role of brain oscillations in speech perception in noise: intelligibility of isochronously retimed speech. *Front. Hum. Neurosci.*, 10(430).

[4] Aubanel, V., Lecumberri, M. L. G., and Cooke, M. (2014). The Sharvard Corpus: A phonemically-balanced Spanish sentence resource for audiology. *International Journal of Audiology*, 53:633–638.

[5] Barker, J. P. and Cooke, M. (2007). Modelling speaker intelligibility in noise. *Speech Commun.*, 49(5):402–417.

[6] Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.*, 67(1):1–48.

[7] Bayard, C., Machart, L., Strauss, A., Gerber, S., Aubanel, V., and Schwartz, J.-L. (2019). Cued Speech enhances speech-in-noise perception in deaf with cochlear implants. *J. Deaf Stud. Deaf Edu.*, 24(3):223–233.

[8] Bench, J., Kowal, Å., and Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *Brit. J. Audiol.*, 13(3):108–112.

[9] Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., and Rauzy, S. (2008). Le CID – Corpus of Interactional Data : Annotation et

exploitation multimodale de parole conversationnelle. *Traitement Automatique des Langues*, 49(3):105–134.

[10] Brysbaert, M., Keuleers, E., and New, B. (2011). Assessing the usefulness of Google Books' word frequencies for psycholinguistic research on word processing. *Front Psych*, 2(27).

[11] Combescure, P. (1981). Vingt listes de dix phrases phonétiquement équilibrées. *Revue d'Acoustique*, 56:34–38.

[12] Content, A., Mousty, P., and Radeau, M. (1990). Brulex. Une base de données lexicales informatisée pour le français écrit et parlé. *L'année psychologique*, pages 1–17.

[13] Cooke, M. and Aubanel, V. (2017). Effects of linear and nonlinear speech rate changes on speech intelligibility in stationary and fluctuating maskers. *J. Acoust. Soc. Am.*, 141(6):4126–4135.

[14] Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., and Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Commun.*, 55:572–585.

[15] Dodelé, L. (2000). L'audiometrie vocale en presence de bruit et le test AVfB. *Cahiers de l'Audition*, 13(6).

[16] Durand, J. and Lyche, C. (2003). Le projet Phonologie du français contemporain (PFC) et sa méthodologie. In Durand, J., Laks, B., and Lyche, C., editors, *Corpus et variation en phonologie du français : méthodes et analyses*, pages 213–278. Hermès, Paris.

[17] Fournier, J. E. (1951). *Audiométrie vocale. Les épreuves d'intelligibilité et leurs application au diagnostic, à l'expertise et à la correction prothétique des surdités*. Maloine, Paris VI.

[18] Giraud, A.-L. and Poeppel, D. (2012). Speech Perception from a Neurophysiological Perspective. In Poeppel, D., Overath, T., Popper, A. N., and Fay, R. R., editors, *The Human Auditory Cortex*, pages 225–260. Springer New York, New York, NY.

[19] Goldman, J.-P. (2011). EasyAlign: an automatic phonetic alignment tool under Praat. In *Interspeech*, pages 3233–3236, Florence, Italy.

[20] Greenberg, S. (1997). On the origins of speech intelligibility in the real world. In *Proc. of ESCA-workshop on Robust Speech Recognition for Unknown Communication Channels*, pages 23–32, Pont-à-Mousson, France.

[21] Jansen, S., Luts, H., Wagener, K. C., Kollmeier, B., Del Rio, M., Dauman, R., James, C., Fraysse, B., Vormès, E., and Frachet, B. (2012). Comparison of three types of French speech-in-noise tests: A multi-center study. *International Journal of Audiology*, 51(3):164–173.

[22] Kressner, A. A., Westermann, A., and Buchholz, J. M. (2018). The impact of reverberation on speech intelligibility in cochlear implant recipients. *J. Acoust. Soc. Am.*, 144(2):1113–1122.

[23] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

[24] Lopez-Poveda, E. A., Eustaquio-Martín, A., Stohl, J. S., Wolford, R. D., Schatzer, R., Gorospe, J. M., Ruiz, S. S. C., Benito, F., and Wilson, B. S. (2017). Intelligibility in speech maskers with a binaural cochlear implant sound coding strategy inspired by the contralateral medial olivocochlear reflex. *Hearing Res.*, 348:134–137.

[25] Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *P. Natl. Acad. Sci. USA*, 103(49):18866–18869.

[26] Luts, H., Boon, E., Wable, J., and Wouters, J. (2008). FIST: A French sentence test for speech intelligibility in noise. *International Journal of Audiology*, 47(6):373–374.

[27] Ma, J., Hu, Y., and Loizou, P. C. (2009). Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *J. Acoust. Soc. Am.*, 125(5):3387–3405.

[28] Marantz, A., Miyashita, Y., and O'Neil, W. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain*, pages 95–126. MIT Press.

[29] Nelson, M. J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S. S., Naccache, L., Hale, J. T., Pallier, C., and Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *PNAS*, 114(18):E3669–E3678.

[30] New, B., Brysbaert, M., Véronis, J., and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Appl. Psycholinguist.*, 28(04):283.

[31] New, B., Pallier, C., Ferrand, L., and Matos, R. (2001). Une base de donnees lexicales du français contemporain sur Internet : Lexique. *L'année psychologique*, 101:447–462.

[32] 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.

[33] Qin, M. K. and Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *J. Acoust. Soc. Am.*, 114(1):446–454.

[34] Quilis, A. (1993). *Tratado de fonología y fonética españolas*. Gredos, Madrid, Spain.

[35] Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., Weistock, M., McGee, V. E., Pachl, U. P., and Voiers, W. D. (1969). IEEE Recommended practice for speech quality measurements. *IEEE Trans. Audio Acoust.*, pages 225–246.

[36] Scott, S. K., Blank, C. C., Rosen, S., and Wise, R. J. S. (2000). Identification of a pathway for intelligible speechin the left temporal lobe. *Brain*, pages 2400–2406.

[37] Stajner, S., Ponzetto, S. P., and Stuckenschmidt, H. (2017). Automatic Assessment of Absolute Sentence Complexity. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4096–4102.

[38] Vaillancourt, V., Laroche, C., Mayer, C., Basque, C., Nali, M., Eriks-Brophy, A., Soli, S. D., and Giguère, C. (2005). Adaptation of the HINT (hearing in noise test) for adult Canadian Francophone populations. *International Journal of Audiology*, 44(6):358–361.

[39] Wioland, F. (1985). *Les structures syllabiques du français*. Slatkine - Champion, Genève, Paris.

[40] Wolff, A. (2014). *The French Language worldwide 2014*. Nathan, Paris, France.

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

Sound class		IPA	Frequency (%)	Mergers
Consonant	Plosive	p	4.26	
		t	6.07	
		k	4.08	
		b	2.42	
		d	2.61	
		ɡ	1.20	
	Fricative	f	2.28	
		s	5.48	
		ʃ	1.69	
		v	2.60	
		z	1.16	
		ʒ	1.71	
	Nasal	m	3.07	
		n	2.42	(n: 2.39, ŋ: 0.03)
	Liquid	l	4.51	
		ʁ	11.07	
	Glide	ɥ	0.49	
		j	2.60	
		w	1.35	
Vowel	High	i	4.97	
		y	1.92	
		u	2.58	
	Mid-High	e	4.02	
		ø	0.58	
		o	1.78	
	Mid	ə	1.20	
	Mid-Low	ɛ	5.55	
		œ	0.63	
		ɔ	2.69	
	Low	a	6.93	
	Nasal	ɛ̃	1.21	(ɛ̃: 1.17, œ̃: 0.04)
		ɔ̃	1.69	
		ɑ̃	3.17	

Table 1: Frequency distribution of phonemes-in-keywords ($N = 14,188$) in the Fharvard corpus in International Phonetic Association (IPA) coding. Phonemes' sound class is given in the lefmost two columns. Rightmost column shows the least frequent phonemes, merged with their more frequent allophone or most similar phoneme category.

	F		M	
	mean	SD	mean	SD
SRT (dB)	−4.65	0.74	−3.97	0.64
slope (%/dB)	11.37	1.72	13.71	2.21

Table 2: Mean and standard deviation of SRT and slope for 50 % correct keywords over participants (N=23), for the female (F) and male (M) talker.

N sentence per list	5		10		15		20	
	mean	SD	mean	SD	mean	SD	mean	SD
list	.162	.024	.077	.023	.050	.015	.036	.015
participant	.091	.001	.088	.001	.087	.001	.086	.001

Table 3: Mean and standard deviation of the variance for random effects *list* and *participant*.

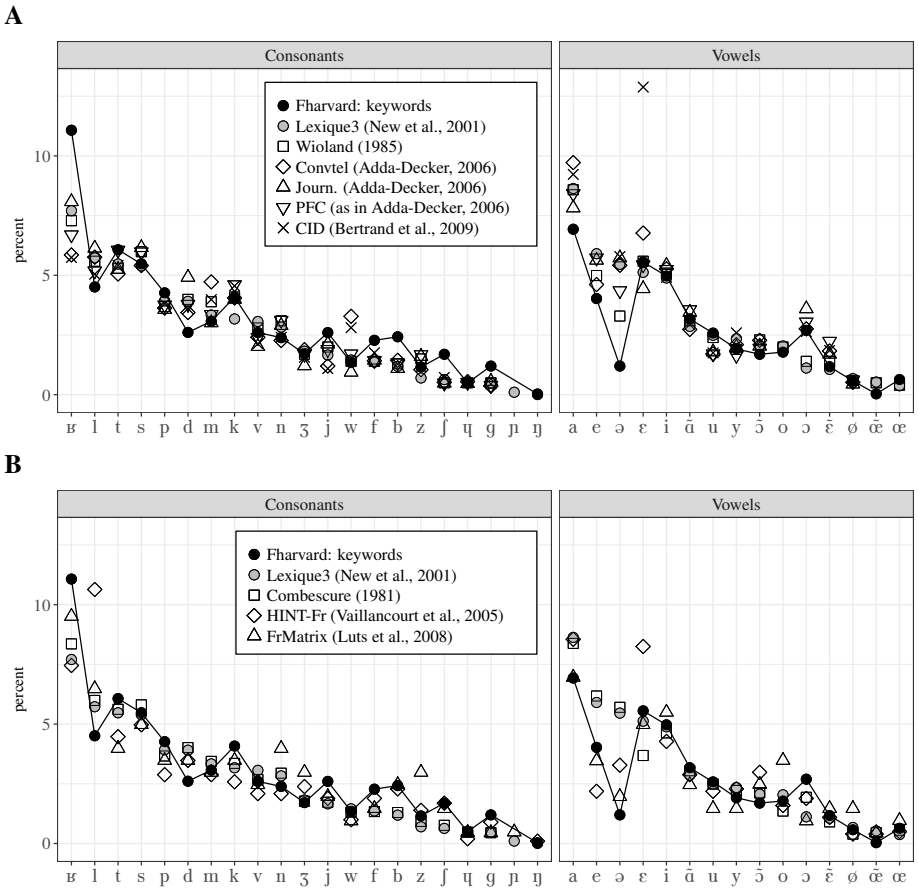

Figure 1: Phoneme frequency distribution in Fharvard keywords compared to that of corpora (**A**) and testing materials (**B**), ordered in decreasing values for the reference distribution Lexique3 and split across consonants (left panel) and vowels (right panel). Numerical values for the Fharvard corpus are found in Table 1. *Vowel mergers*: occurrences of /e/ and /ø/ are merged with /ɛ/ and /ə/ respectively in CID. Occurrences of /o/, /œ/ and /œ̃/, are merged with /ɔ/, /ə/ and /ɛ̃/ respectively in CID, Convtel, Journ. and PFC. *Consonant mergers*: No data is available for /ɥ/ in CID; for /ŋ/ in all corpora except in Fharvard keywords, Lexique3 and HINT-Fr; and for /ɲ/ except in Lexique3 and FrMat.

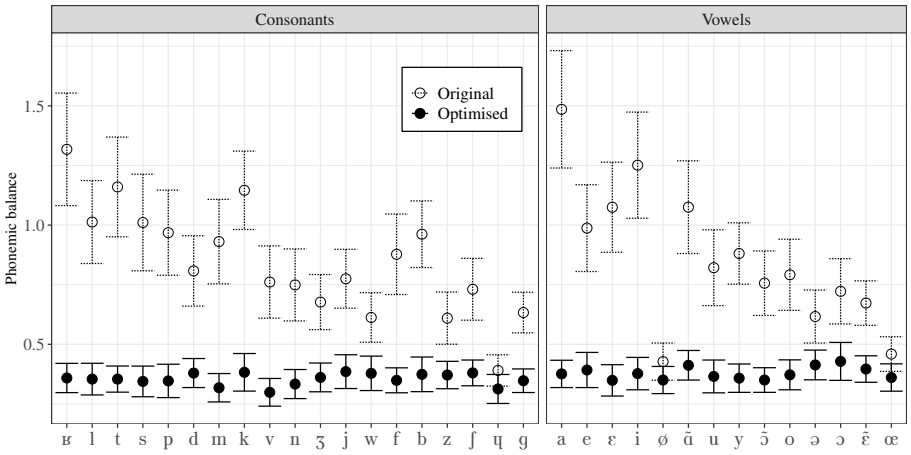

Figure 2: phonemic balance of the Fharvard corpus pre- (dashed lines) and post- (solid lines) optimisation.

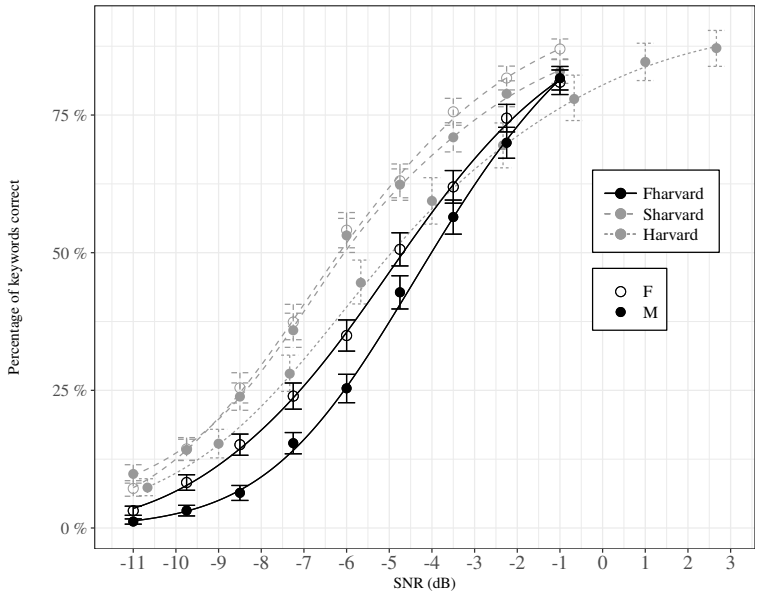

Figure 3: Speech reception thresholds with overlaid psychometric curves for the Female and Male talkers of the Fharvard corpus (solid black lines). Data for the Sharvard and Harvard corpus [4, Fig. 3] were added for comparison.