ELSEVIER

Speech Communication 18 (1996) 381–392

SPEECH COMMUNICATION

The SUS test [1]: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences

Christian Benoît [a,*], Martine Grice [b], Valérie Hazan [c]

[a] Institut de la Communication Parlée, UPRESA CNRS no. 5009, INPG / ENSERG – Université Stendhal, BP 25X, 38040 Grenoble Cédex 9, France

[b] Phonetik, Bau 17, Universität des Saarlandes, 6600 Saarbrücken, Germany

[c] Department of Phonetics and Linguistics, Wolfson House, University College London, 4 Stephenson Way, London NW1 2HE, United Kingdom

Received 23 February 1995; revised 21 February 1996

Abstract

This paper describes the experimental set-up used by the SAM (ESPRIT-BRA Project no. 2589: *Multilingual Speech Input / Output: Assessment, Methodology and Standardisation*) group for evaluating the intelligibility of text-to-speech systems at sentence level. The SUS test measures overall intelligibility of Semantically Unpredictable Sentences which can be automatically generated using five basic syntactic structures and a number of lexicons containing the most frequently occurring mini-syllabic words in each language. The sentence material has the advantage of not being fixed, as words can be extracted from the lexicons randomly to form a new set of sentences each time the test is run. Various text-to-speech systems in a number of languages have been evaluated using this test. Results have demonstrated that the SUS test is effective and that it allows for reliable comparison across synthesisers provided guidelines are followed carefully regarding the definition of the test material and actual running of the test. These recommendations are the result of experience gained during the SAM project and beyond. They are presented here so as to provide users with a standardized evaluation method which is flexible and easy to use and is applicable to a number of different languages.

Zusammenfassung

In diesem Artikel wird das experimentelle Protokol des europäischen Projekts SAM (Projekt ESPRIT Nr.2589: *Multilingual Speech Input / Output: Assessment, Methodology and Standardisation*) für die Beurteilung des Sprachverstehens auf Satzebene von unterschiedlichen Synthetisoren vorgestellt. Der SUS-Test besteht darin, das durchschnittliche Verstehen von "semantisch unvorhersagbaren Sätzen" (Semantically Unpredictable Sentences) zu messen. Diese Sätze werden mit Hilfe mehrerer Lexika von hochfrequenten, kurzsilbigen Wörtern der jeweiligen Sprache mit einer Zufallsauswahl erzeugt, indem fünf elementare syntaktische Strukturen respektiert werden. Der Vorteil dieses Satzkorpus liegt in

[*] Corresponding author.

[1] The SUS ("Semantically Unpredictable Sentences") test has been defined as part of a battery of standardised synthetic speech evaluation tests within the context of a multi-lingual European project on synthesiser and recogniser assessment (ESPRIT "SAM" Project no. 2589).

PII S0167-6393(96)00026-X

der Flexibilität, da bei jedem Test die Wörter für die Bildung neuer Sätze zufällig aus den Lexika gewählt werden. Zahlreiche Vollsynthesesysteme in verschiedenen Sprachen konnten so mit Hilfe dieses Tests beurteilt werden. Die erzielten Resultate zeigen daß der SUS-Test effizient ist und zuverlässige Vergleiche zwischen den Synthetisoren ermöglicht, vorausgesetzt daß eine Reihe von Bedingungen in der Definition und dem Ablauf des Tests selbst berücksichtigt werden. Es handelt sich um Empfehlungen, die auf den Erfahrungen des SAM-Projekts – und darüber hinaus – beruhen, die hier in einer Form dargestellt werden daß sie dem Benutzer dieses Tests als normalisiertes Meßwerkzeug dienen können.

Résumé

Cet article décrit le protocole expérimental retenu par les participants au projet européen SAM (Projet ESPRIT no. 2589: *Multilingual Speech Input / Output: Assessment, Methodology and Standardisation*) pour l'évaluation d'intelligibilité des synthétiseurs de parole au niveau de la phrase. Le SUS test consiste à mesurer l'intelligibilité moyenne d'un jeu de "phrases sémantiquement imprédictibles" (Semantically Unpredictable Sentences), générées aléatoirement à partir de quelques lexiques des mots minisyllabiques les plus fréquents dans chaque langue, selon cinq structures syntaxiques élémentaires. L'avantage de ce corpus de phrases est de ne pas être figé, puisque les mots sont extraits aléatoirement des lexiques pour former de nouvelles phrases à chaque nouveau test. De nombreux synthétiseurs de parole à partir du texte ont été évalués, au moyen de ce test, dans plusieurs langues. Les résultats obtenus ont montré que le SUS test est efficace et permet des comparaisons fiables entre synthétiseurs à condition de respecter un certain nombre de contraintes dans la définition et le déroulement du test lui-même. Ce sont les recommendations issues de l'expérience accumulée pendant le projet SAM et au-delà qui sont présentées ici, de façon à ce que les utilisateurs de ce test puissent se servir d'un outil de mesure "standardisé".

Keywords: Synthesis assessment; Intelligibility; Multilinguality; Standardization

1. Introduction

The development of standardised material and methodologies for the assessment of speech output systems has been given a considerable degree of attention in the last few years (e.g. the European projects SAM, EAGLES, EUROCOCOSDA, etc.). It is well accepted that the assessment of text-to-speech systems should be carried out at different levels (Benoît and Pols, 1992; Fourcin, 1992; Pols et al., 1992) as there is a need both for analytic tests, used during product development to optimize various components of the system, and for more global tests to give a general evaluation of a system for an end-user.

Since connected speech is often synthesized, sentence-length material is required within a larger test battery for quantifying the overall intelligibility of a system, enabling its performance to be evaluated on words, without as well as with sentence accent, and across word boundaries. In addition, sentence-length stretches are the minimum requirement for an evaluation of rules for prosody. However, the use of meaningful sentences as sentence material is problematic for a number of reasons. Meaningful sentences provide semantic and syntactic contextual cues whose effect on intelligibility scores cannot readily be quantified. This makes it virtually impossible to construct lists of sentences which are balanced in terms of their complexity. Finally, most of the text-to-speech systems already on the market reach perfect intelligibility when sentences are simple and meaningful.

Two approaches have been taken to control for the effect of context. SPIN (SPeech In Noise) sentences (Kalikow et al., 1977), which were primarily developed as audiometric speech material, were designed so that the effects of semantic information at sentence level were controlled by presenting test words in either high- (HP) or low-probability (LP) contexts (e.g., HP: "We're lost, so let's look at the *map*" versus LP: "I should have considered the *map*"). The difference between the scores obtained for high-probability words and low-probability words gives an indication of the amount of information provided by the sentence context. The main disadvantages with this material is that it is lengthy to administrate, only tests a single word category (the last noun in the sentence), and that it consists of only ten fixed lists and does therefore not provide enough material for large-scale comparative tests, as sentences cannot be used more than once because of strong learning effects.

A method for presenting words in connected speech with reduced contextual cues was first used in psycholinguistic studies by Miller and Isard (1963) and in the assessment of synthetic speech by Nye and Gaitenby (1974). This involves the use of syntactically acceptable but semantically anomalous sentences. Such sentences respect syntactic rules but violate semantics in that they do not take selection restrictions into account. For example, in a semantically anomalous sentence such as "He drank the wall", the syntactic structure of the sentence is correct. However, as far as semantics is concerned, "drink" would impose a selection restriction to the effect that it requires a liquid direct object. Since the direct object, "the wall", is not liquid, the sentence can be said to be semantically anomalous. The listener will consequently receive cues as to syntactic category only and will be able to make no further predictions about word identity. In a sentence such as Chomsky (1957)'s often quoted "Colorless green ideas sleep furiously.", the listener will probably expect to hear an adverb or adverbial in final position but will not be able to predict more than that. Since their constituting words are selected at random, the sentences discussed below are referred to as *semantically unpredictable sentences* (SUS) because they do not necessarily violate selection restrictions, they simply fail to take account of them. The majority of sentences generated in this way are semantically anomalous. However, should a semantically acceptable sentence be generated, it would also be categorised as unpredictable in the context of such a test.

There are several advantages to this type of material. It controls for the effect of semantic information and this reduction in contextual cues means that ceiling effects are avoided. In addition, because sentences are randomly generated using a fixed vocabulary, it is possible to generate a very large number of different sentences from the same lists. This reduces the strong learning effects known to occur if sentences are listened to more than once. The problem seen with existing material of this type is that it is only available for a restricted number of languages (mainly English and French), and that only one simple syntactic structure is used. Grice and Hazan (1989) report on a pilot test where one structure appeared to impose a rigid template on the listener. It is assumed that, if presented with a variety of structures, the listener can less easily predict the structure and will react afresh to each new sentence.

This work therefore had two aims. The first was to develop material for a number of European languages, which would, as far as possible, be comparable in terms of structure and content: the languages investigated are closely related and are syntactically similar. It was therefore possible to define several structures, rules and constraints for drawing up comparable sets of sentences. The second aim was to construct material which would include a greater variety of prosodic patterns and syntactic categories; to this end, five sentence structures were used in the generation of the sentences. In this paper, a practical description of the structure of the SUS test and recommended test procedures are given, in order to help researchers construct their own material and run assessment experiments.

SUS material produced for three European languages has been assessed in a number of studies with listeners whose mother tongue is the language under investigation (Hazan and Grice, 1989; Grice and Hazan, 1989; Benoît, 1989; Benoît et al., 1989, Jekosch, 1994), as well as with first- and second-language speakers of French (Benoît, 1992; Benoît and Abry, 1995). Those experiments aimed at evaluating the effects of various linguistic factors that could affect the results of an intelligibility test and have led to the procedure specifications presented below. Results are presented below of two studies in which SUS material was used in conjunction with other word- and sentence-level intelligibility material. It is hoped that further use of this test in other languages and in cross-language assessments will provide more information as to its efficiency as compared to existing sentence-length material.

2. General procedure for generating SUS

2.1. The syntactic structures

There are five simple syntactic structures. None of these exceeds eight words, in order to avoid saturation of listeners' short-term memory. The need to determine structures which are comparable across most European languages has led to the use of two methods of categorisation: sentence elements are labelled in terms of (i) functional sentence elements (e.g. **subject – verb – direct object**) which are constant across languages and (ii) syntactic categories (e.g. **determiner – adjective – noun – verb – determiner – noun**). Minor local differences in the linear ordering of these categories across languages (such as **determiner – adjective – noun** for English corresponding to **determiner – noun – adjective** in French) are unavoidable. There is no scientific reason why we selected five syntactic structures. We simply agreed on a limited set of structures which are basic in all European languages and very comparable across them in the type and in the number of rules needed to obey syntax. On the one hand, since they are particularly simple, even non-native listeners would have no difficulty deciphering the syntactic rules applied to those sentences. On the other hand, generation rules are easy to implement so that a computer can generate those sentences in many languages. Examples of sentences produced for each of the syntactic structures are given below in six languages.

2.1.1. Syntactic structures

The following word classes are used: nouns, verbs, adjectives, relative pronouns, prepositions, conjunctions, question-words and determiners.

Structure 1

(a) Subject – Verb – Adverbial: **intransitive** structure

(b) Det + Noun + Verb (intr.) + Preposition + Det + Adjective + Noun

Dutch: *De stoel loopt door een lief huis.*
English: *The table walked through the blue truth.*
French: *La robe entre vers la science rouge.*
German: *Der Bauch boxt mit dem wilden Dreck.*
Italian: *Il padre fugge dentro la voce rara.*
Swedish: *En stol dog till ett tomt hus.*

Structure 2

(a) Subject – Verb – direct Object: **transitive** structure

(b) Det + Adjective + Noun + Verb (trans) + Det + Noun

Dutch: *Een warm bot drinkt de dag.*
English: *The strong way drank the day.*
French: *Le verre vrai ouvre le coin.*
German: *Ein mürbes Blatt schlürft den Rumpf.*
Italian: *La testa sola beve il legno.*
Swedish: *En klar bok sjøng en ko.*

Structure 3

(a) Verb – direct Object: **imperative** structure

(b) Verb (trans.) + Det + Noun + Conjunction + Det + Noun

Dutch: *Eet het boek en de pen.*
English: *Draw the house and the fact.*
French: *Tourne la date ou la main.*
German: *Dränge das Garn und den Fuss.*
Italian: *Copri la zona e la nave.*
Swedish: *Se en bok och en sag.*

Structure 4

(a) Q. word – Verb – Subject – direct Object: **interrogative** structure

(b) Quest. Adv + Aux + Det + Noun + Verb (trans.) + Det + Adjective + Noun

Dutch: *Hoe eet het woord een snel glas?*
English: *How does the day love the bright word?*
French: *Quand le texte pose-t-il la fille crue?*
German: *Wann trinkt der Pelz ein grelles Kind?*
Italian: *Dove il labbro mangia il quadro lungo?*
Swedish: *Hur blev en lukt ett snabbt hus?*

Structure 5

(a) Subject – Verb – complex direct Object: **relative** structure

(b) Det + Noun + Verb (trans.) + Det + Noun + Relat. Pronoun + Verb (intr.)

Dutch: *De vloer sloot de vis die liep.*
English: *The plane closed the fish that lived.*
French: *La chose lance le train qui pense.*
German: *Das Huhn heizt den Mann, der gräbt.*
Italian: *La roba morde il sangue che parla.*
Swedish: *En plan at en fisk som sam.*

2.1.2. Lexical constraints

All constituents are selected from the most frequent words in their syntactic categories using published databases which classify words in terms of their general frequency of occurrence in written texts or spoken recordings. It was decided to reduce contextual cues further by using only ***minisyllabic*** words, i.e. those words containing the smallest number of syllables within a given category for a given language. For many languages, minisyllabic is equivalent to monosyllabic. When there are not enough monosyllabic words in a given category, words with more syllables are used. The words are meaningful and unambiguous in terms of their phonological shape and syntactic category: only one homophone is extracted from the same category (e.g. "mère" from "mer", "mère"), and a word which belongs to two word classes (e.g. "report") is stored in the lexicon in the category corresponding to its most frequent use. In both cases, the category corresponding to the most frequent use is selected.

2.1.3. Syntactic categories

There are several restrictions regarding construction of the lists of items in each syntactic category.

1. Verbs

- The verb list is subcategorised according to transitivity. Verbs which exclusively or mainly occur in one of the two sub-categories "transitive verb" and "intransitive verb" are included in these lists. Those occurring frequently in both are eliminated.
- Auxiliary (*etre*, *to be*, etc.), impersonal (*pleuvoir*, etc.) and reflexive verbs (*se laver*, etc.) are rejected.
- The imperative form for the verb is used in all languages in structure 3. In English in structures 1, 2 and 5, the simple past indicative is used. In structure 4 the infinitive form is used after the auxiliary "does". In other languages, the present indicative is usually used. Such choices depend on their phonetic consequences: factors such as number of syllables and phonetic complexity.
- In the interrogative form (structure 4), an auxiliary (e.g., *does*) is used in English after the Question word and a pronoun *-t-il* is used in French after the verb.

2. Nouns

- Only common singular nouns are included in the noun list.
- Loan words are excluded.

3. Adjectives

- The adjective list only contains adjectives which can be used **attributively** (e.g. in the position "*the* _____ *man*").
- Comparative and superlative forms and adjectives of nationality (which are high on many word-frequency lists) are rejected.
- The adjective precedes the noun in some languages (English, Dutch, German, Swedish, etc.) whereas it generally follows in others (Italian, French). In Italian and French, only the type of adjective which must be placed after the noun is included.
- Adjectives are not inflected according to gender in English and Dutch, but are in other languages.

4. Prepositions

- Only **single word** prepositions which may occur in a prepositional phrase (e.g. *to* as in "*to the house*" are included. Prepositions like "*out*", which can only be used in combination with another preposition in this type of structure, are rejected.

5. Question adverbials

- Only **question adverbials** (e.g. *why*, *when*, in English; *où*, *quand*, in French) fit into the interrogative structure, **question pronouns** (*who*, *whom*, *which*, *what*, in English; *qui*, *quoi*, *quel*, in French) do not.

6.. Determiners

- Definite articles are most often used. Exceptions are structure-specific (e.g. in the German data, articles followed by adjectives are indefinite, as in structures 2 and 4).
- There is a varying number of possibilities for the article (*the* in English; *il*, *la*, *lo*, *l'* in Italian; *der*, *die*, *das*, *den*, *dem* in German).
- Definite articles are inflected according to case in German: nominative, accusative or dative.
- Definite articles undergo elisions before a vowel in French and in Italian, leading to the suppression of a syllable.

7. Adverbs

- The adverb list contains adverbs which may be used as adverbials (e.g. adverbs of time and place such as "always") rather than verb modifiers (local adverbials e.g. "softly")

The number of monosyllabic words in a particular syntactic category differs across languages. If the number is small, necessitating the use of bisyllabic words, then this might constitute a considerable degree of redundancy. It is unclear how such redundancy might affect results. These cross-linguistic differences may introduce discrepancies in the results if one calculates the score in terms of key-words correct instead of computing the percentage of whole sentences correct. It is assumed that the overall redundancy due to syntax and morphology is much more homogeneous across languages when considering the whole sentence rather than individual key-words.

2.1.4. Local congruence

Agreement in number and gender between subject and verb and between determiners, adjectives and nouns must be strictly observed. The number issue is resolved in that only singular words are used.

2.2. Producing the sentence sets

2.2.1. Creating the lexicons

An annex at the end of this article provides bibliographical references for the published word frequency dictionaries that have been used in Dutch, English, French, German, Italian and Swedish to create semantically unpredictable sentences within the SAM project. In each language, as many lexicons must be stored as syntactic categories to be used. There are sub-categories within some of these lexicons. The number of words needed to create 60 sentences (12 for each syntactic structure = 10 for the test and 2 for the training session) are given in parentheses.

N	[Nouns]	(> 120 words)
A	[Adjectives]	(> 36 words)
T	[Transitive Verbs]	(> 48 words)
I	[Intransitive Verbs]	(> 24 words)
Q	[Question Words]	(as many as possible ... generally 3 or 4 monosyllables)
P	[Prepositions]	(as many as possible ... generally > 5 monosyllables)
C	[Conjunction]	(usually 2: *and*, *or*)
R	[Relative Pronoun]	(fixed in most languages)

All these words must obey the constraints defined in the previous paragraph.

2.2.2. Creating the sub-lexicons

From the above lexicons, sub-lexicons must be made so that each word from a lexicon will appear only in one sub-lexicon.

The number of sub-lexicons needed is:

10	for the [Nouns];	N1–N10,
3	for the [Adjectives];	A1 → A3,
4	for the [Transitive Verbs];	T1 → T4 (T2 imperative, T3 infinitive in English),
2	for the [Intransitive Verbs];	I1, I2,
1	for all the others;	Q, P, C, R.

Each sub-lexicon must contain at least as many words as sentences to be generated in each structure (except P, Q, C, R). From our experience with such tests, we recommend twelve sentences per structure so that the test in itself is not too lengthy to administer yet yields sufficient data for statistical analysis.

Certain categories may be stored together in a single lexicon for certain languages, especially when they occur together in the structures. For example, in French and Italian, it is helpful to concatenate the definite article and the noun inside the same lexicon, as the article always precedes the noun. This provides a simple solution to gender agreement and cases where a final vowel of an article is elided before a following vowel.

2.2.3. Generation of the five sets of sentences

(a) Manual generation (lexicon order is given below on the example of English)

Structure 1;	N1 + I1 + P1 + A1 + N2,
Structure 2;	A2 + N3 + T1 + N4,
Structure 3;	T2 + N5 + C1 + N6,
Structure 4;	Q1 + N7 + T3 + A3 + N8,
Structure 5;	N9 + T4 + N10 + R1 + I2.

In each sub-lexicon, the words must be selected at random, avoiding repetition of the same word, once it has been selected.

The only modifications to be done are: concordance in gender between adjective and noun (except for Dutch and English), elision (in French and Italian), a few insertions (if not previously made inside the lexicons or the sub-lexicons; like the auxiliary in the interrogative English form, or the pronoun in the French one, and the punctuation).

In the rare event of a sentence being produced which has unfortunate unacceptable connotations (e.g. "L'homme mûr met la fille"), words can be manually exchanged between one structure and another.

(b) Automatic generation

Within the SAM project a multilingual version of SUS generation software has been designed (in Dutch, English, French, German, Italian and Swedish). It may be run on any PC-workstation. The test generation software can be provided upon request [2]. It is also possible to program a rather simple tool that randomly selects words from each sub-lexicon, and then brings about gender agreement in the article, the adjective and the noun, from a given structure such as A2 + N3 + T1 + N4, for instance.

The problem of agreement may be solved by having multiple forms of each base-form adjective in sub-sections of the [Adjectives] lexicon and by tagging items in the lexicon [Nouns] for gender and case. The program may then easily select the correct form of the adjective according to its preceding (in German and Swedish) or following (in French and Italian) related noun.

3. Recommendations for SUS testing methodology in order to assess a single synthesizer

3.1. Sentence generation

As it is easier to compare test results from different studies when the test material and test procedures used are compatible, we recommend a standard test protocol based upon previous experience of running SUS tests. There are obviously no magic figures for a given test. Nevertheless, compromises between size, test duration, cost in terms of analysis and significance of results usually converge towards a set of optimal numbers. The values recommended below must be thought of in that perspective.

From our own experience, we recommend that users do the following:

- generate sets of 12 sentences per structure, i.e., $5 \times 12 = 60$ sentences,
- in each syntactic structure, randomly select 10 sentences for the actual test and the remaining 2 for training purposes,
- randomly mix the list of 10 training sentences and the list of 50 test sentences,
- record the training and test lists by the synthesizer to be evaluated,
- record the training list by a human speaker in clear condition.

The training set will thus allow subjects to get familiar with (i) the linguistically strange content of the sentences by hearing such SUS clearly uttered by a human speaker, and (ii) the acoustically strange quality of synthetic speech material.

3.2. Recording of test material

- Make a recording on high-quality audio tape or digital medium.
- Use an interstimulus time of 15 seconds and insert a short tone one second before the presentation of each sentence.
- Insert a pause every twenty sentences in order to give subjects a short break.

3.3. Subjects

We recommend that 20 subjects be selected with no experience with synthetic speech and no hearing loss. Subjects should be preferably gratified for participating in the test. They should be of a relatively homogeneous competence in their (native) language knowledge and from a fairly homogeneous age group. Trained phoneticians, or even students in speech science, should be avoided since degree of phonetic training is a factor known to greatly influence intelligibility scores.

3.4. Stimulus presentation

Sentences from the five syntactic structures should be randomly mixed so that subjects do not become aware of the syntactic constraints of the test.

[2] Contact person: C. Benoît, Institut de la Communication Parlée, Université Stendhal, BP 25X, 38040 Grenoble Cedex 9, France; e-mail: benoit@icp.grenet.fr.

It is recommended that, in order to avoid learning effects, each listener hears each sentence once only.

3.5. Training session

An initial training session should be performed so that subjects are well aware of the linguistic and the acoustic content of the tapes. We recommend that subjects first listen to the 10 extra sentences uttered in clear natural and then by those same 10 sentences generated by the synthesizer to be evaluated. In this preliminary session, written feedback should be given to the listeners.

3.6. Test session

Subjects should not be given details about the linguistic structure of the test, apart from the fact that the sentences will not be meaningful although they contain real words. They should be asked to write down what they hear on an answer sheet. The answer sheet must not give any information about the number of words in each sentence. Subjects should preferably wear high-quality headphones in a sound-treated room or cubicle, unless "real condition of usage" is the main issue in the synthesizer evaluation.

4. Recommendations in order to compare two or more synthesizers

To compare two or more synthesizers, we suggest that the same set of sentences is used to evaluate each synthesizer. We thus recommend the following:

- record the 10 training sentences by a human speaker,
- record the 10 training sentences and the 50 test sentences by *each* synthesizer,
- randomize the presentation of synthesizers to the subjects,
- avoid presenting more than 100 sentences per session (approximately one hour),
- take great care when comparing the scores obtained from such an experiment with results from a single presentation, since a training effect is expected from repetition of sentences and leads to an increase in the overall identification scores of sentences (Grice and Hazan, 1989). Improvements in mean intelligibility scores of around 15% over 3 sessions in the above mentioned English study and 30% over 5 sessions in a French study (Benoît, 1990) have been obtained. This procedure allows a fair comparison between the two (or more) synthesizers tested together. However, absolute results should only be compared to previous ones from a similar experiment where the same number of synthesizers were tested. If the objective of the test is to compare one (or more) synthesizer(s) with others previously tested, it is then recommended to run one individual test per synthesizer.

5. How to score the results

The simplest and the fastest way to score results is to only take into account the sentences that are entirely correct. All words (including articles) must be correct. All words must be in their correct position in the sentence (inversions = errors) and must belong to their original syntactic categories (syntactic mistakes = errors). Homophonic words of the same syntactic category and with the same morphological agreement (such as [*la*] *mère* and [*la*] *mer*, but not [*le*] *maire*, in French) are considered as correct.

It has been shown (Benoît, 1989, 1990, 1992) that this easy-to-obtained score is strongly related to the word score, and that it exhibits larger differences between synthesisers.

With 20 listeners, each participating in one session of 50 sentences transcription, the SUS intelligibility score is thus made of the overall percentage of correct sentences for the whole corpus (out of 20 × 50 = 1000 sentences) and for each syntactic structure (out of 20 × 10 = 200 sentences).

6. Comparison of SUS with other word- and sentence-level intelligibility tests

SUS test material was compared to other word- and sentence-level test material used for speech output system evaluation in two different studies. A German study (Jekosch, 1994) compared the use of

SUS with three word-level intelligibility tests for a single German synthesiser. An English study compared the performance of a group of 50 listeners on SUS tests and other sentence-level and word-level tests presented in two conditions: (1) synthetic speech produced by a British English synthesis-by-rule system, and (2) natural speech in noise (Hazan and Shi, 1993).

In the German study (Jekosch, 1994), the following tests were presented: (1) the SAM segmental test (SAM final report, 1992) which consists of consonants presented in VC, CV and VCV word structures; (2) the Cluster Identification test (Jekosch, 1992) in which clusters of n consonants are presented in a CnVCn structure; (3) the Modified Rhyme test (Sotscheck, 1982) which consists of real monosyllabic German words presented with a closed response set, and (4) the SUS sentence test. Different groups of untrained listeners were used for each test. Results showed very significant differences in intelligibility scores for the different test materials synthesised with the same German synthesizer. The lowest scores (19% sentences correct) were obtained for the SUS sentences. Mean scores of around 25% were obtained for single consonants in nonsense syllables (the SAM segmental test), and scores of around 65% were obtained for the consonant cluster test. Not surprisingly, the highest mean scores (79%) were obtained for the MRT test, in which meaningful words were presented in a closed-response set.

In the English study, a homogeneous group of fifty untrained subjects was tested on three types of material. These included: (1) SPIN (SPeech In Noise) sentences (Kalikow et al., 1977) which are designed to evaluate the effect of semantic information at sentence level by presenting test words in either high- or low-probability contexts, (2) SUS sentences, and (3) VCV (Vowel–Consonant–Vowel) utterances constructed using 12 consonants in three vocalic contexts. For the "synthetic speech" condition, the speech material was synthesised through the JSRU synthesis-by-rule system system (Holmes et al., 1964), connected to a parallel formant synthesiser. For the "natural speech in noise" condition, the sentence and VCV material was recorded by a male speaker. Pink noise was added to these recordings at a signal-to-noise level of 6 dB for the sentence material and 8 dB for the VCV material.

Table 1
English study (Hazan and Shi, 1993): intelligibility scores (50 listeners) obtained in the synthetic speech and natural speech in noise conditions for the SPIN test (high- and low-probability sentences), SUS sentences and VCV utterances

%	Synthetic			Natural		
	Mean	S.D.	Range	Mean	S.D.	Range
SPIN (HP)	85.8	7.1	34	87.5	5.3	28
SPIN (LP)	54.1	8.7	34	46.3	7.7	36
SUS	10.4	6.1	28	12.1	5.3	26
VCV	49.8	9.6	47	40.1	7.5	35

Intelligibility scores for the English study are given in Table 1. As in the German study, the lowest mean scores were obtained with the SUS sentences: 10.4% for the synthetic condition and 12.1% for the natural speech in noise condition. Higher intelligibility scores were obtained for the VCV material: 49.8% for the synthetic condition and 40.1% for the natural condition. A similar level of performance was reached for tests words in the low-probability SPIN sentences (54.1% for the synthetic condition and 46.3% for the speech in noise condition). The highest scores were obtained for test words in high-probability sentences: 85.8% for the synthetic condition and 87.5% in the speech-in-noise condition. Within each condition (synthetic or natural speech in noise), greater variability was generally obtained for the sentence material than for the VCV material. A principal component analysis on the results obtained suggested that performance with VCV utterances and sentences did not appear to be related.

In both studies, SUS sentences proved to be the most difficult task, with intelligibility scores of between 10 and 20%. This is probably due to the greater cognitive load involved in correctly identifying a complete string of semantically unconnected words. As predicted, the highest scores were obtained for material in which lexical and/or semantic contextual information was present: e.g., real words in a closed-set response format and words in meaningful sentences.

The large difference in scores between the SUS and SPIN sentences should be weighted by the fact that the SUS score refers to the percentage of sentences where *all* constituting words were correctly identified, whereas the SPIN score refers to the percentage of target words correctly identified (one per sentence). As shown by Benoît (1990), the relation between the proportion of correctly identified words within the SUS p_w and that of entirely correct SUS p_s is $p_s = (p_w)^r$. With French SUS material made of 6.7 words per sentence on average, a mean value of the power $r = 3.8$ has been observed with French subjects (Benoît, 1990) as well as with Ivorian subjects (Benoît, 1992), whatever the synthetic or natural speech used, i.e., whatever the level of acoustic degradation. If we apply this result to the above reported results, we may assume that percentages of correct words within the SUS would be the following: 65% in the German study; respectively 52% for synthetic speech and 57% for degraded natural speech in the English study. These (hypothetical) scores are close to those observed at the word level, e.g., with the SPIN test, although they take into account the overall intelligibility of the sentence and not only that of a single test-word in a carrier sentence. Finally, the scores observed also suggest that SUS tests would be appropriate for high-quality systems for which ceiling effects would be reached for meaningful sentences.

7. Conclusion

The various experiments that were run with the SUS test in order to evaluate the intelligibility of speech synthesizers lead us to believe that it is a highly valuable test for the assessment of text-to-speech synthesizers at the sentence level. This test is suitable for detecting even subtle differences in intelligibility, for example between different prosody modules for a given synthesiser and between two diphone-based synthesisers based on the same speaker (Benoît, 1990). However, like with any kind of 'standardized test', we recommend that great care be taken when setting up the SUS experimental procedure, since even a slight difference between two protocols might affect overall intelligibility scores. One of our objectives is that we encourage researchers in the area of text-to-speech synthesis to use a common standardized procedure in the evaluation of any new version of a system, or of a module of it. Among all the other characteristics of a speech output device, its intelligibility at the sentence level is certainly the very first one which requires perceptual assessment before the device be considered satisfactory. Therefore, it is hoped that a well defined reference test such as the SUS test will greatly help further comparisons of systems in a more objective manner.

One of the aims of this paper has been to propose a test of similar structure for different European languages. However, this does not mean that this test is necessarily suitable for cross-linguistic comparisons. As discussed above, there are many differences across the sentences depending on the language tested, such as the number of different forms taken by the definite article; the agreement in gender between the article, the noun and the adjective; or the minimum number of syllables per language. Linguistic redundancy does not affect all languages in the same manner. For example, French nouns differ in gender whereas English nouns do not. Therefore, comparing English monosyllables with French monosyllables – or even Italian bisyllables – can be rather misleading. Also, identification errors are still rather common with speech synthesizers at the segmental level in the absence of linguistic structure. Thus confusing the vowels /a/ and /œ/ in the definite articles *la* or *le* in French may induce errors in the identification of the following noun when this noun would have been correctly identified in isolation. One approach that may facilitate cross-linguistic comparisons is to measure the index of average linguistic complexity of the SUS material in each language. Initial measurements for French SUS material (Benoît, 1990, 1992) suggested that this index was quite constant across conditions and listener populations and similar to the binomial distribution one could theoretically expect from sequences of 3.8 independent units, whatever the proportion of errors on the units. This result was as expected given the redundancy one could estimate in respect of syntax, morphology and agreement in the French SUS sentences. This approach merits further investigation and the measurement of this index of perceived complexity from the error distributions reported in future SUS material in different languages would allow the experimenters to evaluate the linguistic complexity of their corpus, as perceived by listeners, with those in other languages. This may constitute an efficient protection against the many cross-linguistic differences that may bias a straight comparison of results from the SUS test across languages.

Acknowledgements

We first wish to thank and congratulate Louis Pols for his efficient and enjoyable "democratic authority" while chairing the SAM Speech-Output Group for more than five years. We are indebted to all the SAM colleagues without whom this work would not have been possible, and to Michel Cartier for friendly and scientific advice and support in the publication of this article. Finally, special thanks to Françoise Émerard who first thought of a multilingual version of the "*cadavres exquis*".

Appendix A. Word frequency dictionaries

(a) *Dutch*

Uit den Boogaard, P. (1975), *Woordfrequenties in geschreven en gespoken Nederlands*, Oosthoek, Scheltema and Holkema.

(b) *English*

Nelson, F.W. and Kucera, H. (1982), *Frequency analysis of English usage: lexicon and grammar*, Houghton Miffin.

Johansson, S. and Hofland, K. (1988), *Frequency analysis of English: vocabulary and grammar*, Oxford University Press.

Thorndike, E.L. and Lorge, I. (1968), *The teacher's word book of 30,000 words*, Teachers College Press, New York.

(c) *French*

Boë, L.J. and Tubach, J.P. (1992), *De A à Zut: Dictionnaire phonétique du français parlé*, ELLUG, Université Stendhal, Grenoble, France, 192 pp.

Catach, N. (1984), *Les listes orthographiques de base du français*, Nathan, Paris.

Coll. CNRS (1971), "Dictionnaire des fréquences, vocabulaire littéraire des XIX° et XX° siècles", In: *Trésor de la Langue Française*, CNRS, Nancy.

Gugenheim, G., Michea, R., Rivenc, P. and Sauvageot, A. (1964), *L'élaboration du français fondamental*, Didier, Paris.

Juilland, A., Brodin, D. and Davidovitch, C. (1970), *Frequency dictionary of French words*, Mouton.

(d) *German*

Ortmann, W.D. (1976), *Hochfrequente Deutsche Wortformen* I/II, Kemmler, München.

(e) *Italian*

Bortolini, Tagliavini and Zampoli (1977), *Lessico di Frequenza della Lingua Italiana Contemporanea.*

Sciarone, G.A. (1977), *Vocabolario Fondamentale della Lingua Italiana*, Minerva Italica.

(f) *Swedish*

Almqvist and Wiksell (1970), *Sture All'en "Nusvensk frekvensordbok"*, Vol. 1, Stockholm.

References

C. Benoît (1989), "Intelligibility test for the assessment of French synthesizers using Semantically Unpredictable Sentences", *Proc. ESCA Workshop on Speech Input / Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, pp. 1.7.1–1.7.4.

C. Benoît (1990), "An intelligibility test using semantically unpredictable sentences: Towards the quantification of linguistic complexity", *Speech Communication*, Vol. 9, No. 4, pp. 293–304.

C. Benoît (1992), "The intelligibility of the French spoken in France compared across listeners from France and from the Ivory Coast", *Proc. 2nd Internat. Conf. on Spoken Language Processing*, Banff, Alberta, Canada, October 1992, Vol. 2, pp. 999–1002.

C. Benoît and C. Abry (1995), "De l'impertinence, ou comment relier complexité linguistique et qualité acoustique", in *Levels in Speech Communication: Relations and Interactions*, ed. by C. Sorin et al. (Elsevier, Amsterdam), pp. 127–136.

C. Benoît, A. van Erp, V. Hazan, M. Grice and U. Jekosh (1989), "Multilingual synthesizer assessment using Semantically Unpredictable Sentences", *Proc. Eurospeech'89 Conf.*, ESCA, Paris, Vol. 2, pp. 633–636.

C. Benoît and L.C.W. Pols (1992), "On the assessment of synthetic speech", in *Talking Machines: Theories, Models and Designs*, ed. by G. Bailly and C. Benoît (Elsevier, Amsterdam), pp. 435–441.

N. Chomsky (1957), *Syntactic Structures* (Mouton, The Hague).

A.J.F. Fourcin (1992), "Assessment of synthetic speech", in *Talking Machines: Theories, Models and Designs*, ed by G. Bailly and C. Benoît (Elsevier, Amsterdam), pp. 431–434.

M. Grice and V. Hazan (1989), "The assessment of synthetic speech intelligibility using Semantically Unpredictable Sentences", *Speech, Hearing and Language: UCL Work in Progress*, Vol. 3, pp. 107–122.

V. Hazan and M. Grice (1989), "The assessment of synthetic speech intelligibility using Semantically Unpredictable Sentences", *Proc. ESCA Workshop on Speech Input/output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, pp. 1.6.1.–1.6.4.

V. Hazan and B. Shi (1993), "Individual variability in the perception of synthetic speech", *Proc. Eurospeech'93 Conf.*, ESCA, Berlin, Vol. 3, pp. 1849–1852.

J.N. Holmes, I.G. Mattingly and J.N. Shearme (1964), "Speech synthesis by rule", *Language and Speech*, Vol. 7, pp. 127–143.

U. Jekosch (1992), "The cluster-identification test", *Proc. 2nd Internat. Conf. on Spoken Language Processing*, Banff, Alberta, Canada, October 1992, Vol. 1, pp. 205–209.

U. Jekosch (1994), "Speech intelligibility testing: On the interpretation of results", *J. American Voice I/O Society*, Vol. 15, pp. 63–79.

D.N. Kalikow, K.N. Stevens and L.L. Elliott (1977), "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability", *J. Acoust. Soc. Amer.*, Vol. 61, pp. 1337–1351.

G.A. Miller and S. Isard (1963), "Some perceptual consequences of linguistic rules", *J. Verbal Learning and Verbal Behavior*, Vol. 2, pp. 217–228.

P.W. Nye and J. Gaitenby (1974), "The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences", *Haskins Lab. Stat. Rep. on Speech Research*, Vols. 37/38, pp. 169–190.

L.C.W. Pols and SAM-partners (1992), "Multi-lingual synthesis evaluation methods", *Proc. 2nd Internat. Conf. on Spoken Language Processing*, Banff, Alberta, Canada, Vol. 1, pp. 181–184.

J. Sotscheck (1982), "Ein Reimtest für Verständlichkeitsmessungen mit deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachübertragungsgüte", *Der Fernmeldeingenieur*, Vol. 36, pp. 1–84.