SUS-based Method for Speech Reception Threshold Measurement in French

Alexander Raake[*†], **Brian FG Katz**[†]

[*]Deutsche Telekom Laboratories, Berlin University of Technology
Ernst-Reuter-Platz 7, 10405 Berlin, Germany
alexander.raake@telekom.de
[†] LIMSI-CNRS, Orsay, France
brian.katz@limsi.fr

Abstract

We propose a new method for measuring the threshold of 50% sentence intelligibility in noisy or multi-source speech communication situations (Speech Reception Threshold, SRT). Our SRT-test complements those available e.g. for English, German, Dutch, Swedish and Finnish by a French test method. The approach we take is based on semantically unpredictable sentences (SUS), which can principally be created for various languages. This way, the proposed method enables better cross-language comparisons of intelligibility tests. As a starting point for the French language, a set of 288 sentences (24 lists of 12 sentences each) was created. Each of the 24 lists is optimized for homogeneity in terms of phoneme-distribution as compared to average French, and for word occurrence frequency of the employed monosyllabic keywords as derived from French language databases. Based on the optimized text material, a speech target sentence database has been recorded with a trained speaker. A test calibration was carried out to yield uniform measurement results over the set of target sentences. First intelligibility measurements show good reliability of the method.

1. Introduction

For studying human speech perception performance in noisy environments, intelligibility threshold measurements are often used. They allow the performance differences between a large number of acoustical conditions to be expressed in a compact manner. For example, advantages related to certain configurations, such as the spatial unmasking enabled when switching from monaural to binaural hearing, can be quantified in a sensitive way (Bronkhorst, 2000). Another application domain is that of relative speech quality assessment, where the intelligibility threshold can serve as a quality measure. The sensitive measurement is achieved based on the steep psychometric function of speech identification in noise. For the 50% intelligibility threshold, the so-called speech reception threshold (SRT), slopes between 10 and 20% per dB signal-to-noise ratio (SNR) have been reported in the literature (Brand and Kollmeier, 2002).

The Speech Reception Threshold is typically determined using an adaptive procedure that employs lists containing a certain number of sentences: Each list corresponds to one acoustical test condition. For each sentence of a given list, the speech reproduction level is chosen as a function of the number of keyword identification errors made on the previous sentence, targeting 50% intelligibility. The SRT is defined as the SNR at the 50% intelligibility threshold, i.e. the speech level vs. the level of the distracting signal(s).

Speech material for SRT tests has to be similarly intelligible across sentences, and across lists. In terms of phonetic, syntactic and semantic complexity, the different lists, and the sentences composing the lists should thus be comparable. Such sentence material has been developed for several languages, like English, Dutch, German and Finnish (Rothauser et al., 1969; Plomp and Mimpen, 1979; Wagener and Kollmeier, 2004; Vainio et al., 2005). In spite of numerous studies of speech quality in French, a French method for SRT measurement has not been developed to date. The available phonemically balanced French sentence material lacks the desired comparable complexity across sentences and lists, and thus cannot be used for the type of tests we aimed at (Combescure, 1981).

2. Test Development

Our goal was to assess the intelligibility linked to different configurations of a multi-user virtual speech-chat environment. Consequently, a method was needed enabling a large number of different conditions to be assessed in one test. This requirement is bound to a rather large number of different sentence lists, in order to minimize a potential training effect that ultimately could enhance intelligibility over the time. Moreover, we wanted our method to easily be portable to other languages. The test methods developed for other languages fulfil at least some of these criteria. Sentence material limited in the size of the underlying lexicon may lead to a comparable complexity over lists and may more easily be translated into other languages, but is typically accompanied by a measurable training effect (Wagener and Kollmeier, 2004). In turn, sentences that better reflect the actual usage of the language — e.g. by employing a far larger lexicon and more conversation-typical topics — reduce training effects, but are not easily portable to other languages.

2.1. SUS Database

As a compromise between training effect and homogeneous sentence complexity, we based our test method on the framework of semantically unpredictable sentences (Benoît et al., 1996). The underlying syntactic structures are very similar and thus of comparable complexity, and are available for different languages.

2.1.1. Text Material

Typically, the error-rate underlying an adaptive SRT-test is determined based on wrongly identified keywords. In order to achieve four keywords per sentence, only four of

Nr.	question word	noun	adj.	trans. verb	intr. verb	prepos.	noun	adj.	rel. pron.	intr. verb	punct.
1		Le chien			lutte	sous	la plage	rouge			.
2		La robe	sourde	voit			l'ours				.
4	Quand	l'or		prend-il			la peur	beige			?
5		Le blé		tente			l'heure		qui	tremble	.

Table 1: Examples for the syntactic structures of the SUS text material with original indexing as in (Benoît et al., 1996).

the five syntactic structures available from (Benoît et al., 1996) were used for our method (words considered as keywords are nouns, transitive and intransitive verbs, and adjectives). Examples for the employed structures (and some of our sentences) are shown in Table 1.

Due to the considerably larger number of sentences to be created for our tests, the original word-Lexika and sentence material from (Benoît et al., 1996) had to be extended. Additional monosyllabic, singular French words were selected from the keyword-categories. A specific database was created based on existing French word databases, according to an automatized protocol. The information contained in the final database is summarized in Table 2. In all cases, the direct article is used, and the third person singular. All verbs are in present tense, 3^{rd} person indicative, active. Since especially for verbs the form used by the SUS structures differs in the number of syllables from the canonical form, the database was initially created for all mono- and bi-syllabic words from the three keyword categories. The database was then reduced to words that are monosyllabic in their relevant form (except from a sometimes pronounced schwa at the end). In addition, word occurrence frequency thresholds were set in order to limit the number of the remaining entries to about three times the required number of entries per lexical category. For 288 sentences of the chosen structures, with three repetitions per word, the numbers aimed at are 192 nouns, 108 verbs, and 72 adjectives. Moreover, in case of homophones between word categories, only the most frequent one in terms of word occurrence in French was kept, in order to reduce possible ambiguities.

The initial database was optimized and further reduced to the target number of entries based on a Chi-square criterion used to yield a high agreement in phoneme-distribution with a phoneme distribution from average French. For each word category as well as on average, the phoneme distribution was compared to an average French phoneme distribution derived from the database-entries for orthographic frequency from BRULEX and LEXIQUE, and a grapheme-to-phoneme converted form obtained using Graphon[1]. The average phoneme distribution was derived as the arithmetic mean of three relative distributions, one from BRULEX, and two from LEXIQUE (one from the database Frantext and one from a web-related text-database). The procedure was iterative, and the words finally kept were the ones showing the best fit to the reference phoneme distribution (10000 iterations). The different phoneme-distributions are shown in Figure 1.

The sentence lists were created to yield — per list — an equal number of sentences of each of the four syntactic structures. Thus, for the twelve sentences of each list, three samples of the four syntactic structures were employed. Sentence creation was performed automatically, following an iterative approach with (1000 iterations):

1. The keywords necessary for each list were drawn at random from the Lexicon, with three repetitions of each word between different lists.
2. The words were initially represented by numbers and category indices. These were assigned in such a manner, that the number of re-occurrences of word-combinations between lists was minimized. To this aim, four matrices were created containing the real indices of the words of each keyword category (nouns, adjectives, transitive and intransitive verbs). These matrices were used to map a randomly created set of second indices to the actual words.
3. The fit to the average French phoneme distribution mentioned above was determined (per list and on average).
4. Auxiliary words were randomly selected from appropriate additional word-lists.
5. From the selected auxiliary- and keywords, a set of twelve sentences is created based on the syntactic structures outlined in Table 1.
6. For all keywords chosen for a given list, the sum of the word-frequencies in general French is calculated (per list and per lexical category). If the deviation between the current list and previous lists is larger than 20%, the current iteration is restarted from the beginning.
7. From the given number of 1000 iterations, the one with the best Chi-square statistics was kept.
8. Finally, keywords from the same category were interchanged within lists if the randomly created sentences coincidentally made sense (with the help of two French native speakers).

The characteristics of the resulting French sentence material can be summarized as follows:

- Word Lexicon of the most frequent monosyllabic words based on word frequencies tabulated in database BRULEX, and with optimized phoneme distribution as compared to a reference distribution calculated from the databases BRULEX and LEXIQUE.

[1]Graphon is a grapheme-to-phoneme conversion engine showing less than 1% word error rate (Boula de Mareüil, 1997; Yvon et al., 1998).

Index	Nouns (n.)	Verbs (v.)	Adjectives (a.)	Source
1	canonical form			Morphalou
2	gender	Transitive?	male form	Morphalou
3	-	3^{rd} pers. sing. pres. ind. act.	female form	Morphalou
4	frequency of lexical entry			BRULEX
5	frequency of orthographic entry			BRULEX
6	frequency of orthographic entry			LEXIQUE (text)
7	frequency of orthographic entry			LEXIQUE (web)
8	number of syllables			LEXIQUE (n., v.); BRULEX (a.)
9	phonemic transcriptions of canonical and non-canonical forms			Graphon

Table 2: Information contained in word database underling the SUS text material. Resources: Morphalou (Romary et al., 2004; Salmon-Alt et al., 2004), BRULEX (Content et al., 1990), and LEXIQUE (New et al., 2004). Graphon is a grapheme-to-phoneme conversion engine showing less than 1% word error rate (Boula de Mareüil, 1997; Yvon et al., 1998).

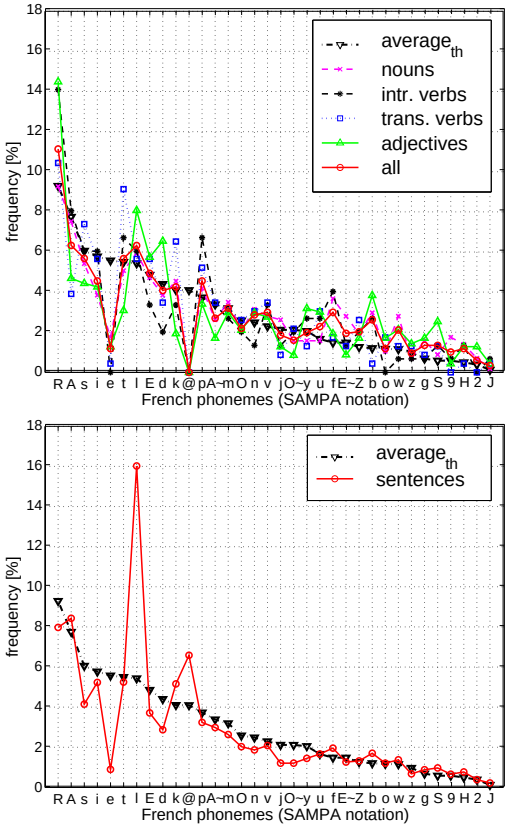

Figure 1: Phoneme distributions, in SAMPA notation (Gibbon et al., 1997). Top: Phoneme distributions of word Lexika per lexical category vs. reference (general French phoneme distribution — "th."; theoretical reference — from databases BRULEX and LEXIQUE (Content et al., 1990; New et al., 2004)). Bottom: Phoneme distribution for final SUS sentence corpus vs. reference.

- Avoidance of ambiguities where words from one category are misinterpreted as words from another category.
- Lexicon of 192 nouns, 108 verbs, 72 adjectives.
- Three repetitions of each word from the underlying lexicon.
- 24 lists of 12 sentences, leading to a total of 288 sentences. Each list contains 48 keywords.
- Minimized re-occurrence of word-pairs in another list.
- Chi-square-based maximization of the agreement between the phoneme-distribution of each list and the phoneme-distribution characteristic of the French language.
- Equalization of the word-frequencies per lexical category and per list.

Examples of the sentences are provided in Table 1.

2.1.2. Speech Material

The resulting sentences were recorded with a professional speaker of medium voice timbre. The speaker was instructed to read the sentences clearly but with a natural intonation reflecting the syntactic structures in an effort to avoid the rather artificial reading style often used by untrained readers of SUS-sentences, which reflects the lack of semantic predictability (Raake, 2002). The recordings were made with high-quality audio hardware in a soundproof acoustically treated environment. The samples were directly recorded to hard-disk at 48 kHz, 16 bits. After recording, the sentences were adjusted to an equal RMS (root mean square) level of -22 dB rel. overload of the digital system.

As reference distracter, a 60 s long speech-shaped stationary noise sample was created by twenty times overlaying and scaling of the original 288 sentences, with randomly selected, faded start and end instances. The resulting noise sample shows the same long-term spectrum as the underlying speech. The reference distracter was scaled to the same RMS as the target sentences.

3. Test System

The SRT-test presented in this paper is fully automatic. The sentences are entered by the subjects on the test PC

screen. The test program notifies the subject when it detects a typographical error by comparing the word entries to a large French vocabulary. This mechanism helps preventing at least those typing errors that yield non-existent French words (much larger than the keyword Lexicon). In order not to emphasize certain words, the subjects are only informed that one or more typographical errors were detected, but not where in the sentence the error was found. The typographical error detection was implemented in a simple fashion using the option of the Unix `grep` command of providing different words to be found in a given file.

```
grep [opt.] word1|...|wordN file
```

The corresponding dictionary file contains all up to three-syllabic words from the database LEXIQUE (New et al., 2004). With the automatic notification of typographical errors, the effect of spelling errors was minimized. Since the keywords are monosyllabic frequent French words, the occurrence of spelling difficulties were further limited.
The number of wrongly identified keywords is determined based on an alignment of the subject's entry with the target sentence. The alignment is achieved using NIST's sclite, which is based on a dynamic programming approach (NIST, 2005). In order to avoid orthographic ambiguities typical of French (e.g. la mer vs. la mère, which are phonemically identical), the alignment is carried out on automatically created phonetic transcriptions of the subject's entries and of the actual target sentence using Graphon.
The approach of automatically detecting and counting keyword identification errors, which is not commonly used in adaptive SRT-tests, was chosen due to the disadvantages of the two alternative methods:

1. E.g. in the tests by (Hawley et al., 2004), the correct target sentence is presented to the subjects on screen after they have delivered their transcript. Keywords are highlighted, and the subjects are instructed to identify the keywords they did not understand correctly. However, if the test subjects count the number of errors themselves, a potential training effect of the word corpus may be assisted by the visual presentation. Moreover, this method is more time-consuming than our automatic approach, and thus reduces the number of conditions that can be presented in one test run. Also, such a method relies on the "honesty" of the test subjects, who may re-iterate on what they think they heard based on the presented correct target sentence.

2. Another method often employed is to have the test-supervisor sit in the listening booth and the subjects verbally repeat what they understood. The experimenter marks the wrongly identified keywords on a prepared sheet. This method requires a high proficiency in the test-language from the supervisor, and is considerably more time- and resource-consuming for the experimenter than our approach.

The actual SRT-test is conducted as follows:

- The distracter signal is always played out at a fixed level throughout one list. The distracter signal is picked randomly from the distracter sound file, which was much longer than the target sentences. In addition, trailing periods of 500 ms were added in the beginning and the end of the target sentence to determine the necessary distracter duration.

- For the first of the twelve sentences of each list, the subject can repeatedly listen to the combined target and distracter samples. At each repetition, the target level is increased by 3 dB (starting at -25 dB signal-to-distracter-ratio). The subject switches to the next sentence, when she/he has the impression of understanding at least 50% of the sentence. The corresponding target level is stored and used as the starting play-out level for the following adaptive procedure, i.e. as the level of sentence # 2/12. Since it only serves as the starting point of the procedure, the possible inaccuracy of this subject-decision is of minor importance for the remaining test.

- For sentences $i \in [3, 12]$ the level is determined based on the number of wrongly detected keywords according to Equations (1)-(3) (Brand and Kollmeier, 2002):

$$L_k = L_{k-1} + \Delta L_k \quad (1)$$

$$\Delta L_k = -\frac{f(i) \cdot (prev - 0.5)}{slope} \quad (2)$$

$$f(i) = 1.5 \cdot 1.41^{-i} \quad (3)$$

  Here, L_k is the level for the current target sentence. ΔL_k is the level difference to the previous target sentence. For its derivation, the ratio of correctly identified keywords from the previous sentence is used ($prev$), and an estimated slope of the psychometric function of intelligibility over signal-to-distracter-ratio of $slope = 0.15$, as it was proposed by (Brand and Kollmeier, 2002). The function $f(i)$ steers the convergence of the method: The higher the number of level-inversions i, the lower $f(i)$, and thus the lower the amount of level-change. For the reference distracter, i.e. the speech-shaped stationary noise, the slope of approximately 15%/dB was verified based on the calibration tests described in Section 4. Since this slope reflects a good compromise for different tests described in the literature (Brand and Kollmeier, 2002), it was used throughout our tests.

- The SRT is determined as the average level difference between target and distracter over the last 8 sentences (# 5 – 12).

4. Test Calibration

In a two-step optimization procedure, the speech material was adjusted to homogeneous speech intelligibility in speech-shaped noise:

1. The average SRT was determined in an adaptive SRT intelligibility test for a sample of six of the 24 lists.

2. The estimated SRT was used as the SNR for all list/noise combinations, and all sentences were presented to a number of six subjects in a simple intelligibility test. From the word-errors determined in the

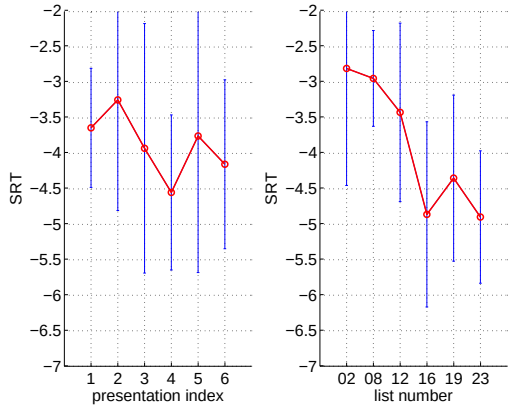

Figure 2: Test results of first calibration test. Left: Results plotted over presentation index (i.e. averaged over different lists and over subjects). Right: Results plotted over list index (i.e. averaged over the same lists and over subjects).

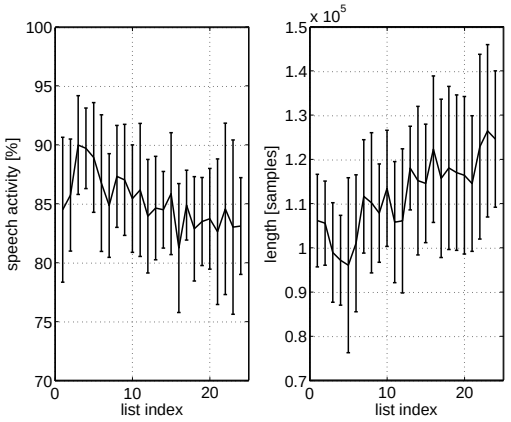

Figure 3: Mean speech activity (left) and sentence duration (right) as a function of list index (error-bars represent standard deviations).

test, a level correction was derived in order to achieve a more homogeneous intelligibility of the different sentences.

4.1. First SRT estimate

The first calibration test was run with 12 subjects. Six of the 24 lists were presented to the subjects employing a digram-balanced test design according to (Wagenaar, 1969), with n=12. As distracter, the speech-shaped stationary noise was employed in all cases. As in all other tests, the sound samples were presented via Headphones (Sennheiser HD 600). In this test, an average SRT of -4.37 dB was obtained. From the intelligibility-score/target level combinations collected in the test, a slope of approximately 15%/dB could be observed. Moreover, the SRT test results indicate both a strong subject-dependence, and a list index effect. In turn, with an initial training phase using two unscored runs with one training list each, no training effect was observed. In Figure 2, the SRT results for this first calibration test are plotted averaged over the test subjects. The left picture shows the results in the order of presentation, i.e. averaged over different lists. As can be seen from this picture and is further verified in the actual SRT tests, no significant training effect appears. The right graph depicts the results in the order of the underlying list indices. Here, each entry is averaged over the same list, showing the previously mentioned list effect.

In order to further investigate the source of the list-effect (i.e. that the measured SRT decreases with increasing list index or recording duration), the average speech activity and sentence-sample duration were determined for each list. Therefore, a simple voice activity detection was employed, which is based on a fixed level threshold. The trailing pauses at the beginning and end of each sentence-sample were excluded from the analysis, since these depend on the sample preparation rather than on the sentences themselves. Then, the speech activity of each sentence file was derived as the ratio of samples above the predefined level threshold vs. the overall number of samples. The second measure used for analyses simply was the overall number of samples. The results of these recording analyses are illustrated in Figure 3. As is depicted in the two graphs, speech activity and sample length change over recording duration (i.e. list index). Obviously, the reading style of the speaker slightly changed over time, with increasing pauses and thus increasing sentence durations. This observation is in line with the observed decrease of the SRT with increasing list index, since intelligibility is facilitated by the increasingly slower reading style.

4.2. Level Calibration

The second calibration test was conducted in order to reduce the SRT-differences between lists observed in the first calibration phase. Therefore, the target level was fixed to the SRT measured in the first test, i.e. $SRT = -4.37$. With this setting, a classical speech intelligibility test against the stationary speech-noise was run. Here, we assumed that the keyword intelligibility lies around the threshold of 50%. From the intelligibility scores obtained from the six subjects who participated in this second test, an error-rate-dependent level-correction was determined for each sentence, similarly to (Plomp and Mimpen, 1979). Corrections were only employed when the average intelligibility score for a given sentence was below 35% or higher than 65%, according to Equation (4). The corrections were limited to at most $\pm 2\ dB$.

$$LevCor(i) = \begin{cases} (0.35 - I(i))/0.15, & I(i) < 0.35; \\ (0.65 - I(i))/0.15, & I(i) > 0.65. \end{cases} \quad (4)$$

5. Test Application

A number of SRT tests on speech intelligibility in multi-source configurations in virtual auditory listening spaces have been carried out with the described method. All in all, three test series with 16 test conditions each were conducted, with 10 normal hearing subjects per test series. For clarity and brevity, we here will restrict ourselves to the reference condition with the speech-shaped stationary noise distracter and without e.g. spatial processing. The reference was used as the first test condition in all three tests.

The average SRT for this condition over the three tests is $SRT = -4.7\ dB$. Moreover, the results for this condition show a standard deviation below 1.2 dB for all three tests. Thus, the test method we have developed delivers accurate SRT-estimates in this case, which are comparable to or better than those obtained in other studies (Hawley et al., 2004; Vainio et al., 2005).

6. Conclusions

We have demonstrated a new method for measuring the speech reception threshold in French. It is based on a phonemically balanced keyword corpus used as the basis for automatically generated semantically unpredictable sentences. After pre-tests and calibration, the method delivers highly reliable estimates of the SRT in case of a stationary speech-shaped noise source ($SRT = -4.6$). For our tests, we employed a new fully automatic procedure in order to reduce the considerable effort typically linked to adaptive SRT-tests. Due to the design of the method, error-sources like spelling errors by the subjects have been reduced to a far extent. Future work will address a detailed analysis of the effect of typographical errors on the accuracy of our automatic method. In addition, the method will be translated into other languages such as German and English, in order to investigate cross-language validity and reliability. The French method and SRT-corpus will further be used in our future studies of speech intelligibility in real and virtual environments. It is our aim to make the SUS text and speech corpora and, if desirable, the automated test available to interested parties. Please contact the authors for more information.

Acknowledgement

This work was conducted in the framework of the French Ministry of Research funded RNTL-project (Réseau National des Technologies Logicielles) OPERA (Optimisation PErceptive du Rendu Audio — Application au chat sonore 3D multi-utilisateurs et aux environnements virtuels réalistes: `http://www-sop.inria.fr/reves/OPERA`). The authors wish to thank Patrick Paroubek, Philippe Boula de Mareüil and Marie-Neige Garcia for fruitful discussions and their support during the setting up of the SUS text database.

7. References

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

P. Boula de Mareüil. 1997. *Étude linguistique appliquée à la synthèse de la parole à partir du texte*. Ph.D. thesis, University of Paris XI, Orsay.

T. Brand and B. Kollmeier. 2002. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J. Acoust. Soc. Am.*, 111:2801–2810.

A. Bronkhorst. 2000. The Cocktail Party phenomenon: A review of research on speech intelligibility in multi-talker conditions. *Acta Acustica utd w. Acustica*, 86:117–128.

P. Combescure. 1981. 20 listes de dix phrases phoétiquement équilibrées. *Revue d'acoustique*, 56:34–38.

A. Content, P. Mousty, and M. Radeau. 1990. BRULEX : Une base de données lexicales informatisée pour le Français écrit et parlé. *L'Année Psychologique*, 90:551–566.

Dafydd Gibbon, Roger Moore, and Richard Winski. 1997. *Handbook on Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, D–Berlin.

M. L. Hawley, R. Litovsky, and J. Culling. 2004. The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *J. Acoust. Soc. Am.*, 115:5.

B. New, C. Pallier, M. Brysbaert, and L. Ferrand. 2004. Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers*, 36:516–524.

NIST. 2005. Speech recognition scoring toolkit (sctk) v.2.1.1. http://www.nist.gov/speech/tools/index.htm.

R. Plomp and A. M. Mimpen. 1979. Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 18:43–52.

A. Raake. 2002. Does the content of speech influence its perceived sound quality? In *Proceedings 3rd Int. Conf. on Language Resources and Evaluation (LREC 2002)*, volume 4, pages 1170–1176, Las-Palmas, Spain.

L. Romary, S. Salmon-Alt, and G. Francopoulo. 2004. Standards going concrete : from LMF to Morphalou. In *Proceedings Workshop on Electronic Dictionaries (Coling 2004)*, Geneva, Switzerland.

E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock. 1969. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio and Electroacoustics*, 17:225–246.

S. Salmon-Alt, L. Romary, J.-M. Pierrel, I. Falk, E. Petitjean, P. Bernard, J.-P. Chauveau, C. Jadelot, and M. Valette. 2004. Lexique morphologique ouvert du Français (version 1.0.1). http://actarus.atilf.fr/morphalou/.

M. Vainio, A. Suni, H. Jrvelinen, J. Jrvikivi, and V.-V. Mattila. 2005. Developing a speech intelligibility test based on measuring speech reception thresholds in noise for english and finnish. *J. Acoust. Soc. Am.*, 118:1742–1750.

W. A. Wagenaar. 1969. Note on the construction of digram-balanced Latin Squares. *Psychological Bulletin*, 72:384–386.

K. Wagener and B. Kollmeier. 2004. Göttinger und oldenburger satztest. *Z Audiol*, 43:134–141.

F. Yvon, P. Boula de Mareüil, C. d'Alessandro, V. Aubergé, M. Bagein, G. Bailly, F. Béchet, S. Foukia, J.-F. Goldman, E. Keller, D. O'Shaughnessy, V. Pagel, F. Sannier, J. Véronis, and B. Zellner. 1998. Objective evaluation of grapheme to phoneme conversion for text-to-speech synthesis in french. *Computer Speech and Language*, 12:393–410.