Mutual intelligibility of American, Chinese and Dutch-accented speakers of English tested by SUS and SPIN sentences

Hongyan Wang[1] *and Vincent J. van Heuven*[2]

[1]School of Foreign Languages, Shenzhen University, Shenzhen, P.R. China
[2]Phonetics Laboratory, Leiden University Centre for Linguistics, Leiden, The Netherlands
wanghongyan0069@hotmail.com, v.j.j.p.van.heuven@hum.leidenuniv.nl

Abstract

This paper investigates the mutual intelligibility of Chinese, Dutch (both foreign-language learners) and American (native language) speakers of English using SUS (Semantically Unpredictable Sentences) and SPIN (Speech in Noise) materials. We test the hypothesis that speakers and listeners who share the same native-language background have an advantage (interlanguage speech intelligibility benefit).

Index Terms: Mutual intelligibility, Intelligibility tests, Inter-language, SUS test, SPIN test, Mandarin, English, Dutch

10.21437/Interspeech.2013-129

1. Introduction

Normally sounds occur in meaningful words. Typically, when sounds occur in the context of a word in a sentence, the listener needs to get only a few of the constituent segments to piece the word together, using lexical redundancy. For instance, the last two sounds in the word *elephant* are perfectly predictable once the listener has heard *elef…*; there are simply no other words in the English lexicon than *elephant* that begin with this sequence. When the target word is embedded in a meaningful context sentence, segments in short, monosyllabic words will also be predictable. If the listener misses the initial consonant in *I heard the _at mew*, the listener will know that the mutilated word must be *cat* rather than a *rat* (or *bat* or *gnat*), let alone *mat*. In In this paper we deal with the intelligibility of meaningful words in several kinds of sentence contexts.

The first type of sentence context is a syntactically correct structure, but the words that are filled in the various slots in the structure do not yield a meaningful sequence. For instance, in *The state sang by the long week*, it is at least odd that an inanimate subject *The state* should perform an action normally only manageable by humans (i.e. singing); also, the choice of the preposition *by* would seem to be ungrammatical. These sentences were called Semantically Unpredictable Sentences or just SUS sentences [1]. They were originally constructed for the purpose of evaluating the quality of text-to-speech systems. The claim would be that the SUS test will discriminate in a highly sensitive way between small differences in speech quality, when the subjects are native listeners of the stimulus language. The test was not developed to discriminate excellent from not-so-excellent speakers and listeners.

The second type of test we used in our materials is the SPIN test, which stands for SPeech In Noise test [2]. The SPIN test requires listeners to fill in the last word of a short sentence; the final word is either highly predictable (HP) from the preceding words in the sentence (e.g. *She put her broken arm in a sling*) or not predictable from the context (low predictability, LP, e.g. *We should consider the map*). The SPIN LP sentences are more or less comparable with the SUS sentences in that the target words appear in grammatically correct word sequences, may benefit from the presence of a precursor utterance (through phonetic adaptation) but not from any semantic constraints.

We decided to include all three types of sentences in our test battery (i.e., SUS, SPIN-LP, SPIN-HP,) precisely because together they would seem to cover a very large range of listener abilities, large enough to adequately discriminate all nine combinations of speaker and listener nationalities in our study. More specifically, since the purpose of the SPIN audiology test was to discriminate between listeners from a wide range of hearing ability and that of the SUS test was to differentiate between better and poorer talking machines, one would expect therefore that the SPIN test will be rather more sensitive to differences between listeners, whilst the SUS test would be susceptible to differences between speakers.

2. Methods

We used a set of 30 SUS sentences and 50 SPIN test sentences. The materials were spoken by one male and one female Dutch, Mandarin and American speaker of English. These speakers were selected from a larger group of ten male and ten female speakers for each native language (L1) background, such that the designated speakers were closest to mean of their peer group in terms of consonant and vowel intelligibility (for details on the selection procedure see [3]). All non-native speakers were university undergraduates who did not specialize in English language and literature, who had not spent time in an English-speaking country and who had no English-speaking relatives. Speakers were digitally recorded in individual sessions in a quiet room using a Shure SM10A close-talking microphone, reading the sentences from paper and repeating the procedure as often as the experimenter deemed necessary to obtain fluent tokens of each sentence.

The materials were then presented to 36 native Dutch listeners (tested in Leiden, the Netherlands), 36 Chinese listeners (tested at Jilin University, China) and 36 American listeners (tested at the University of California at Los Angeles, USA). Within each group there were 18 male and 18 female listeners. Listeners volunteered, had no self-reported hearing problems, and were paid (the equivalent of) 10 Euros.

Stimuli were presented in a small lecture room over head-phones. Every listener heard 30 SUS sentences. These were evenly distributed over five different syntactic frames with each speaker donating one sentence to each syntactic frame. Speakers were blocked over sentences such that any listener heard each sentence only once, and every speaker donated each sentence as often as any of the other speakers.

Copyright © 2013 ISCA

In the SUS test, the entire sentence was made audible once. Then the utterance was incrementally repeated such that the utterance was truncated after the first content word on the first repetition, after the second content words in the second repetition, and so on, until the final content word was made audible. The listeners had answer sheets before them with the functions words printed for each sentence but with the content words replaced by a line of constant length, as follows: *Why does the ___ ___ the ___ ___?* After each repetition the listener was given 3 seconds to fill in the next content word in the sentence. Then the entire sentence was repeated one more time to allow the listener to make any last-minute changes that he deemed necessary.

In the SPIN test the listeners' task was just to fill in the last word of each successive sentence. No printed version of the sentences was provided. The SPIN test was developed as a diagnostic tool to determine the severity of hearing loss in audiological settings. SPIN sentences should be administered at various signal-to-noise levels. In our application we did not do this, as we noted in pilot versions of our test that the range of intelligibility across the various speaker and listener types was more or less fully covered; had we presented stimuli in noise, some of our listener groups would not have understood a single word.

The SPIN test presents sentence-final target words in high-predictability (HP) and in low-predictability (LP) contexts (see introduction). In the LP contexts the results should be roughly similar to those obtained in the SUS sentences. In both type of tests, the target words have to be understood purely from bottom-up acoustic information contained in the word itself; syntactic and semantic cues in the preceding context are useless. In the HP sentences, the words in the preceding context strongly constrain the identity of the sentence-final target word. In this condition, the SPIN test comes rather close to real-life speech recognition, where the outcome of the processing task is the result of interaction between acoustic bottom-up information and top-down semantic and syntactic information. It seems a reasonable hypothesis that the interaction between the two information sources makes heavier demands on the listener, so that the native listeners will benefit substantially from the contextual information but that the non-native listeners will be hindered by the dual-processing task – having to attend to two non-automatized processing tasks at the same time.

3. Results

3.1. SUS test

A broad phonemic transcription was produced for all the stimulus (input) and response (output) forms. To this effect the orthographic input and output forms were converted to broad IPA by hand. Non-aligned (extra) phonemes were not included in the analysis. Differences between the aligned input and output transcription were detected automatically; the scoring of the responses was done by computer. When even a single mismatch was found between input and output form, the entire word was scored as an error. In other words, every single segment in the word had to be reported correctly or else the word was not counted as a correct response.

Figure 1 presents the overall percentages of correctly reproduced words broken down first by L1 of the listener and broken down further by the L2 of the speaker. The effect of listener L1 is highly significant by a two-way ANOVA with listener and speaker L1 as fixed factors, $F(2, 312) = 669.0$ ($p < .001$). Post-hoc Scheffé tests reveal that the Chinese listeners (mean = 41% correct) performed more poorly than the Dutch (78%) and the American (79%) listeners, who did not differ from each other. There is a smaller effect of speaker L1, $F(2, 312) = 240.0$ ($p < .001$) by which Chinese speakers are poorest (52%), Dutch speakers are intermediate (70%) and Americans are best (77%). All three speaker language backgrounds differ from each other (Scheffé, $p < .05$). The effect of listener L1 is appreciably stronger than that of speaker nationality (roughly in a 3:1 ratio). The speaker × listener interaction also reached significance, $F(4, 312) = 45.9$ ($p < .001$). The interaction is clearly the result of what has been called the interlanguage speech intelligibility benefit [3, 4, 5]. For Dutch and American listeners, Chinese speakers are difficult to understand but Chinese listeners have word-recognition scores for fellow Chinese speakers which are not less than for the Dutch or American speakers. By the same token, Dutch listeners do relatively better for Dutch speakers than for speakers of other nationalities. Similarly, even American speakers have a small advantage when listening to their own speaker type.

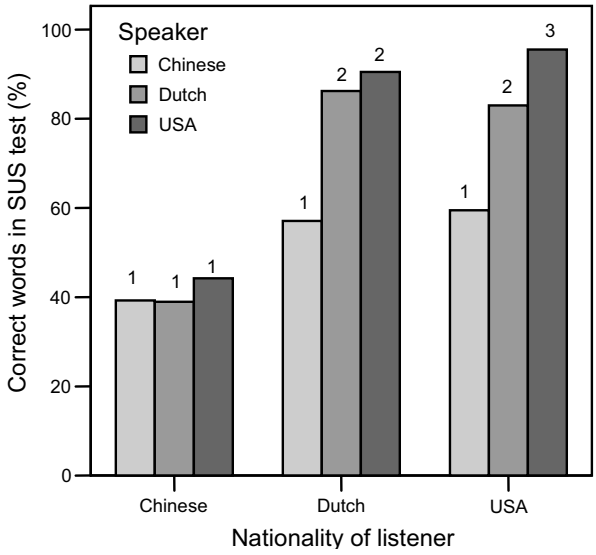

Figure 1: *Percent correct word identification in SUS test for Chinese, Dutch and American listeners broken down by accent of speakers. Numbers above the bars indicate the subgroup membership as determined by the Scheffé procedure.*

3.2. SPIN test

We will first present the results in terms of overall word recognition, once across both predictability conditions, and then separately for LP and HP sentences. In this part of the data presentation a word will be counted as an error if any component of it was not correctly reported by the listener, whether a coda consonant, a vocalic nucleus of some part of the coda.

Figure 2 presents the percentages of correctly recognized target words as defined here, broken down by L1 of listener and of speaker. The data have been accumulated over the two predictability conditions.

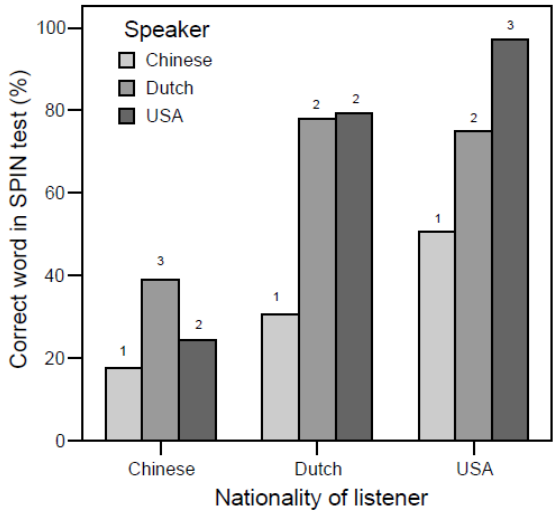

Figure 2: *Percent correct word identification in SPIN test for Chinese, Dutch and American listeners broken down by accent of speakers. Further see Figure 1.*

The data in Figure 2 were subjected to a three-way ANOVA with predictability (LP versus HP) of the targets, L1 of speaker and L1 of listener as fixed effects. The effect of listener was largest, $F(2, 630) = 807.6$ ($p < .001$), with Chinese listeners scoring 27% correct word recognition, Dutch listeners 63% and Americans 77%. All three listener groups differed significantly from each other (Scheffé, $p < .05$). A smaller effect was obtained for speaker L1, with Chinese speakers performing significantly poorer (32%) than the Dutch and American speakers (both at 67%), $F(2, 630) = 500.4$ ($p < .001$). The effect of contextual predictability is much smaller, with 52 versus 60% correct words for LP and HP, $F(1, 630 = 58.8$ ($p < .001$). There was significant interaction between speaker and listener L1, $F(4, 630) = 71.7$ ($p < .001$), which to some extent reflects interlanguage or native language benefit. However, there is one remarkable instance of foreign-language benefit: the Chinese listeners perform significantly better when the speakers are Dutch than when the speakers are either Chinese or American. Possibly, the Dutch non-natives speak more slowly and deliberately than the American native speakers, which may have helped the Chinese listeners to get more useful information from the signal than with other speaker nationalities.

There is also significant interaction between the predictability of the targets and listener L1 (but not with speaker L1), $F(2, 630) = 22.6$ ($p < .001$). Also the three-way interaction was significant, $F(4, 630) = 18.0$ ($p < .001$). We will first analyze the two-way interaction (in Figure 3), and then we will analyze the three-way interaction by presenting the results for LP and HP separately (in Figure 4A-B).

Figure 3 shows the interaction between predictability and listener L1 in detail. The figure shows that there is no effect of contextual predictability for the non-native listening groups, whether Chinese or Dutch. However, the difference is significant for the American listeners; here HP targets get better recognition scores than their LP counterparts. It seems, therefore, as if only the Americans profit from the contextual information. This would be in line with our suggestion above that non-native listeners do not recognize enough of the context to use it to their advantage.

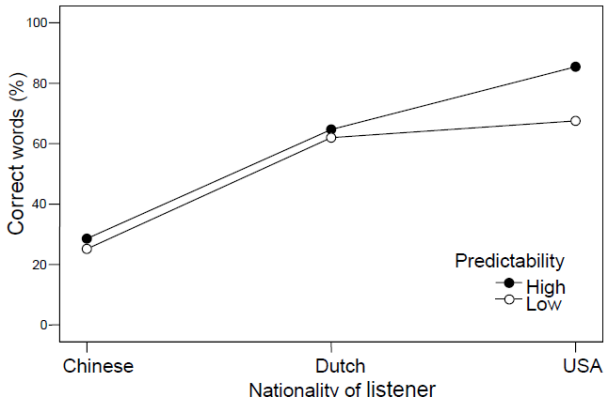

Figure 3: *Percentage of correctly recognized words in SPIN test broken down by listener nationality and by contextual predictability of targets.*

We will now present the word recognition scores for the LP and HP conditions separately. This is done in Figure 4A-B.

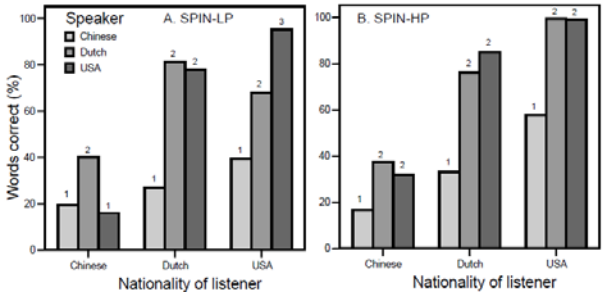

Figure 4: *Percentage of correct word identification in SPIN test for Chinese, Dutch and American listeners broken down by L1 of speakers, for low-predictability words (left) and for high-predictability words (right). Further see Figure 1.*

Comparing the two panels by listener L1, we may observe, first of all, that the Chinese listeners benefit from HP words somewhat but only if the speakers are American. Also the gain in percent correct is counteracted by a small loss of intelligibility in the HP utterances of Chinese and Dutch speakers. The Dutch listeners have no advantage of HP words at all. Apparently, they fail to use the semantic information contained in the meaningful context preceding the targets. The American listeners present an altogether different configuration of scores. If the speakers are American it does not really matter whether the words are LP (95% correct) or HP (97% correct). The quality of the pronunciation is such that recognition is close to ceiling in both conditions; there is no room for improvement due to HP. However, when the speakers are non-native, the pronunciation is relatively poor, in fact very much poorer for Chinese speakers (37% correct) and rather poorer for Dutch speakers (77%). When these speakers are tested with HP words, the Americans get so much useful information from the context that they improve their recognition scores by roughly 25 percent for Dutch speakers and by 20 percent for Chinese speakers. So, the significant three-way interaction mentioned above is due to the fact that contextual information is only used by native listeners, and only if there is room for improvement, that is, when the speakers are foreign.

4. Conclusions and discussion

Two types of sentence test were used: SUS and SPIN. The first test presented words in syntactically correct but semantically anomalous sentences, in which function words correctly constrained the content words in terms of part of speech category but not in terms of meaning. One would expect words in such sentences to be difficult to understand. The second test contained syntactically and semantically correct sentences, which were constructed such that the sentence-final target word was either highly predictable from the preceding context (HP) or not. In the low-predictability sentences (LP) the context was neutral as to the identity of the targets, i.e. they were neither made more nor less predictable than when they had been presented as citation forms. All else being equal, the order of difficulty between the three types of sentences would be SUS > SPIN-LP > SPIN-HP. Table 1 summarizes the scores for the three tests, overall and broken down by speaker and listener groups.

Table 1: *Word recognition scores for SUS, SPIN-LP and SPIN-HP sentences broken down by L1 of listener and of speaker. See appendices A9.1-2 in [2] for number of listeners, and values of SD and Se.*

L1 of		SUS scores by		SPIN scores	
Listener	Speaker	word	sentence	LP	HP
Chinese	Chinese	39	5	19	17
Chinese	Dutch	39	6	39	38
Chinese	American	44	5	18	32
Dutch	Chinese	57	17	27	33
Dutch	Dutch	86	60	81	76
Dutch	American	91	71	78	85
American	Chinese	60	18	39	58
American	Dutch	83	52	68	99
American	American	96	85	95	99
Overall		**66**	**36**	**52**	**60**

The table shows that the overall prediction does not hold: the SUS sentences are the easiest type. However, within the two types of SPIN sentences the prediction is correct: words in HP sentences are easier than words in LP sentences but the difference is rather small (but significant).

Overall word recognition scores tend to be more extreme for the SPIN sentences than for the SUS sentences. The least and most favorable speaker/listener combinations in the SUS test are Chinese/Chinese and American/American with 39 and 96% correct, respectively. Comparable numbers for the SPIN-LP test are 19 and 95%, and for the SPIN-HP test 17 and 99. Tests seem to discriminate better as they come closer to real-life speech perception, i.e. words in normally constrained, meaningful sentences. Interestingly, although the SPIN sentences were developed as audiological test materials to be presented in a range of signal-to-noise ratios, no degradation by added noise was needed in order to create a sufficiently wide range of scores in the present application of the test. Clearly, the suboptimal performance of the non-native speakers and listeners compensated for the absence of added noise.

For all three types of test (SUS, SPIN-LP, SPIN-HP) we find that the largest effect is that of listener L1. It is stronger than the effect of speaker L1 by a factor 3. For both listener and speaker effects we find that the Americans obtain the highest scores, closely followed by the Dutch nationals, while the Chinese subjects performed much more poorly. The effects of context, as determined by comparing the SPIN-LP and HP sentences, are generally minimal, except for American native listeners; only native listeners use the information contained in earlier words in the sentence to predict the identity of the sentence-final target word.

We observed clear effects of the interlanguage benefit, showing that listeners who hear speakers of their own native language background obtain better scores than when they are exposed to speech of speakers from a different nationality.

There is a remarkable discrepancy between our results and those reported by Hazan and Shi [6]. In both studies a comparison can be made of the results obtained with SUS sentences and with SPIN sentences. Hazan and Shi found word recognition scores of 12, 48 and 84 percent correct for SUS, SPIN-LP and SPIN-HP sentences, respectively. Our results reveal not the slightest difference between the scores on the SUS sentences and those on the SPIN-LP materials. Moreover, although the overall effect of LP versus HP sentences in the SPIN test is preserved in my study, the effect of context was only found for American listeners when the speakers were non-native.

Hazan and Shi recorded the materials from one male British English speaker and presented the materials to 50 native listeners. The materials were presented with a signal to noise ratio of 6 dB. It is possible, therefore, that the degradation due to the poorer signal-to-noise ratio (SNR) caused the enormous differentiation between the three tests in Hazan and Shi. We presented all our materials in quiet. As a result percent correct word recognition is close to ceiling in all three tests – but only if American native listeners respond to American speakers. When our speakers and/or listeners are non-native, the scores are rather more in the middle of the range. However, in our edition of the tests, there was virtually no difference between the LP and the HP word in the SPIN sentences (except when American listeners responded to American speakers) and the SUS sentences were some 10% better than the SPIN sentences for all conditions involving a non-native party. We must assume that the relative ease of the SUS test was caused by the way we presented the materials, i.e. not just once but repeatedly using a gating method incrementing the utterance in word-sized chunks.

The most important reason, however, why the mean SUS scores in Hazan and Shi were so low would seem to lie in the fact that these authors used the sentence as the scoring unit, whereas we computed word-recognition scores. In [6], if even one word in a SUS sentence was wrong, then the entire sentence was wrong. In order to check whether my results would be more comparable to those of Hazan and Shi, we recomputed the SUS scores using the sentence as the scoring unit. The results in terms of the sentence-based scores have been listed in Table 1, along with the word recognition scores.

Overall, the SUS scores drop from 66 to 36% when the sentence is used as the scoring unit instead of the word. As a result of this, the SUS scores are closer to those reported by Hazan and Shi (18% correct sentence recognition) but they are still considerably better. Moreover, the discriminatory power of the SUS sentence-based scores is better than that of the word-based scores. This property of the SUS test has been reported earlier by the designers of the SUS test ([1]: 388).

5. References

[1] Benoît, C., Grice M. and Hazan, V., "The SUS test: A method for the assessment of text-to speech synthesis intelligibility using Semantically Unpredictable Sentences." Speech Communication 18, 381–392, 1996.

[2] Kalikow, D. N., Stevens, K. N., and Elliott L. L., "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability." Journal of the Acoustical Society of America 61, 1337–1351, 1977.

[3] Wang, H., English as a lingua franca. Mutual intelligibility of Chinese, Dutch and American speakers of English. LOT dissertation series nr. 147, Utrecht: LOT, 2007.

[4] Bent, T. and Bradlow, A. R., "The interlanguage speech intelligibility benefit." Journal of the Acoustical Society of America, 114, 1600–1610, 2003.

[5] Heuven, V. J. van and Wang, H., "Quantifying the interlanguage speech intelligibility benefit." in Barry, W. & Trouvain, J. [Eds.] Proceedings of the 16th International Congress of Phonetic Sciences, Saarbrücken. 1729-1732, Universität des Saarlandes, 2007.

[6] Hazan, V. and Shi, B., "Individual variability in the perception of synthetic speech*." Proceedings of Eurospeech* 1993, Berlin. 1849-1852, 1993.