Speech intelligibility of English, Polish, Arabic and Mandarin under different room acoustic conditions

Laurent Galbrun[*] and Kivanc Kitapci

School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Edinburgh, EH14 4AS, UK

ABSTRACT

This paper examines the impact of room acoustic conditions on the speech intelligibility of four languages (English, Polish, Arabic and Mandarin). Listening test scores (diagnostic rhyme tests, phonemically balanced word tests and phonemically balanced sentence tests) of the four languages were compared under four room acoustic conditions defined by their speech transmission index (STI = 0.2, 0.4, 0.6 and 0.8). The results obtained indicated that there was a statistically significant difference between the word intelligibility scores of languages under all room acoustic conditions, apart from the STI = 0.8 condition. English was the most intelligible language under all conditions, and differences with other languages were larger when conditions were poor (maximum difference of 29% at STI = 0.2, 33% at STI = 0.4 and 14% at STI = 0.6). Results also showed that Arabic and Polish were particularly sensitive to background noise, and that Mandarin was significantly more intelligible than those languages at STI = 0.4. Consonant-to-vowel ratios and languages' distinctive features and acoustical properties explained some of the scores obtained. Sentence intelligibility scores confirmed variations between languages, but these variations were statistically significant only at the STI = 0.4 condition (sentence tests being less sensitive to very good and very poor room acoustic conditions). Overall, the results indicate that large variations between the speech intelligibility of different languages can occur, especially for spaces that are expected to be challenging in terms of room acoustic conditions. Recommendations solely based on room acoustic parameters (e.g. STI) might then prove to be insufficient for designing a multilingual environment.

1. Introduction

In a modern and globalised world, the interaction between multilingual and multicultural people in public, commercial and social spaces is gaining importance, and oral communication is at the centre of this interaction. In the literature, only few studies have been comparing differences between physical measures and subjective measures of speech intelligibility for native speakers of varying languages [1-5], and most of these focused on comparisons between English and Chinese (i.e. Mandarin) [2-5]. Additionally, design guidelines used for speech intelligibility always focus on physical parameters only (e.g. speech transmission index (STI), reverberation time, signal-to-noise ratio (S/N)), disregarding the possibility of having interactions between room acoustics parameters and languages. Investigating the relations between commonly used objective speech intelligibility measures and subjective intelligibility scores of different languages may clarify how each language performs in a given acoustics condition, and help designing the acoustic environment appropriately for a specific language, or a combination of languages.

Houtgast and Steeneken [1] investigated the speech intelligibility of various languages by examining differences between rank orders obtained across the languages, for different room acoustic conditions. The research examined 11 western languages (English, Finnish, French, German, Hungarian, Italian, Dutch, Maori, Polish, Swedish and Slovak) under 16 acoustic conditions which were varied in terms of reverberation time and signal-to-noise ratio. The main purpose of this study was to validate the rapid speech transmission index (RASTI), which is a simplified version of the STI, by comparing this physical measure of speech intelligibility with the articulation index (AI) obtained from listening tests. Differences between the test materials used for each language did not make it possible to compare word intelligibility percentages obtained from the different languages. However, correlations between rank orders were carried out, and these highlighted differences in speech intelligibility between the languages. It was suggested that these may be caused by several effects, including talker specific effects, phoneme or language specific effects, as well as absence of (or subtle differences among) the carrier phrases, and level mismatch between the tests [1]. The research presented here focuses on language specific effects.

[*] Corresponding author. Tel.: +44 131 4513145
E-mail address: L.G.U.Galbrun@hw.ac.uk

Another highly relevant study was conducted by Kang [2], who compared the intelligibility of English and Mandarin in two spaces (a seminar room and a corridor), under different room acoustic conditions. It was found that for a relatively high STI, the word intelligibility of Mandarin was better than English (around +5% at STI = 0.6), and for a low STI, the intelligibility of English was better (around +10% at STI = 0.2). It is interesting to note that these significant differences were observed in the corridor, but not in the seminar room (almost no differences for STIs below 0.5 and only around +2% for Mandarin at STI = 0.6 and above). Converted sentence intelligibility showed even more marked differences both in the corridor and in the seminar room, especially at low STI values. This led the author to state that Mandarin is slightly better than English under reverberant conditions, and English is considerably better than Mandarin under noisy conditions. Kang suggested that the greater dynamic range of English might explain its better scores at low STI values, while the tonality of Mandarin might have been helpful at high STI values. Peng [3] also compared the word intelligibility of Mandarin and English as a function of the STI, and found English to be more intelligible than Mandarin across most STI conditions (+2-4%), with the exception of STIs of approximately 0.3 and below, where Mandarin was marginally more intelligible. More recently, Zhu et al. [4] found that the word intelligibility of English is slightly better than that of Mandarin up to an STI of 0.7 (typically around +2-3%, with a maximum difference of +4.5% at STI = 0.4), after which the scores are very similar. Overall, the studies [2-4] indicate that English tends to be slightly more intelligible than Mandarin under most room acoustic conditions, although some contradictions are observed between the findings of these studies, especially for either very poor or very good room acoustic conditions. These contradictions have been mainly attributed to the use of different test materials [4].

Ji et al. [5] investigated the correlation between objective measures of speech intelligibility and subjective intelligibility scores of Chinese, Japanese and English. The research found that the objective measures providing the best correlations varied depending on the language considered, suggesting that a single objective measure cannot accurately predict the intelligibility of different languages. Unlike the work presented here, the research focused on correlations and did not examine variations between the subjective scores of the three languages examined.

A number of other researchers also examined native and non-native speech intelligibility [6-9], main findings being that non-native speakers tend to perform lower under any type of masking condition [6, 8] and that the linguistic content of background noise can also affect speech intelligibility [7, 9].

Overall, the review of previous work shows that the number of studies that investigated the relationship between languages and speech intelligibility is quite limited, most comparisons having been made between English and Mandarin. Although it is known that there can be speech intelligibility variations between languages, little is known about the extent of these variations and their statistical significance. The present study aims to develop this knowledge by comparing the speech intelligibility of four languages representative of a wide range of linguistic properties (English, Mandarin, Polish, and Arabic) under various room acoustic conditions. The comparisons have been based on a physical measure of intelligibility (STI) and word/sentence intelligibility scores. More specifically, these four languages have been tested under four room acoustic conditions (varying in terms of reverberation time and signal-to-noise ratio), and diagnostic rhyme tests (DRT), phonemically balanced word tests (PB word), and phonemically balanced sentence tests (PB sentence) have been used to determine speech intelligibility scores. It is important to point out that both word and sentence tests have some limitations with regard to comparisons between languages. For example, Kang [2] pointed out that English PB words, especially monosyllabic ones, represent the English words with relatively few phonemes and letters, unlike Mandarin PB words that represent all type of words in Mandarin. In that sense, the use of sentences provides a more direct way to compare the speech intelligibility of different languages, but sentence scores tend to be high under good acoustic conditions and not very sensitive to small changes in listening conditions [10], i.e. less sensitive to identifying variations across languages. For these reasons, both word and sentence tests have been used in the research; their respective limitations should however be kept in mind when analysing results.

The paper first presents the methodology used in the study, followed by the illustration and analysis of results, a discussion, and conclusions.

2. Methodology

This section describes the selection of languages, the word and sentence lists, the recording procedure, the post-processing, and the listening tests used in the research. All the intelligibility tests were carried out under four different room acoustic conditions that were defined in terms of different speech transmission index values (STI = 0.2, 0.4, 0.6 and 0.8). According to the STI qualification ratings of ISO 9921 [11], these corresponded respectively to “bad”, “poor”, “good” and “excellent” speech intelligibility conditions (Bad: STI 0-0.3; Poor: STI 0.3-0.45; Fair: STI 0.45-0.6; Good: STI 0.6-0.75; Excellent: STI 0.75 – 1.0).

2.1 Selecting the languages

Languages representative of a wide range of linguistic properties were selected from different language families such as the Indo-European (e.g. English, German, Polish, Spanish, and Farsi), Uralic (e.g. Turkish), Afro-Asiatic (e.g. Arabic), and Sino-Tibetan (e.g. Mandarin) language families. Five criteria were applied for identifying the languages to be tested:

(1) The selected languages had be representative of real multilingual environments, such as those often found in large western cities.
(2) A significant variability between the consonant-to-vowel ratios of the languages was aimed for, as the speech intelligibility is affected by the loss of consonants [12], and as such variability would allow examining whether languages with a high consonant-to-vowel ratio are more sensitive to poor room acoustic conditions. Consonant-to-vowel ratios of languages are calculated from consonant and vowel inventories which are elements of phonology of a language [13]. Inventories are not limited to the letters specified as consonants and vowels in an alphabet, as a combination of several letters might produce a single consonantal or vowel speech sound, such as ‘th’ or ‘ch’ in English. The total numbers of such sounds create the consonant and vowel inventories. Depending on the language, the number of consonants in a consonant inventory varies between 6 and 122, and the number of vowels in a vowel inventory varies between 2 and 14 [13]. Consonant-to-vowel ratios are calculated by dividing the number of consonants by the number of vowels in an inventory, resulting in a number between 1 and 29. The results are divided into 5 categories, which have been used when selecting the languages of the research presented: low (smaller than or equal to 2), moderately low (between 2 and 2.75), average (between 2.75 and 4.5), moderately high (between 4.5 and 6.5), and high (larger than or equal to 6.5) consonant-to-vowel ratio [13].
(3) Tonality was identified as a linguistic factor that can clearly differentiate languages [14, 15], which is why at least one tonal language had to be selected. Tone is the change of the meaning of a word by the change of pitch, and in that respect languages can be subdivided into three categories: no tones, simple tonal system, and complex tonal system [14]. Languages with a simple tonal system utilise only two-way contrast in terms of tones (i.e. high pitch - low pitch), but languages with a complex tonal system, such as Mandarin, can also use an ascending or descending pitch. 307 out of 527 languages utilise no tones, whilst 132 have a simple tonal system and 88 have a complex tonal system [14].
(4) The native speakers’ population of each language also had to be taken into account, as the research aimed to be representative of a wide range of people.
(5) The availability of native speakers had to be considered, as the languages selected had to comply with high numbers of participants that could be found at Heriot-Watt University.

Based on those five criteria of real environment depiction, consonant-to-vowel ratio, tonality, native speakers’ population, and availability of subjects, four languages were selected. These were English (low consonant-to-vowel ratio [13], wide-spread usage around the world), Mandarin (complex tonal system [16], average consonant-to-vowel ratio [13], high native speakers’ population), Arabic (moderately high consonant-to-vowel ratio [13], high native speakers’ population), and Polish (high consonant-to-vowel ratio [13] and availability of speakers).

2.2 Word and sentence lists

To assess the speech intelligibility of each language, diagnostic rhyme tests (DRT), phonemically balanced (PB) word lists and phonemically balanced sentence lists were used. DRT and PB word tests were employed to examine word intelligibility, whilst PB sentence tests were used for the analysis of sentence intelligibility. Phonemically balanced word and sentence lists represent a language by having approximately the same phonemes of that language [17], where a phoneme is any one of the set of smallest units of speech in a language that distinguish one word from another (e.g. in English, the /s/ in sip and the /z/ in zip represent two different phonemes) [18]. More specifically, phonemically balanced words or sentences match approximately the frequency of phonemes as they appear on average in ordinary conversations in that language [17]. It should be noted that in order to represent a specific language, all of the word lists must be phonemically balanced. Therefore, the DRT is a phonemically balanced test as well. The difference between the DRT and PB word tests is that the former focuses on discrimination of consonants, and the latter focuses on the intelligibility of the whole word [19]. Furthermore, the DRT test is based on the assumption that the sounds of languages can be identified by using a set of distinctive features, which does not exceed twelve distinctive features [20]. These distinctive features are representative of the phonological properties of speech (how speech sounds are used in a given language), rather than the phonetic properties of speech (how speech sounds are physically produced) [18]. Therefore, the present study focuses on the phonological properties of languages rather than their phonetic properties.

The DRT consists of a list of words arranged in pairs, e.g. 192 words arranged in 96 pairs for English [21]. The words are common, monosyllabic words, and most of them have three sounds ordered in a consonant-vowel-consonant (CVC) sequence. The word pairs differ only in their initial consonants, so that discrimination of consonants of a given language can be analysed. DRT lists were used for English [21], Arabic [22] and Mandarin [23], the same list being used for each acoustic condition. In order to minimise prior learning effects, the words heard from the DRT pairs were randomised across all STI conditions, as well as the talkers pronouncing the words (i.e., the sequence of words and talkers pronouncing them were different for each of the acoustic conditions tested). PB word tests were used for Polish [24], because of the lack of DRT material in Polish. The Polish PB word lists consisted of 4 sets of 48 monosyllabic CVC words, with one set used for each acoustic condition (i.e., no prior learning effects possible), and no carrier sentences used in the tests. Different word tests can easily be responsible for significant variations within a single language (e.g. nonsensical vs. meaningful words) [11], which is why the use of comparable test materials is crucial when comparing intelligibility across languages. The lack of DRT material in Polish is in that sense a limitation of the current study, but comparisons between DRT and PB English words data (the former being taken from the current study and the latter from ref. [25]) indicate that the variability between DRT and PB scores tends to be fairly small (Figure 1), suggesting that comparisons between DRT and PB results are acceptable (as also pointed out in ANSI/ASA S3.2-2009 [19]). Figure 1 shows that DRT and PB scores of English have an average difference (calculated from absolute values) of 2.4% across the four STI conditions considered, with a maximum difference of 5.5% observed at STI = 0.4. This is well below the large differences observed between languages that are presented in section 3 (which are as high as 33% at STI = 0.4), indicating that these inaccuracies are not expected to have affected the main findings obtained when comparing Polish PB word scores to DRT scores of the other languages. It is however accepted that some inaccuracies should be expected and are unfortunately not quantifiable for Polish, and that the variations between DRT and PB word scores of Polish could be higher than those presented for English in Figure 1. The data taken from ref. [25] was based on the standard Harvard PB word test, which is commonly used in the United States. It should also be noted that comparability of DRT and PB scores can be achieved only by removing the effect of guesswork in the calculation of DRT scores (see section 2.3), as rhyme tests are closed tests that are otherwise expected to provide higher scores [3].

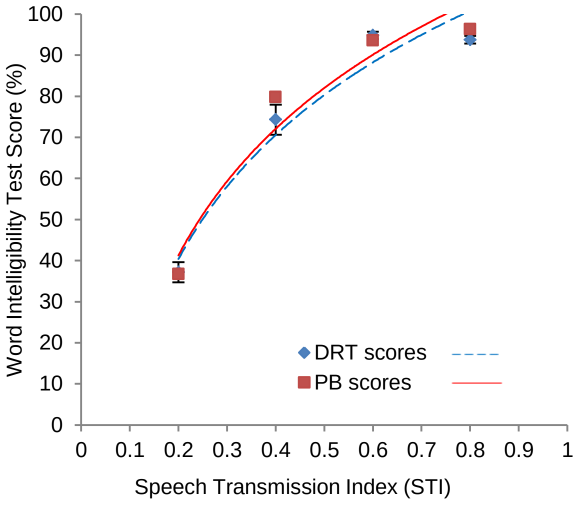


Fig. 1. DRT (present study) and PB (ref. [24]) word intelligibility scores of English obtained under different STI conditions (data markers, standard errors of the means (for DRT data only) and logarithmic regression lines).

Table 1. Distinctive features tested in the Diagnostic Rhyme Tests of English [21], Arabic [22] and Mandarin [23].

English	**Arabic**	**Mandarin**
Voicing (voiced / unvoiced)	Tenseness (tense / lax)	Airflow (airflow / no airflow)
Nasality (nasal / oral)	Nasality (nasal / oral)	Nasality (nasal / oral)
Sustention (continuant / interrupted)	Mellowness (strident / mellow)	Sustention (continuant / interrupted)
Sibilation (sibilant / non-sibilant)	Flatness (flat / plain)	Sibilation (sibilant / non-sibilant)
Graveness (grave / acute)	Graveness (grave / acute)	Graveness (grave / acute)
Compactness (compact / diffuse)	Compactness (compact / diffuse)	Compactness (compact / diffuse)

Furthermore, it is important to clarify the reasons for having chosen the Diagnostic Rhyme Test in the current work, and to illustrate its principles in some detail. The decision of using the DRT followed guidance given in the standard ANSI/ASA S3.2-2009 [19]: the DRT specifically allows examining distinctive features of speech through the discrimination of phonemes, and comparing those features across languages (unlike other tests). Furthermore, this is one of the few tests for which materials have been developed for several languages using consistent procedures (pairs of words based on a Consonant-Vowel-Consonant (CVC) sequence, with words varying in the first consonant only) [19] and the test is known to give stable intelligibility scores [21]. The DRT results reveal errors in the initial consonant only and the test does not need a carrier sentence, i.e. this is a very simple test that is not context dependent. More specifically, the DRT analyses a language by dividing sentences into sound level units (phonemes), rather than dividing them into grammatical components such as words or syllables [20]. For instance, the comparison between the words ‘bill’ and ‘pill’ focuses on isolating the phonemes /b/ and /p/ in English and can be used to discriminate the “voiced” sound in which vocal cords vibrate, /b/, against the “unvoiced” sound that does not require vocal cords vibration, /p/. In this example, the discrimination between these two phonemes identifies the intelligibility of the distinctive feature called “voicing”. The method is universal and can be applied to any language, and the concept of distinctive features was indeed developed through the analysis of multiple languages [20]. DRT lists are language specific, as the six distinctive features of DRT lists are selected to be representative of the specific language considered, and the words included in the lists are phonemically balanced.

DRT lists are made of 6 distinctive features that do not need to be identical across languages, as some distinctive features might be relevant in one language but irrelevant in another, and this is why different distinctive features might need to be considered to correctly represent a language. All of these aspects are carefully taken into account when developing DRT lists, as described in the references of the DRT lists used in the present study [21-23]. Table 1 illustrates the distinctive features tested in the present research for English, Arabic and Mandarin. These DRT lists include 16 pairs of words per feature for English [21] and Mandarin [23] (16 × 6 = 96 pairs of words) and 12 pairs of words per feature for Arabic [22] (12 × 6 = 72 pairs of words).

Sentence intelligibility was tested using PB sentence lists. These consisted of a total of 10 sentences per language, and from these, 2 sentences were randomly selected for the STI = 0.8 and STI = 0.6 conditions, and 3 sentences were randomly selected for STI = 0.4 and STI = 0.2 (no prior learning effects possible, as each sentence was tested only once). The English PB sentence pool [26] consisted of 6 high predictability and 4 low predictability sentences (last word out of context), with 6-7 words per sentence. Sentences’ pools used for Polish (5 words per sentence) [27], Arabic (3-6 words) [28], and Mandarin (7 words) [29], consisted of sentences that represent an everyday conversation. It is important to note that a variety of factors influence sentence intelligibility, such as context, familiarity, predictability, prosody and number of words [19], and many of these were not clearly defined in the materials used. Therefore, sentence intelligibility comparisons between languages were expected to be less reliable than word tests, which were well defined and comparable.

2.3 Calculation of intelligibility scores

DRT scores were calculated using Voiers’ equation [30], which eliminates the effect of guesswork:

$$P_c = 100\frac{N_r - N_w}{T} \qquad (1)$$

where N_r is the number of correct responses, N_w is the number of incorrect responses, T is the total number of test items, and P_c is the percentage correct score. Phonemically balanced words were counted as correct only when the word's spelling was exact, and the results obtained from all the words tested were converted into percentages of correct scores. As mentioned earlier, the use of equation (1) allows comparing DRT and PB word scores by removing guesswork.

PB sentence scores were calculated by counting the number of correct words in a given sentence, and this was then converted into percentages of correct scores. In all cases, the arithmetic average of all participants' results was computed for each of the STI conditions examined.

2.4 Recording and post-processing

The word lists were recorded in the anechoic chamber of Heriot-Watt University using native speakers for each language (3 males and 3 females). In the standard ANSI/ASA S3.2 [19], the minimum number of speakers is stated as 5; in order to achieve equal gender representation, 6 speakers were used in the current study. Because of the significant variety of accents within languages, attention was given to the origin of the speakers. The English speakers had to speak English with Received Pronunciation (RP) [31], which is normally associated with formal speech and tends to be spoken in the south of England. The Arabic speakers were selected from Syria, although the origin of Arabic speakers was not crucial, as the Arabic material was written and recorded in modern standard Arabic (al-fuṣḥá) [22, 28], for which the pronunciation tends to be independent from accents and dialects. Care was also taken in the selection of Polish and Mandarin speakers, so that they could produce formal speech material. Before the actual recordings, a practice list was read by each speaker, to make them familiar with the process, and to train them in producing normal vocal effort and normal rate of talking. All the lists were also read by the speakers prior to the actual recordings. The speaking rates and average sentences' durations were comparable across languages. These were, respectively, 0.34 s and 2.22 s for English, 0.32 s and 1.62 s for Polish, 0.39 s and 1.77 s for Arabic, and 0.25 s and 1.81 s for Mandarin.

The word and sentence recordings were then calibrated in terms of sound pressure level, by using a custom made head and torso model with microphones (Brüel & Kjaer 4176 (Naerum, Denmark)) placed inside its ears and connected to a sound level meter (Brüel & Kjaer 2231). The material to be calibrated was played through Beyerdynamic DT 150 (Berlin, Germany) closed headphones placed over the head of the model. Audio files were then prepared for the listening tests, including randomisation in the sequencing of words and editing of gaps between words. For the DRT tests (English, Arabic, and Mandarin), the word selected between a pair was simply ticked on a list provided, and the word frequency was set to one word per 1.4 seconds, following guidance by Cohen [32]. For the Polish PB word tests, the gap between words was set to 5 seconds, to give a convenient amount of time for writing down the whole word. Although there is no standard for the frequency of words in PB word tests, Diaz et al. [33] suggested the frequency of one word per 4 seconds for Spanish PB word tests. This was adapted to 5 seconds for Polish, based on trial and error. For sentence tests, each new sentence was played after the listener had finished writing down the sentence just heard (no predefined frequency/duration).

2.5 Listening tests

The listening tests were conducted in one of the chambers of the acoustic laboratory of Heriot-Watt University. The dimensions of the chamber were 6.8 m (length) × 4.0 m (width) × 3.0 m (height). All the surfaces were made of reflective materials (brick walls, concrete floor and ceiling), and the room had no windows.

The minimum number of listeners stated in the standard ANSI/ASA S3.2 [18] is 5, but 3 male and 3 female listeners were selected from native speakers of each language, in order to achieve equal gender representation. The listeners of each language were selected from the same regions/countries of the speakers (see section 2.4), and their age distribution was as follows: English participants ranged from 23 to 42 yr (mean 32.3 yr and standard deviation 6.7 yr), Polish from 24 to 33 yr (mean 29.3 yr and standard deviation 3.1 yr), Arabic from 30 to 33 yr (mean 31.7 yr and standard deviation 1.4 yr) and Chinese from 21 to 32 yr (mean 26.2 yr and standard deviation 5.2 yr). The hearing threshold level of the participants was tested using the simple *AudioCheck* online hearing test [34], results showing that all the participants had normal hearing. Hearing tests were carried out in the anechoic chamber of Heriot-Watt University using Beyerdynamic DT 150 closed headphones.

It can also be noted that all the listeners used as participants had one native language only, and most of the Polish, Arabic and Chinese participants were students who had been living in their native country until recently. However, these participants also knew English, and this might have affected their intelligibility scores at the lower STI levels tested (STI = 0.2 and STI = 0.4). In fact, decreases in the intelligibility of a first language can occur under noisy conditions: Tabri et al. [35] showed that monolinguals perform better than bilinguals since birth in noisy conditions only, and Weiss and Dempsey [36] also showed that bilinguals with greater experience of their second language were poorer at perceiving their first language in noise. Therefore, the use of both monolingual and multilingual participants is a limitation of the current study. This is discussed further in section 3.1.

The recorded material was presented through a loudspeaker (KEF Coda III (Maidstone, UK)) placed at 1 m from one of the 4 m wide walls, and positioned over a small table with a propagating height of 1.2 m (mid-way between the woofer and tweeter). Listeners were seated at a distance of 2 m from the loudspeaker, and the speech level was adjusted to 65 dBA, 1 m on axis from the loudspeaker and 1.2 m above floor level. The level was calibrated using uninterrupted speech material (gaps removed between words) and the sound level meter Brüel & Kjaer 2250.

For DRT tests, listeners had to identify the spoken words within the pairs of words provided on a list (by ticking), whilst for PB words and PB sentences, these had to be written down. Each listening test was repeated for four different acoustic conditions (STI = 0.2, 0.4, 0.6 and 0.8), by changing the reverberation time and signal-to-noise ratio. The order of the acoustic conditions tested was always highest (STI = 0.8) to lowest (STI = 0.2), in order to minimise auditory fatigue (exposure to high levels of white noise at STI = 0.4 and in particular STI = 0.2, having been found to be aurally tiring). Testing from STI = 0.2 to any higher STI condition would have required reasonable breaks to compensate from auditory fatigue: this was excluded in order to reduce the testing time. Furthermore, the fixed order also reduced the time needed for setting up the room across the conditions tested (absorption panels used at STI = 0.8 and 0.6 and removed for STI = 0.4 and 0.2). Although the procedure was consistent and therefore guaranteed comparable results, the drawback of this fixed order is that it might have included an order effect that could have been excluded by randomising the STI conditions tested. This fixed order was however not expected to be responsible for learning effects for three reasons: 1) Word familiarity can be neglected in DRT tests [20]; 2) The sequence of DRT words and talkers was randomised (as explained in section 2.2); 3) The number of DRT words heard in each condition was quite large (96 for English and Mandarin and 72 for Arabic), so that words were unlikely to be easily learnt.

The reverberation time was controlled by adding or removing foam and glass-wool panel absorbers on the walls. The use of different absorbers was due to not having enough identical panel absorbers for achieving the STI = 0.8 condition (details about the absorbers used are given in Table 2). The panel absorbers were distributed evenly across the room and were used only at the STI = 0.8 and STI = 0.6 conditions. The signal-to-noise ratio was controlled by adding artificial noise to the speech signal, using the white noise generator Brüel & Kjaer 1405 (S/N = +5 dB for STI = 0.4, and S/N = -5 dB for STI = 0.2). No artificial noise was used at the STI = 0.6 and STI = 0.8 conditions. The STI conditions could then be described as follows; STI = 0.8: no artificial noise and low reverberation time; STI = 0.6: no artificial noise and medium reverberation time; STI = 0.4: S/N = +5 dB and high reverberation time; STI = 0.2: S/N = -5 dB and high reverberation time. In practice, the number of absorption panels and amount of white noise were adjusted to achieve the exact STI values of 0.2, 0.4, 0.6 and 0.8. Details of the reverberation time and direct-to-reverberant ratio (DRR) present during the tests are given in Table 2.

Table 2. Reverberation time (*T*) and direct-to-reverberant ratio (DRR) at the listener's position for all the STI conditions tested. No absorption panels were used at the STI = 0.2 and STI = 0.4 conditions. 12 foam panels (1.2 m × 0.6 m × 0.05 m) were used at the STI = 0.6 condition, and an additional 16 glass-wool panels (8 of dimensions 1.2 m × 0.6 m × 0.05 m, and 8 of dimensions 1.2 m × 0.6 m × 0.1 m) were used at the STI = 0.8 condition (28 panels in total).

	No absorption panels (STI = 0.2 and 0.4)		With 12 absorption panels (STI = 0.6)		With 28 absorption panels (STI = 0.8)	
Frequency (Hz)	*T* (s)	DRR (dB)	*T* (s)	DRR (dB)	*T* (s)	DRR (dB)
125	3.52	-15.7	2.18	-13.5	1.41	-11.5
250	3.36	-13.7	1.15	-8.7	0.73	-6.5
500	3.25	-11.3	1.06	-6.1	0.62	-3.4
1000	3.37	-9.0	1.05	-3.6	0.62	-0.9
2000	2.80	-4.1	0.90	1.2	0.58	3.5
4000	1.94	-2.4	0.75	2.1	0.48	4.5
8000	1.21	2.6	0.59	6.2	0.41	8.3

The physical evaluation of speech intelligibility was made using the speech transmission index (STI), which was measured using the commercial Maximum Length Sequence System Analyzer (MLSSA) software (DRA Laboratories, Sarasota, USA). MLSSA's measurement of the STI is language independent. The computer used to run MLSSA was connected via its sound card to the loudspeaker KEF Coda III and to a half inch microphone Brüel & Kjaer 4190, which was in turn connected to a microphone power supply Brüel & Kjaer 2804. MLSSA measurements showed a maximum change in the STI of around ±0.001 (on a 0-1 scale) when measurements were repeated several times, demonstrating the reliability of the STI measurements.

The data gathered from MLSSA calculations were compared to the word/sentence speech intelligibility scores, and the results obtained are given in the next section.

3. Results

In this section, results of the DRT/PB word tests (overall scores and distinctive features' scores), and PB sentence tests are presented and analysed, followed by the comparison between word and sentence intelligibility scores. All the statistical analysis presented in this paper has been made using Rationalized Arcsine Units (RAU) [37] (i.e., rationalized arcsine transformed data), to ensure that the homogeneity assumption of ANOVA was not violated. Furthermore, the *p*-values given have not been corrected for multiple comparisons. All the statistical analysis has been carried out using the Statistical Package for Social Sciences (SPSS), and all the results given in figures include standard errors of the mean and logarithmic regressions.

Subjects' consistency across all tests presented in this section (word scores, distinctive features' scores and sentence scores) was analysed using the Intra-Class Correlation Coefficient (ICC). The absolute agreement average measures ICC analysis with the two-way mixed model revealed that the answers of participants agree with each other for English (ICC = 0.973), Mandarin (ICC = 0.948), Arabic (ICC = 0.925), and Polish (ICC = 0.991), where ICC > 0.720 is usually considered as an acceptable value for social sciences [38]. This confirms that the use of only 6 listeners per language was appropriate and that the results presented are reliable.

3.1 Word intelligibility tests - Overall results (DRT and PB word tests)

This section examines word intelligibility scores tested under four room acoustic conditions (STI = 0.2, 0.4, 0.6 and 0.8). The results are presented in Figure 2, where the horizontal axis shows the STI values, and the vertical axis shows the word intelligibility scores for all languages. As stated previously, the word intelligibility scores correspond to DRT results for English, Arabic and Mandarin, and to PB results for Polish.

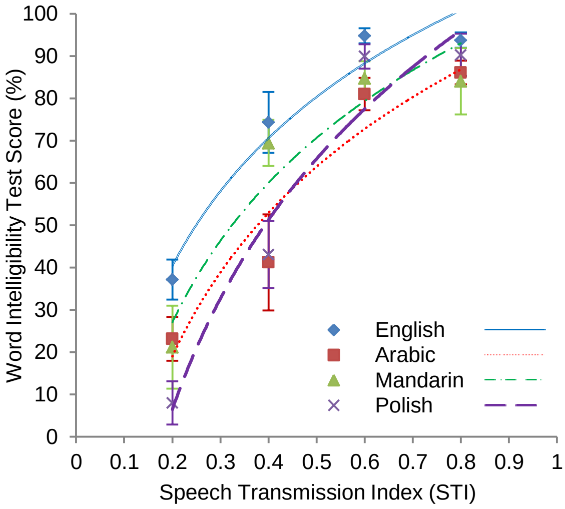


Fig. 2. Word intelligibility scores of English, Arabic, Mandarin, and Polish obtained under different STI conditions (data markers, 95% confidence intervals and logarithmic regression lines).

Figure 2 illustrates that there were differences between speech intelligibility scores of English, Polish, Arabic and Mandarin. English was the most intelligible language under all acoustic conditions, with scores that were at least 4% (STI = 0.8) to 14% (STI = 0.2) higher than those obtained for all the other languages. Results also show that differences between languages were more conspicuous under poor acoustic conditions (STI = 0.4 and STI = 0.2), as there was an approximate maximum difference between language scores of 10% at the STI = 0.8 condition and 14% at STI = 0.6, but this increased to much larger differences of 33% at STI = 0.4 and 29% at STI = 0.2. Of particular interest is the STI = 0.4 condition, which shows the largest difference between scores. At this condition, participants were first introduced to the artificial background noise and the reverberation time was increased: English and Mandarin were significantly more intelligible than Arabic and Polish at STI = 0.4, the word intelligibility scores of English and Mandarin being approximately 30% and 25% higher, respectively, than the word intelligibility scores of Arabic and Polish. This indicates that Arabic and Polish were the languages most sensitive to the introduction of artificial background noise and the increase in reverberation time. For Arabic, the reduction in word intelligibility scores between STI = 0.6 and STI = 0.4 was 40%, and for Polish it was 46%. Although bilingualism might be partly responsible for a decrease in first language intelligibility under noisy conditions (STI = 0.2 and STI = 0.4), this alone cannot justify the low scores observed for Arabic and Polish. The differences found between monolingual and bilingual speakers by Tabri et al. [35] were of around 9% for a signal-to-noise ratio of + 5 dB and of around 7% for a signal-to-noise ratio of 0 dB, well below the 25-30% differences observed here with English and Mandarin. Furthermore, the differences of Tabri et al. [35] represent a worst case scenario, as they were observed for bilinguals since birth, rather than late bilinguals for which differences tend to be lower [36]. One justification of the results obtained can instead be found in the languages' consonant-to-vowel ratios, as a statistically significant negative correlation was found between the consonant-to-vowel ratios of languages and the word intelligibility results for the most challenging room acoustic conditions, Spearman's correlation analysis results being ρ = -0.73 (p < 0.01) for STI = 0.2, and ρ = -0.76 (p < 0.01) for STI = 0.4. The negative sign indicates that word intelligibility decreased with increasing consonant-to-vowel ratio, as expected [13].

Factorial ANOVA showed that there was a main effect (p < 0.01) of language [F(3, 80) = 26.09, p = 0.000] and a main effect (p < 0.01) of STI conditions [F(3, 80) = 339.45, p = 0.000] on word intelligibility, as well as an interaction (p < 0.01) of language and STI conditions [F(9, 80) = 6.55, p = 0.000] on word intelligibility.

One-way ANOVA tests were also carried out for each STI condition, and these clarified that the word intelligibility scores of the four languages examined were significantly different (p < 0.01) at STI = 0.6 [F(3, 20) = 16.35, p = 0.0000], STI = 0.4 [F(3, 20) = 16.38, p = 0.000] and STI = 0.2 [F(3, 20) = 11.45, p = 0.000], whilst differences were not significant (p > 0.05) at STI = 0.8 [F(3, 20) = 2.99, p = 0.055]. In other words, word intelligibility of different languages is comparable under excellent room acoustic conditions, but is not comparable under all other conditions. PB Polish word scores were then removed from the statistical analysis, to check whether differences in test methods affected findings. Statistically significant differences (p < 0.05) between DRT scores were then found at all conditions: at STI = 0.8 [F(2, 15) = 4.67, p = 0.027], STI = 0.6 [F(2, 15) = 23.10, p = 0.000], STI = 0.4 [F(2, 15) = 16.67, p = 0.000] and STI = 0.2 [F(2, 15) = 4.75, p = 0.025]. This confirms that the main findings are not affected by the different word test used for Polish.

The results obtained here confirm the higher intelligibility of English compared to Mandarin, similar to the results obtained in previous research [2-4], although previous work occasionally found slightly higher intelligibility of Mandarin at either very good [2] or very poor [3] room acoustic conditions. Such contradictions have been mainly attributed to the use of different test materials, but it should be noted that the use of different spaces can also affect comparisons [4], and not all languages' comparisons made in previous studies [3, 4] correspond to identical spaces (unlike the work presented here and in ref. [2]).

3.2 Word intelligibility tests – Distinctive features' results (DRT tests)

In this section, DRT distinctive features' scores are analysed for English, Arabic, and Mandarin. Distinctive features of Polish could not be examined, as no DRT tests were available in Polish.

DRT lists examine six different distinctive features that vary with the language tested, and these were listed in Table 1 for the languages examined here. For illustration purposes, an example list of distinctive features tested in the English DRT is also given in Table 3.

In order to understand the effects of room acoustic properties on distinctive features and overall intelligibility of languages, DRT scores of each linguistic property have been first compared and analysed within each language (Figures 3, 4, and 5), and then across languages (Figure 6).

Table 3. Example list of the English DRT distinctive features [21].

Voicing		**Nasality**		**Sustention**	
Voiced	Unvoiced	Nasal	Oral	Continuant	Interrupted
Veal	Feel	Meat	Beat	Vee	Bee
Bean	Peen	Need	Deed	Sheet	Cheat
Gin	Chin	Mitt	Bit	Vill	Bill
Sibilation		**Graveness**		**Compactness**	
Sibilant	Non-sibilant	Grave	Acute	Compact	Diffuse
Zee	Thee	Weed	Reed	Yield	Wield
Cheep	Keep	Peak	Teak	Key	Tea
Jilt	Gilt	Bid	Did	Hit	Fit

By looking at Figure 3, it is seen that for the English language, nasal/oral consonants were discriminated easily under all acoustic conditions, in line with previous work [39]. Even when STI = 0.2, the acoustic condition in which there is a high level of artificial background noise (S/N = -5 dB), the nasality score was still as high as 85%. Additionally, sibilation was the second most intelligible distinctive features for English at STI = 0.4, and compactness also showed high scores at most conditions. Based on these results and the results discussed in the previous section, it can therefore be assumed that the low consonant-to-vowel ratio, and the intelligibility of nasal/oral, sibilant/non-sibilant and compact/diffuse consonants, are the main factors increasing the overall intelligibility of English under all room acoustic conditions.

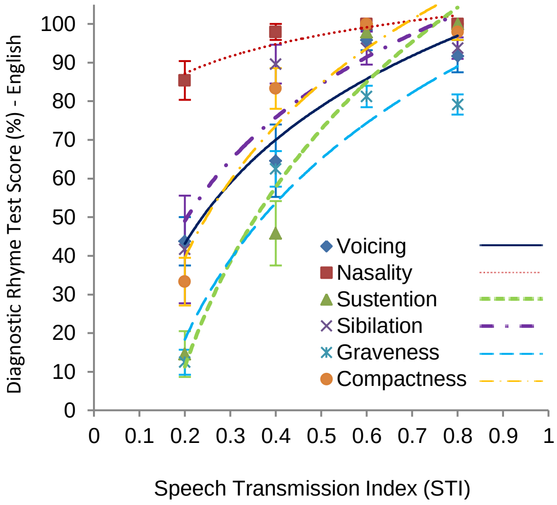


Fig. 3. English distinctive features' scores
(data markers, standard errors of the means and logarithmic regression lines).

The Arabic DRT results of distinctive features (Figure 4) prove that the moderately high consonant-to-vowel ratio of Arabic is not the sole reason for its low speech intelligibility scores. In Figure 4, it is clearly shown that there was a large decrease of intelligibility between STI = 0.6 and STI = 0.4. At STI = 0.6, the DRT scores for all of the distinctive features varied between 67% and 97%, whilst at STI = 0.4, the DRT scores decreased considerably to a range between 22% and 58%. Graveness was the most sensitive distinctive feature, with a decrease as large as 75% between STI = 0.6 and STI = 0.4, followed by compactness which showed a decrease of approximately 45% between these conditions. According to independent sample *t*-tests, the changes in Arabic intelligibility between STI = 0.6 and STI = 0.4 were statistically significant ($p < 0.01$) for graveness [$t(10) = 8.62$, $p = 0.000$], compactness [$t(10) = 3.86$, $p = 0.003$] and mellowness [$t(10) = 4.50$, $p = 0.001$].

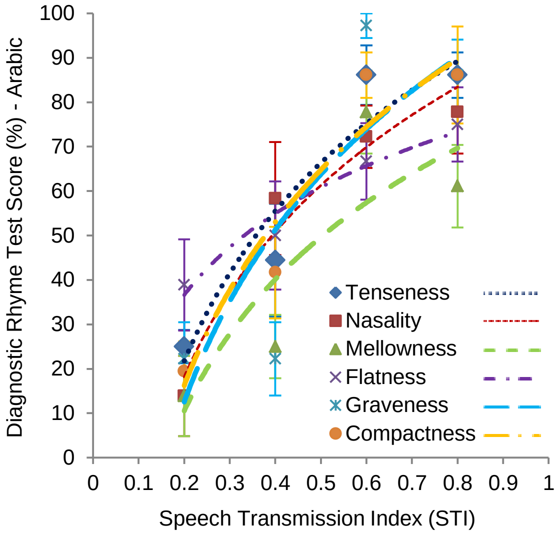


Fig. 4. Arabic distinctive features' scores
(data markers, standard errors of the means and logarithmic regression lines).

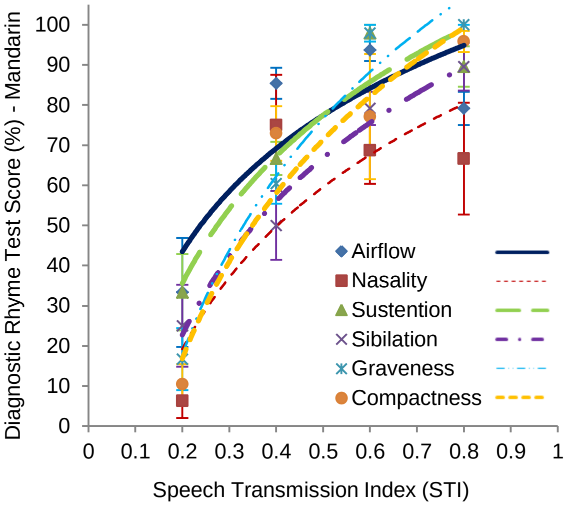


Fig. 5. Mandarin distinctive features' scores
(data markers, standard errors of the means and logarithmic regression lines).

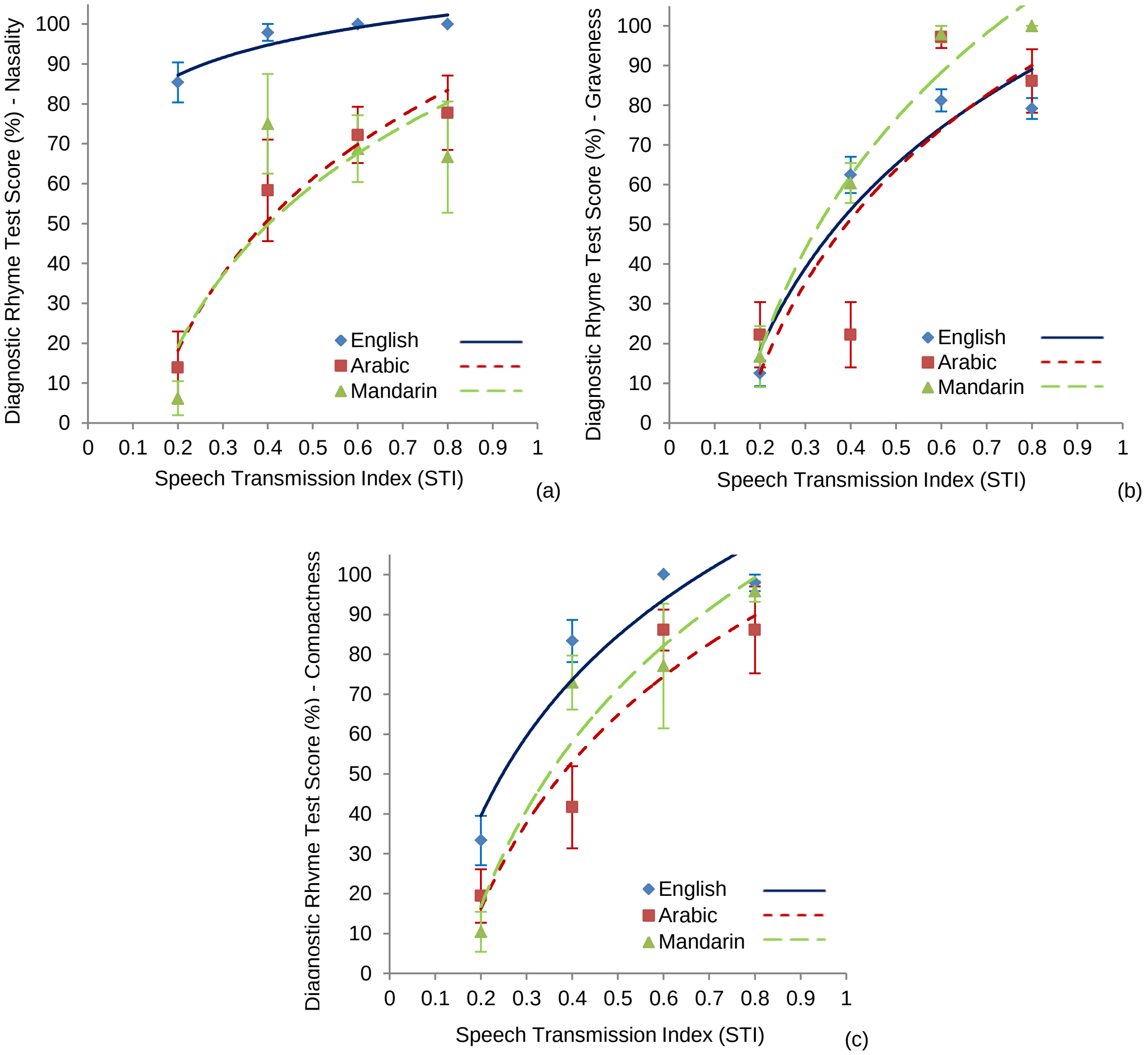


Fig. 6. Comparison of distinctive features' scores between English, Arabic, and Mandarin (data markers, standard errors of the means and logarithmic regression lines). (a) Nasality, (b) Graveness, (c) Compactness.

In section 3.1, it was stated that Mandarin was significantly more intelligible than Arabic and Polish, especially when artificial noise and an increase in reverberation time were introduced (STI = 0.4). The distinctive features' DRT results reveal that for Mandarin, the intelligibility of airflow/no-airflow consonants was high at most conditions (Figure 5). Even under poor conditions such as STI = 0.4, the intelligibility of these consonants was as high as 85%; however, this effect diminished under very high background noise levels (STI = 0.2), to a much lower score of 33%.

Incidentally, it can be noted that Figures 3, 4 and 5 show an unexpected increase in intelligibility for some distinctive features from STI = 0.8 to STI = 0.6 (for voicing and graveness in Figure 3, graveness and mellowness in Figure 4 and airflow, nasality and sustention in Figure 5). However, independent sample *t*-tests showed that none of these changes are statistically significant ($p > 0.05$), so that no conclusions can be drawn from these unexpected variations.

The comparison of distinctive features common to English, Arabic and Mandarin can be seen in Figure 6, which highlights the significantly higher intelligibility of nasality for English (up to 70% higher than Arabic and 80% higher than Mandarin at STI = 0.2), the good intelligibility of graveness for Mandarin, and also the good intelligibility of compactness for English. These results further justify the higher intelligibility scores of English and Mandarin compared to Arabic.

Factorial ANOVA revealed that there was a main effect ($p < 0.01$) of language for nasality [$F(2, 60) = 34.85$, $p = 0.000$], graveness [$F(2, 60) = 6.39$, $p = 0.003$], and compactness [$F(2, 60) = 7.66$, $p = 0.001$], as well as a main effect ($p < 0.01$) of STI conditions for nasality [$F(3, 60) = 22.25$, $p = 0.000$], graveness [$F(3, 60) = 104.25$, $p = 0.000$], and compactness [$F(3, 60) = 47.07$, $p = 0.000$]. Furthermore, an interaction ($p < 0.05$) between languages and STI conditions was found for nasality [$F(6, 60) = 2.40$, $p = 0.038$] and graveness [$F(6, 60) = 6.12$, $p = 0.000$], but not for compactness [$F(6, 60) = 1.59$, $p = 0.166$].

3.3 PB sentence test results

In this section, phonemically balanced sentence test results for English, Polish, Arabic and Mandarin are presented and analysed (Figure 7). By looking at Figure 7, it is seen that Arabic tended to be less intelligible than the other three languages at most conditions. Furthermore, unlike word scores, English was not noticeably more intelligible than all the other languages, which might be partly explained by the low predictability sentences used in English.

The differences between the highest and lowest PB sentence test scores were larger at STI = 0.4 (38%) and STI = 0.6 (11%) compared to STI = 0.2 (6%) and STI = 0.8 (3%). Therefore, it can be stated that PB sentence tests were less sensitive than word tests in identifying differences between languages when the acoustic condition was either very challenging (STI = 0.2), or very good (STI = 0.8). It can also be seen that the variance of intelligibility was the largest at STI = 0.4, confirming what was already observed for word intelligibility. As stated in the analysis of DRT and PB word tests results, Arabic and Polish appeared to have a high sensitivity to the artificial noise and high reverberation time introduced from STI = 0.4, whereas Mandarin and English were less sensitive to these factors. No statistically significant correlations were found between the consonant-to-vowel ratios of languages and the sentence intelligibility results (Spearman test, $p > 0.05$).

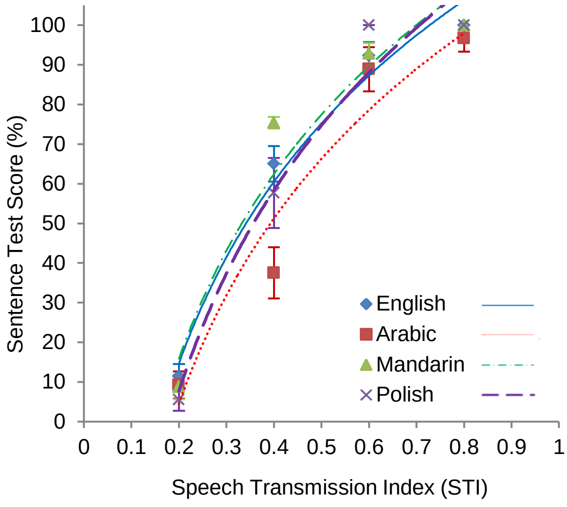


Fig. 7. Sentence intelligibility scores of English, Arabic, Mandarin, and Polish obtained under different STI conditions (data markers, standard errors of the means and logarithmic regression lines).

The smaller variations observed for sentence intelligibility compared to word intelligibility can be explained by the followings: 1) Sentence scores tend to be high under good acoustic conditions, regardless of language [10]; 2) Under very noisy and reverberant conditions the boundaries between syllables can disappear [41] and sentence scores can then become very low across all languages. Smaller variations between languages are therefore to be expected for sentence intelligibility at either very good or very challenging room acoustic conditions (STI = 0.8 and STI = 0.2 respectively), justifying the fact that only the STI = 0.4 and STI = 0.6 conditions show comparable variations between the word and sentence scores. In that respect, the study of Kang [2] represents an anomaly, as it found large differences between sentence scores even under poor and good room acoustic conditions, and it is not clear why this occurred.

Factorial ANOVA showed that there was a main effect ($p < 0.05$) of language [$F(3, 80) = 3.87$, $p = 0.012$] and a main effect ($p < 0.01$) of STI conditions [$F(3, 80) = 361.75$, $p = 0.000$] on sentence intelligibility, as well as an interaction ($p < 0.01$) of language and STI conditions [$F(9, 80) = 2.85$, p = 0.006] on sentence intelligibility. These factorial ANOVA findings are identical to those obtained in the analysis of word intelligibility scores.

However, one-way ANOVA tests carried out for each STI condition, indicated that the sentence intelligibility scores of the four languages examined were significantly different ($p < 0.01$) only at STI = 0.4 [$F(3, 20) = 6.99$, $p = 0.002$], whilst differences were not significant ($p > 0.05$) at STI = 0.8 [$F(3, 20) = 1.00$, $p = 0.413$], STI = 0.6 [$F(3, 20) = 2.07$, $p = 0.137$] and STI = 0.2 [$F(3, 20) = 5.71$, $p = 0.641$]. In other words, the sentence intelligibility of different languages was comparable under most conditions, with the exception of the poor room acoustic condition represented by STI = 0.4.

3.4 Comparison of word and sentence intelligibility

Further analysis was achieved by comparing the sentence and word intelligibility scores for each of the four languages tested (Figure 8). The comparison graphs illustrate that there was a threshold where word and sentence

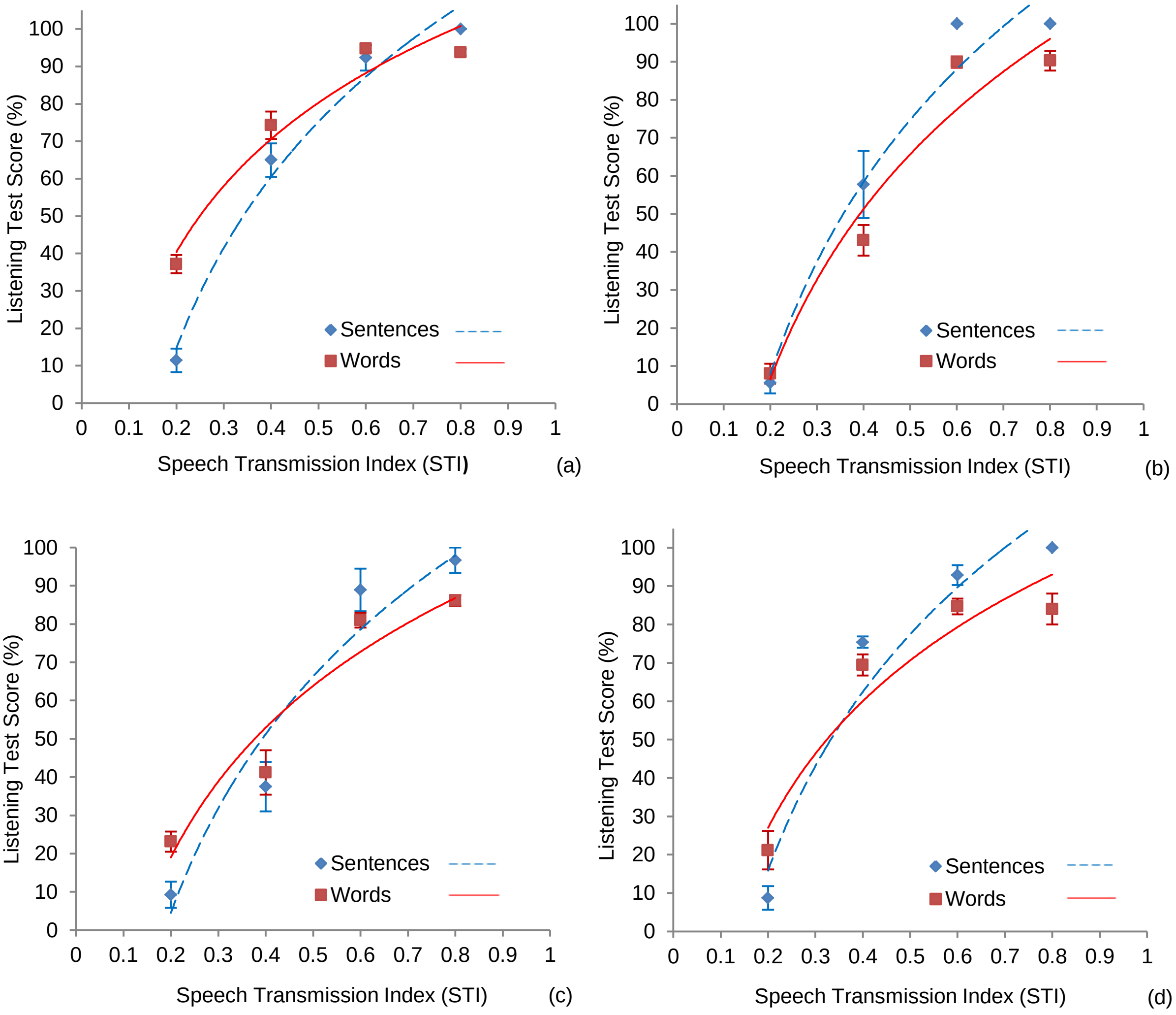


Fig. 8. Comparison between sentence and word intelligibility scores (data markers, standard errors of the means and regression lines). (a) English, (b) Polish, (c) Arabic, (d) Mandarin.

intelligibility scores intercepted. Sentence tests tended to show higher intelligibility scores than the word scores above the threshold, whilst below the threshold the word intelligibility scores were higher than the sentence scores. The STI threshold value varied between languages. For English, the threshold was STI ≈ 0.6, it was STI ≈ 0.25 for Polish, STI ≈ 0.45 for Arabic and STI ≈ 0.35 for Mandarin. Although these results are informative, it is important to note that the accuracy of these thresholds is limited, due to the limited reliability of sentence materials and variations shown by the standard errors of the means of word intelligibility scores.

The difference between word and sentence intelligibility scores varies with the distance from the threshold value, and the threshold can be interpreted as the STI level where context becomes intelligible enough [41]. When the context becomes intelligible, even if not all the words can be understood, context can be transferred from the speaker to the listener, and the sentences can ultimately become 100% intelligible. The results obtained here indicate that Polish and Mandarin have a lower threshold compared to English and Arabic. Furthermore, it can be noted that under high reverberation time and low signal-to-noise ratio, the boundaries between syllables can disappear [40]: this might justify the lower scores obtained for sentence intelligibility compared to word intelligibility below the threshold, as all the word tests used were based on monosyllabic words, unlike sentences.

Finally, it is worth pointing out again that sentence intelligibility is influenced by many factors that were not clearly defined in the sentence material used here. Therefore, sentence intelligibility comparisons between languages, as well as comparisons between word and sentence scores should be considered with caution.

4. Discussion

This section examines possible reasons for the differences in intelligibility observed between languages. In section 3.1, correlations showed that consonant-to-vowel ratios can justify variations observed under poor room acoustic conditions, but not variations observed under good room acoustic conditions. Furthermore, distinctive features identified which types of phonemes are more easily discriminated across languages, but no explanation was given of potential reasons for such differences. Analysis of the spectral content and temporal variability of the speech signals are discussed in this section, to provide a further insight into the differences observed.

First of all, spectral analysis (Figure 9) of uninterrupted speech (word test materials used and all talkers included in the signals analysed) indicates that for an identical sound pressure level of 65 dBA, high-frequencies (and in particular 4 kHz and 8 kHz) are more pronounced for English (up to +5 dB). Such high frequencies contribute to the clarity of consonants and might justify the better consonantal discrimination observed for English. By contrast, Arabic has the lowest high frequency content. It should however be noted that spectral content only provides a limited insight into the acoustical properties of languages.

A more in depth analysis can be carried out using spectrograms, which allow examining frequency content, temporal variability and signal amplitude at the same time. This has been done here to compare nasal and oral words, in order to identify possible reasons for the excellent nasality scores observed for English. Spectrograms were produced using the software *RavenLite* and are shown in Figure 10, with four words displayed per graph. The words selected represent a wide range of nasal/oral sounds within each language [21-23]. The spectrograms shown correspond to male speakers, although it can be noted that identical findings were found for female speakers. Most of the English monosyllabic word show a clear drop in high frequency amplitude between their initial and final parts (Figure 10(a)), unlike Arabic (Figure 10(b)) and Mandarin (Figure 10(c)). The drops observed in the English words correspond to vowel sounds contained between consonants (CVC sequences used), and could

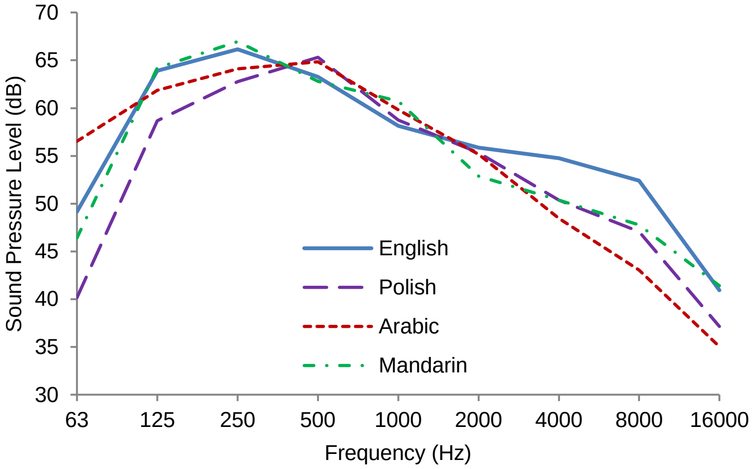
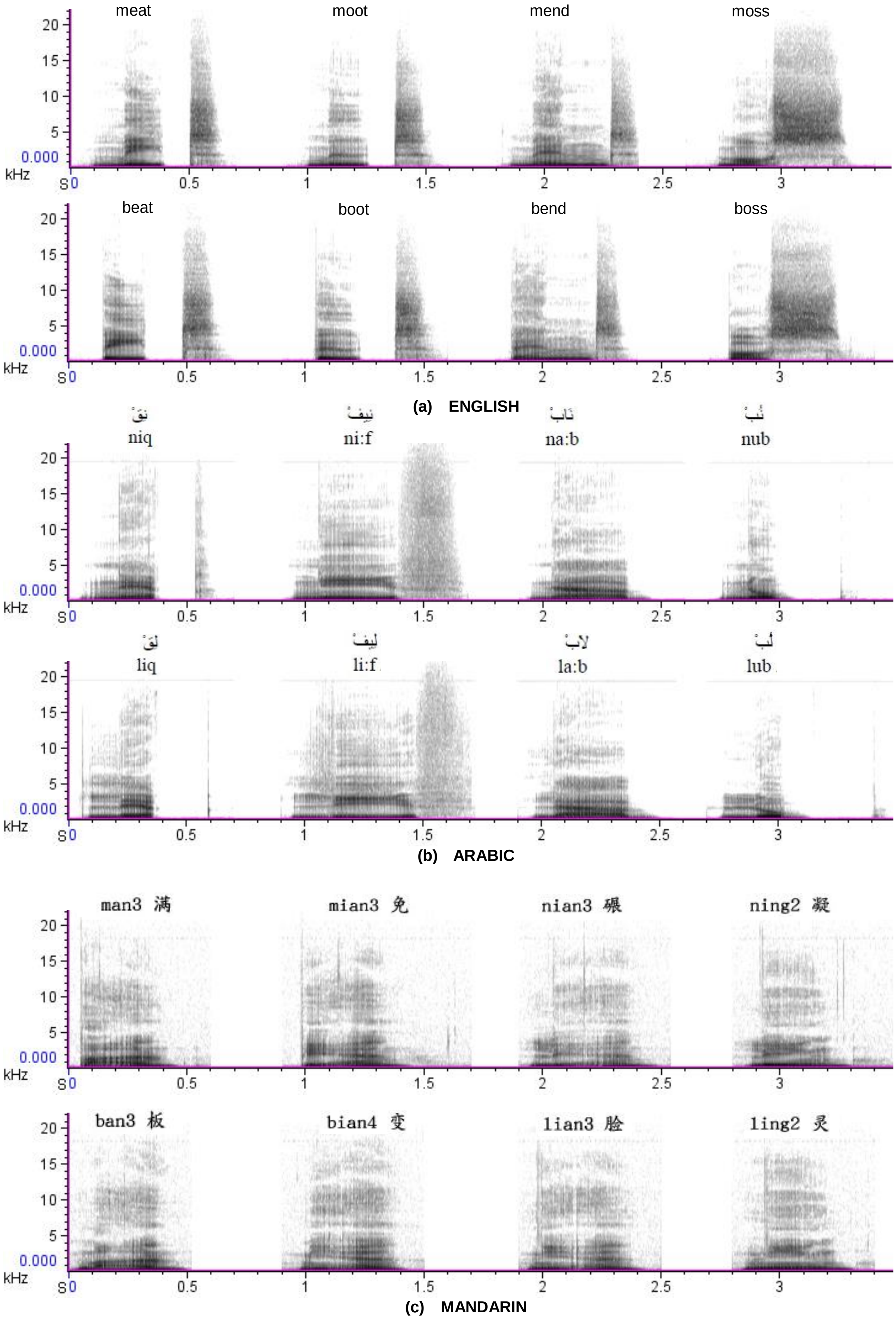


Fig. 9. Spectra of the languages tested (word tests).



(a) ENGLISH

(b) ARABIC

(c) MANDARIN

Fig. 10. Spectrograms of four nasal (top) and four oral (bottom) words, for English (a), Arabic (b) and Mandarin (c). The horizontal axis corresponds to time in seconds, while the vertical axis corresponds to frequency in kHz. The darker areas represent larger amplitude of the signal.

help better discriminate the initial consonants tested in the DRT method. Furthermore, English nasal consonants show an increase in the amplitude of high frequencies towards the beginning of the word, unlike oral consonants. By contrast, nasal and oral words of Arabic and Mandarin show similar frequency contents. The differences observed for English might help discriminate nasal vs. oral consonants, unlike Arabic and Mandarin that do not exhibit significant differences between the spectrograms of their nasal and oral words. Additional spectral analysis also highlighted a larger temporal variability (quantified by $L_{10} - L_{90}$) for English nasality around 8 kHz (+5 dB compared to Arabic and +10 dB compared to Mandarin). Furthermore, when all the distinctive features were taken into account, the temporal variability of high frequencies of both English and Mandarin were on average higher than what was found for Polish and Arabic (from +2 dB up to +7 dB across the 4-8 kHz range), again with a slightly higher $L_{10} - L_{90}$ for English at 8 kHz (+ 2 dB compared to Mandarin). A larger temporal variability means a larger dynamic range, a property that can contribute to better intelligibility by picking up of the higher peaks [2]. Spectrograms' analysis of other distinctive features showed that English consistently exhibits a drop in high frequency amplitude in the middle part of its words (vowel sounds), but the duration of this drop tends to be shorter than what was observed for nasality.

To summarise, the better intelligibility of English appears to be justified by its low consonant-to-vowel ratio, its larger high frequency content, as well as its larger temporal variability and dynamic range at high frequencies. Mandarin can also take advantage of an average consonant-to-vowel ratio and fairly high temporal variability at high frequencies, previous work having also pointed out that tonality can improve its intelligibility [15]. By contrast, the low word intelligibility of Arabic and Polish appears to be related to moderately high and high consonant-to-vowel ratios respectively, as well as low high frequency content and temporal variations. All of the above findings have been obtained from the acoustical analysis of word test materials used in the present work. In order to confirm these findings, further analysis will need to be carried out on additional test materials as well as on a larger number of talkers.

5. Main findings and limitations

The word intelligibility results revealed that statistically significant differences ($p < 0.01$) between languages occurred at the STI = 0.6, STI = 0.4 and STI = 0.2 conditions. Furthermore, word scores indicated that English was the most intelligible language. It was found that the intelligibility of nasal/oral, compact/diffuse and sibilant/non-sibilant consonants contributed to the high word intelligibility of English under all room acoustic conditions. Further analysis suggested that the better intelligibility of English might be justified by its low consonant-to-vowel ratio, its larger high frequency content, as well as its larger temporal variability and dynamic range at high frequencies. For Mandarin, results showed that the intelligibility of its airflow/no-airflow and grave/acute consonants, together with its average consonant-to-vowel ratio, made Mandarin highly intelligible under most acoustic conditions. Furthermore, it was found that Mandarin can take advantage of a fairly high temporal variability at high frequencies, previous work having also pointed out that tonality can improve its intelligibility [15]. In particular, Mandarin was highly intelligible at a relatively low signal-to-noise ratio of +5 dB (STI = 0.4), but quickly became unintelligible under extreme room acoustic conditions (STI = 0.2). Arabic and Polish were found to be the most sensitive languages to artificial noise and increased reverberation time introduced at STI = 0.4. The distinctive features' DRT results of Arabic gave an insight to this, as the decreases in intelligibility between the STI = 0.6 and STI = 0.4 conditions were as high as 75% for grave/acute consonants, and 45% for compact/diffuse consonants. The high consonant-to-vowel ratio of Polish was also found to be detrimental to its word intelligibility at poor (STI = 0.4) and very poor (STI = 0.2) conditions, as negative correlations were found between word intelligibility scores and consonant-to-vowel ratios at such conditions. Furthermore, Polish and Arabic were found to have low high frequency contents and low temporal variations, both of which might be responsible for their reduction in speech intelligibility.

Sentence scores confirmed differences in intelligibility between languages, but these were statistically significant ($p < 0.01$) only at the STI = 0.4 condition, due to the low sensitivity of sentence intelligibility at either very good or very challenging room acoustic conditions. Furthermore, comparisons of word and sentence scores showed that there is a threshold where sentence scores become higher than word scores, and this threshold varied with language.

Word intelligibility of Polish was assessed using PB words, unlike other languages for which DRT lists were available. This represents an important limitation of the current study. English data suggested that variations between DRT and PB results are small and therefore acceptable, but this alone cannot guarantee the same conclusion for Polish.

The multiple factors affecting sentence intelligibility varied across the languages used (e.g. context, familiarity, predictability, prosody and number of words), making sentence tests less comparable than word tests. To obtain a further insight into sentence intelligibility, future work could compare sentences translated across different languages. It might be difficult to obtain phonemically balanced material across all the languages tested, but this approach could at least maintain context and provide useful comparisons of real life scenarios.

Although word tests were more sensitive to room acoustic conditions than sentence tests, it is important to remember that representing a language through words only is a limitation, as the PB words used might only represent a fraction of the type of words available in a language (as this is for example the case for English, as opposed to Mandarin).

The work did not account for monolingual vs. multilingual speakers' effect, which might be partly responsible for some of the variations observed, although this effect alone cannot justify the large differences observed between languages. The fixed order of STI conditions tested might also have been responsible for order effects that could have been excluded through randomisation.

Finally, it should be noted that white noise was used in all the tests for the STI = 0.2 and STI = 0.4 conditions, but research has shown that the type of background noise used affects DRT scores. Kondo [42] found that, for identical signal-to-noise ratios, white noise produced lower DRT scores (for Japanese) than pseudo-speech noise and babble noise, and Astolfi et al. [43] also found variations in DRT scores of Italian for a variety of noise sources in primary school classrooms (traffic vs. babble vs. fan-coil vs. impact). Further research will therefore need to examine whether different types of background noise can also affect languages' intelligibility differently.

6. Conclusions

The research presented examined the impact of room acoustics on the speech intelligibility of four languages (English, Polish, Arabic and Mandarin). The study found that there was a significant difference between the word intelligibility scores of these languages. Under the same acoustic conditions (reverberation time and S/N ratio), the word intelligibility scores of each language differed between each other. The differences were found to be statistically significant for all conditions but the excellent room acoustic condition (STI = 0.8), indicating that the word intelligibility of different languages was comparable under excellent room acoustic conditions, but was not comparable under any other condition. The largest difference between word intelligibility scores (33%) was observed at STI = 0.4, in which the listeners were presented to artificial background noise for the first time, and in which reverberation time was increased. As the acoustic conditions improved, the difference decreased to 10% at STI = 0.8. It was found that distinctive features of the selected languages have an impact on the overall intelligibility, nasal/oral consonants being particularly intelligible in English. Acoustical analysis of the languages suggested that the latter might be related to the greater high frequency content of English, as well as its larger temporal variability and dynamic range at high frequencies. Furthermore, a significant correlation was found between the consonant-to-vowel ratios and the word intelligibility scores of languages at poor room acoustic conditions (STI = 0.2 and STI = 0.4), highlighting the better intelligibility of languages with lower consonant-to-vowel ratios. English, Arabic and Mandarin were tested using DRT lists, whilst Polish was assessed using PB words, because of the lack of DRT material in Polish. This is a limitation of the current study, although it can be noted that removing Polish from the analysis did not affect the main findings.

In contrast to word scores, sentence scores showed statistically significant differences between languages only at the STI = 0.4 condition, but this was justified by the lower sensitivity of sentence tests to either very good or very challenging room acoustic conditions.

Overall, the results of the study revealed that each language is affected differently by room acoustic properties, and these variations can be significant and are due to differences between the linguistic and phonological properties of each language. As the STI is affected by reverberation time and signal-to-noise ratio only, a single STI value might then be insufficient for designing a multilingual environment, or even for designing the same type of space within different countries (as previously pointed out by Li et al. [5]). Guidance values based on the STI, or on the STI qualification ratings of ISO 9921 [11], might then be appropriate for some languages (e.g. English, because of its higher intelligibility scores) but not for others. The same can also be said for any guideline solely based on acoustic parameters, e.g. recommended values of reverberation time and background noise. From the results obtained here, this appears to be particularly true for spaces that are expected to be more challenging in terms of intelligibility, e.g. open-plan spaces where excellent room acoustic conditions are difficult to achieve in practice. Furthermore, even under the "excellent" STI = 0.8 condition [11], differences between word intelligibility scores can still be non-negligible (10%), suggesting that variability across languages should be considered anyway.

The variance of properties between languages therefore demands spaces specially designed for a specific language or for a multilingual context, by taking into account the relationships between languages' intelligibility and room acoustic properties.

Acknowledgement

The authors would like to thank Professor Bernadette O'Rourke and Professor Graham H. Turner for the valuable help provided in the establishment and development of the research.

References

[1] T. Houtgast and H. J. M. Steeneken: A multi-language evaluation of the RASTI-method for estimating speech intelligibility in auditoria. Acustica 54(4) (1984) 185-199.

[2] J. Kang: Comparison of speech intelligibility between English and Chinese. Journal of the Acoustical Society of America 103(2) (1998) 1213-1216.

[3] J. X. Peng: Relationship between Chinese speech intelligibility and speech transmission index in rooms based on auralization. Speech Communication 53 (2011) 986-990.

[4] P. Zhu, F. Mo and J. Kang: Relationship between Chinese speech intelligibility and speech transmission index under reproduced general room conditions. Acta Acustica united with Acustica 100 (2014) 880-887.

[5] J. Li, R. Xia, D. Ying and Y. Yan: Investigation of objective measures for intelligibility prediction of noise-reduced speech for Chinese, Japanese, and English. Journal of the Acoustical Society of America 136(6) (2014) 3301-3312.

[6] M. L. Garcia Lecumberri and M. Cooke: Effect of masker type on native and non-native consonant perception in noise. Journal of the Acoustical Society of America 119(4) (2006) 2445-2454.

[7] K. J. Van Engen and A. R. Bradlow: Sentence recognition in native- and foreign-language multi-talker background noise. Journal of the Acoustical Society of America 121(1) (2007) 519-526.

[8] M. L. Garcia Lecumberri, M. Cooke and A. Cutler: Non-native speech perception in adverse conditions: A review. Speech Communication 52 (2010) 864-886.

[9] K. J. Van Engen. Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. Speech Communication 52 (2010) 943-953.

[10] L. L. Beranek: Acoustic Measurements, Chapter 13, 625-634, Chapter 17, 761-792, Wiley, New York, 1949.

[11] EN ISO 9921: Ergonomics – Assessment of speech communication. 2003.

[12] V. M. A. Peutz: Articulation loss of consonants as a criterion for speech transmission in a room. Journal of the Audio Engineering Society 19(11) (1971) 915-919.

[13] I. Maddieson: Consonant-Vowel Ratio. In: M. S. Dryer and M. Haspelmath, Martin (Eds.) The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2008. (Available online at http://wals.info/chapter/3, Accessed on 01/12/2012).

[14] I. Maddieson: Tone. In: M. S. Dryer and M. Haspelmath, Martin (Eds.). The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. (Available online at http://wals.info/chapter/13, Accessed on 16/06/2015).

[15] J. L. Zhang, S. Q. Qi, M. Z. Song and Q. X. Liu: On the important role of Chinese tones in speech intelligibility. Acta Acustica united with Acustica 1981 (7) 237-241 (in Chinese).

[16] C.-C. Cheng: A Synchronic Phonology of Mandarin Chinese. The Hague: Mouton, 1973.

[17] J. Kingston: The Phonetics-Phonology Interface. In: Paul DeLacy (Ed.). The Cambridge Handbook of Phonology. Cambridge University Press, New York, 2007.

[18] A. S. Hornby: Oxford Advanced Learner's Dictionary. 9th edition, Oxford University Press, Oxford, 2015.

[19] ANSI/ASA S3.2: Method for measuring the intelligibility of speech over communication systems. American National Standards Institute, 2009.

[20] R. Jakobson, G. Fant, M. Halle: Preliminaries to speech analysis. MIT Press, Cambridge, MA, 1952.

[21] W. D. Voiers: Diagnostic evaluation of speech intelligibility. In: M.E. Hawley (Ed.). Speech Intelligibility and Speaker Recognition. Stroudsburg (PA.): Dowden, Hutchinson and Ross, 1977.

[22] B. Boudraa, M. Boudraa, B. Guerin: Arabic diagnostic rhyme test using minimal pairs. Proceedings of Acoustics'08, Paris, France (2008) 2329-2333.

[23] Z. Li, E. C. Tan, I. McLoughlin, T. T. Teo: Proposal of standards for intelligibility tests of Chinese speech. IEE Proceedings Vision Image Signal Processing 147(3) (2000) 254-260.

[24] A. Pruszewicz, G. Demenko, L. Richter, T. Wika: New articulation lists for speech audiometry. Part II. Otolaryngologia Polska 48(1) (1994) 56-62.

[25] B. W. Anderson and J. T. Kalb: English verification of the STI method for estimating speech intelligibility of a communication channel. Journal of the Acoustical Society of America 81(6) (1987) 1982-1985.

[26] D. N. Kalikow, K. N. Stevens: Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. Journal of the Acoustical Society of America 61(5) (1977) 1337-1351.

[27] E. Ozimek, D. Kutzner, P. Libiszewski, A. Warzybok, J. Kocinski: The new Polish tests for speech intelligibility measurements. Proceedings of Signal Processing Algorithms, Architectures, Arrangements, and Applications (SPA) (2009) 163-168.

[28] M. Boudraa, B. Boudraa, B. Guerin: Twenty lists of ten Arabic sentences for assessment. Acta Acustica united with Acustica 86(5) (2000) 870-892.

[29] Q. Fu, M. Zhu, X. Wang: Development and validation of the Mandarin speech perception test. Journal of the Acoustical Society of America 129(6) (2011) EL267-EL273.

[30] W. D. Voiers: Evaluating processed speech using the Diagnostic Rhyme Test. Speech Technology 1(4) (1983) 30-39.

[31] D. Jones: An English Pronouncing Dictionary. Dent, London, 1917.

[32] M. F. Cohen: Effects of stimulus presentation rate upon intelligibility-test scores. Journal of the Acoustical Society of America 37 (1965) 1206.

[33] C. Diaz, C. Velazquez: A live evaluation of the RASTI method. Applied Acoustics 46(4) (1995) 363-372.

[34] AudioCheck. Online Audiogram Hearing Test. (Available online at http://www.audiocheck.net/testtones_hearingtestaudiogram.php, Accessed on 16/06/2015).

[35] D. Tabri, K. M. Abou Chacra, T. Pring T.: Speech perception in noise by monolingual, bilingual and trilingual listeners. International Journal of Language & Communication Disorders 46(4) (2011) 411-422.

[36] D. Weiss, J. J. Dempsey: Performance of bilingual speakers on the English and Spanish versions of the Hearing in Noise Test (HINT). Journal of the American Academy of Audiology 19(1) (2008) 5-17.

[37] G. A. Studebaker: A “rationalized” arcsine transform. Journal of Speech, Language, and Hearing Research 28(3) (1985) 455-462.

[38] P. Shrout, J. Fleiss: Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86 (2) (1979) 420-428.

[39] W. D. Voiers: Uses of the diagnostic rhyme test (English version) for evaluating multilingual operability in aviation communications: An exploratory investigation. Workshop on Multilingual Interoperability in Speech Technology (MIST) (September 1999) 55-60.

[40] A. Cutler, S. Butterfield: Word boundary cues in clear speech: A supplementary report. Speech Communication 10(4) (1991) 335-353.

[41] G. A. Miller, G. A. Heise, W. Lichten: The intelligibility of speech as a function of the context of the test materials. Journal of Experimental Psychology 41(5) (1951) 329-335.

[42] K. Kondo: Estimation of Speech Intelligibility Using Perceptual Speech Quality Scores. In: Speech and Language Technologies. Ivo Ipsic (Ed.), June 2011. (Available online at http://www.intechopen.com/books/speech-and-language-technologies/estimation-of-speech-intelligibility-using-perceptual-speech-quality-scores, Accessed on 01/12/2015).

[43] A. Astolfi, P. Bottalico, G. Barbato: Subjective and objective speech intelligibility investigations in primary school classrooms. Journal of the Acoustical Society of America 131(1) (2012) 247-257.