AUTOMATIC CREATION OF TTS INTELLIGIBILITY TESTS

Fabio Tesser[1,2], Giacomo Sommavilla[2], Giulio Paci[2], Piero Cosi[1,2]
[1]Istituto di Scienze e Tecnologie della Cognizione
Consiglio Nazionale delle Ricerche - Unità Organizzativa di Supporto di Padova, Italy
[2]MIVOQ S.R.L., Padova, Italy
fabio.tesser@pd.istc.cnr.it,giacomo.sommavilla@mivoq.it
giulio.paci@mivoq.it,piero.cosi@pd.istc.cnr.it

Abstract

This work presents a new method to automatize the creation and the analysis of subjective intelligibility tests for Text To Speech systems. In order to reach this goal, two main ideas have been adopted: the employment of multiple-choice answers and the use of algorithms able to automatically create the appropriate test set for intelligibility tasks. The reasons that led to the design of this new kind of intelligibility tests and the choices regarding the design of the algorithms are described. A practical example of the application of this new methodology has been experimented in the context of an online subjective test for intelligibility evaluation of two TTS voices.

1. INTRODUCTION

In speech communication, *intelligibility* is a measure of how *comprehensible* is speech under particular *conditions*. In the context of Text To Speech (TTS) Synthesis the goal is to evaluate how much the speech signal generated from text is *comprehensible*. This measure can give an indication about the goodness of the synthesizer, its voices and its modules. In addition, the results can help pointing out how to improve the quality of a TTS system.

The *conditions* under which intelligibility is evaluated are generally environmental acoustic conditions, and then they can be specified by the particular audio device used or the environment in which the test is carries out.

Looking to the literature, different speech intelligibility measures can be defined. It can be useful to measure the *segmental intelligibility* of a TTS system (Venkatagiri, 2003) in order to understand if the system is less intelligible in correspondence of particular phonemes; in other cases, it can be sufficient the use of the intelligibility at *word level* (Yu et al., 2010).

To simulate real life conditions, some methods add noise to the speech signal in order to measure intelligibility depending on the level of added noise (Venkatagiri, 2005). In this way it is possible to test which TTS systems or technologies are more resistant to noise. Moreover researchers started investigating on strategies to make the speech signal more intelligible if disturbed by a particular kind of noise (e.g.: babble noise) for example using particular speech enhancement techniques (Zorila et al., 2012) or by formants shifting (Godoy et al., 2013).

This problem affects also the speech generated by a TTS; as a matter of fact it was noticed that the intelligibility of a TTS system immersed in noise is lower than the natural voice in the same conditions (King & Karaiskos, 2010).

The scientific and technological communities are aware of this phenomenon, so that, in the context of a *speech-in-noise* intelligibility challenge, a dedicated task was reserved only for TTS systems (Cooke et al., 2013) and researchers in the field of speech synthesis have proposed many ideas to increase the intelligibility of TTS systems (Nicolao et al., 2013; Valentini-Botinhao et al., 2014; Erro et al., 2014).

All these examples suggest that intelligibility is a very important area for both speech technology and voice science.

Objective measures of intelligibility has been proposed, but some of them do not correlate well to subjective intelligibility scores. Moreover most of them mainly measure the audibility of a signal without taking into account of actual phonetic content of the signal (Valentini-Botinhao et al., 2011).

Although the use of an objective measure for intelligibility would enable to fully automate the procedure, because of the limitations mentioned above, currently the most reliable method to evaluate intelligibility are subjective tests.

However, organizing and performing a subjective intelligibility test is a task that needs a lot of resources. In fact traditional subjective intelligibility tests ask the subject to listen to a sentence and then to transcribe what she/he understood. Drawbacks of this method are: 1) it requires a lot of effort from the user; 2) the results are difficult to analyse due to user typos and different ways of writing the same word.

The new method presented here is a proposal to eliminate the aforementioned issues.

The paper is organized as follow: Sections 2 and 3 present the proposed method, Section 4 illustrates the web-oriented intelligibility tool designed to dispense the test, Section 5 describes the possible data analysis and Section 6 reports on an experiment of intelligibility executed with this method. Finally Section 7 concludes the paper.

2. METHOD

We designed a novel type of subjective intelligibility test, based on multiple-choice answers. The benefits of this methodology with respect to the traditional ones are: 1) less effort is required for the subject; 2) the results are easy to analyse (also in an automatic way); 3) all the process, from the data creation to the data analysis can be *automatized*.

The first two advantages are obtained thanks to the Multiple-choice answers, while the last one, the automation, is explained in Section 3.

Multiple-choice answers intelligibility tests are not a novelty; Modified Rhyme Test (MRT) (Logan et al., 1989; Goldstein, 1995) was used in the past to measure intelligibility of speech synthesis systems. However, that method was designed to present to the listener only isolated words and then it is not suitable for any kind of overall sentence evaluation, and it is focused only on consonants.

Our method does not suffer from these limitations because it considers all the phonemes and it makes use of full sentences instead of isolated words.

The procedure of the test is as follows:

a) the system presents a sentence to the subject and the user can play the sentence as many time she/he likes;

b) when the subject feels ready to give an answer, the sentence is displayed on the screen with the exception of one word called the *missing word*;

c) the subject is asked to select the missing word she/he heard, by choosing from a list of acoustically-similar words.

To make a visual example, Figure 1 shows a screen-shot of our intelligibility *web tool* taken when the subject is selecting the missing word.

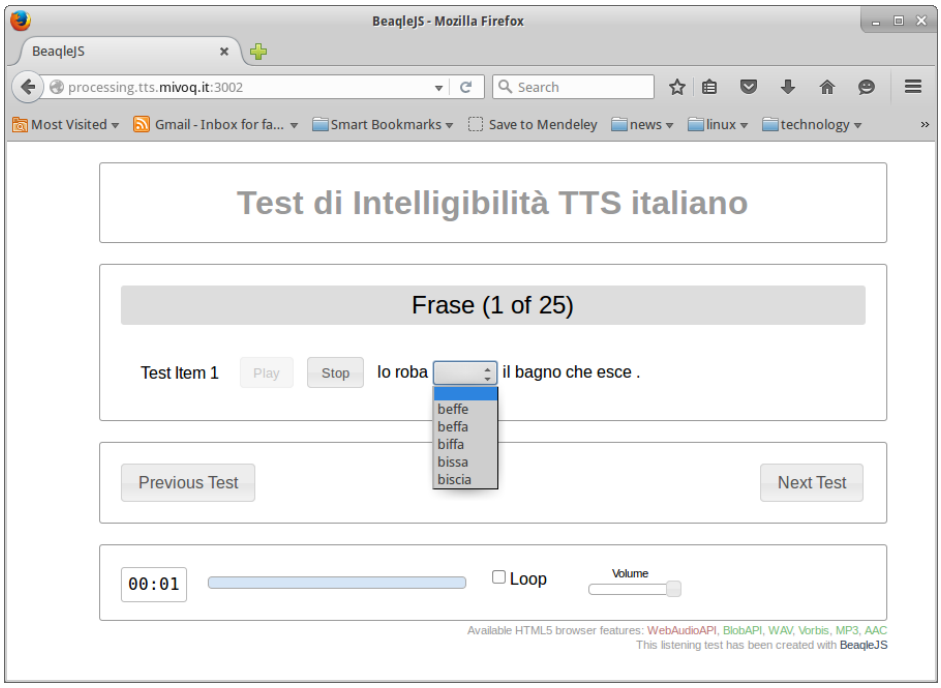

Figure 1: Intelligibility web evaluation tool: selection of the missing word.

3. AUTOMATIC CREATION OF THE TEST'S DATA SET

Automatic creation of test's data set is a very useful feature if the experiment has to be replicated for different speakers, systems or even for different languages. When designing the test, in fact, it is considered a constraint the fact of being able to test and prepare the data for other languages with little effort.

In this paper the term *automatic* means that there is a procedure able to compute the data or analyse the results for different languages and different subjects, with the only requirement of having the necessary data and linguistic modules for that language.

Going into detail, the *automatic creation* of the data set for this intelligibility test includes three main points:

1) the automatic generation of a Phonetically Balanced Word List;

2) the automatic generation of Acoustically Similar Words;

3) the automatic generation of Semantically Unpredictable Sentences.

The next three paragraphs explain the motivations and methods concerning these three points.

3.1. Automatic generation of the Phonetically Balanced Word List

In order to evaluate the intelligibility of the entire phonetic inventory of a language, of course it is necessary to prepare a test that contains sentences and words that globally contains all the sounds of that language.

Because the test is based on guessing a *missing word*, our goal is to find a list of words that contains all the phonemes defined for that language.

The first *Phonetically Balanced Word List* was developed during the Second World War (Egan, 1948) and reformulated later (Logan et al., 1989; Goldstein, 1995), precisely for the purpose of measuring intelligibility, and it was a 50 words list manually computed for the English language.

Today it is possible to automatize the computation of such a list, using algorithms able to do extensive search and computations on big textual corpora available for that language (e.g. wikipedia). We developed this procedure using the same algorithms created to select the sentences that maximize the phonetic coverage for the purpose of building a balanced TTS corpus (Pammi et al., 2005), with the difference that in our case the sentences were composed only of a single word.

The procedure to derive the Phonetically Balanced Word List is the following:

- extract the words of the considered language from wikipedia;
- use the before-mentioned algorithm to select the set W_N of the N words that maximize the phonetic coverage.

While the choice of M and N is left to the test's designers and may vary according to the aim of the test, it is critical to choose N so that the set of words includes all the phonemes of the language.

The procedure can be easily replicated for different languages, in fact the maximum phonetic coverage algorithm only needs a LTS (Letter To Sound) module and a big textual corpus available for the language taken into consideration.

3.2. Automatic generation of acoustically similar words

During the process of understanding a verbal message, it can happen of not having understood a particular word. In this circumstance it is very common to try to mentally find the missing word among the words more acoustically similar to the perceived sound. For this reason it is desirable that the words the subject has to choose among are similar from the acoustic point of view.

The task that we want to solve here is then: starting from a word w_X (taken from the set of Phonetically Balanced Word List W_N), to find the R words more acoustically similar to a word w_X looking in a dictionary D.

To do this in an automatic way, we need to compute an acoustic distance between two words. Since at this stage of the procedure we do not know the acoustic realization of the words but only their phonetic representation, we can approximate the distance with acoustic phonetics distance.

In order to compute the phonetic distance between the pronunciations of two words, we make use of the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970), an algorithm used in bio-informatics to align DNA sequences. The distance between two words is determined by the number of moves necessary to convert a phonetic sequence into the other one.

The procedure is weighted by a substitution matrix that is responsible of weighting the substitution between phonemes taking into account the similarities between them.

Each element (i, j) of this matrix represents the penalty to be introduced by the algorithm in the replacement of the phoneme i with phoneme j. Knowing the acoustic-phonetic characteristics of each phoneme (e.g.: vowel, consonant type, place of articulation, ...) of the language it is possible to automatically build this matrix for every language.

An example of a generated substitution matrix for the Italian language is shown in Figure 2. As an example, looking at the figure, it is possible to verify that the penalty for a substitution of a vowel with a consonant is high (darker colour) and, on the contrary, the diagonal elements have a penalty equal to zero (white colour).

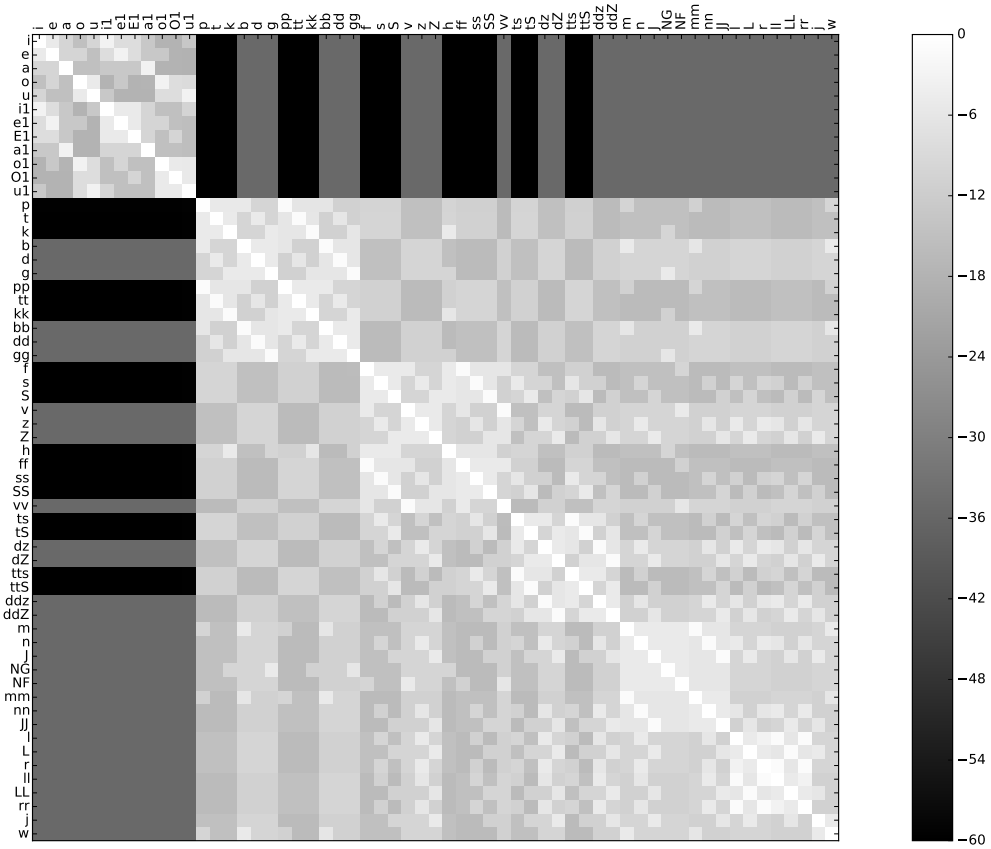

Figure 2: Example of the substitution matrix used in the Needleman-Wunsch algorithm for the Italian language. Rows and columns are indexed by Italian phonemes (coded with SAMPA). Dark-gray elements represents an high penalty, light-gray elements represent low penalty.

Also for this task, the proposed procedure can be easily ported to different languages. In fact for each language the algorithm needs: the definition of the phonetic characteristics (for the generation of the substitution matrix), the LTS (Letter To Sound) module and the dictionary D.

3.3. Automatic generation of SUS sentences

SUS (Semantically Unpredictable Sentences) sentences (Benoît et al., 1996) are generated randomly from permissible grammatical structures by a NLG (Natural Language Generation) program. Such structures are language-dependent, so in order to run an experiment for a particular language, a generative grammar must be provided.

Since they are randomly generated, most SUS sentences are semantically anomalous; also, they can have no meaning or be perfectly correct (from a semantical point of view). For this reason the user does not have prior knowledge about the semantical correctness of the sentence he is about to listen, so they are called "semantically unpredictable".

In the context of intelligibility tests, SUS sentences are used to prevent the subject to be able to "guess" the missing word thanks to the semantic context.

Furthermore, it is worth noting that since the words of the sentences are randomly picked, there can be no agreement in gender and number between nouns, modifiers and verbs. For example, consider the following SUS Italian sentence *"Lo roba beffa il bagno che esce"* that has the following literal English translation *"The stuff jokes the bathroom that exits"*. In this example, "Lo" is a masculine article, while "roba" is a feminine noun.

In order to generate the sentences for our intelligibility test, the SUS sentences generation program has been modified so that, in the random structure of the sentences, one word was replaced by a missing word. In this way we have ensured that the missing word was not always in the same position. It could be at the beginning, in the middle or at the end of the sentence.

As an example, taking the SUS sentence used before, the third word was replaced with the missing word:

Lo roba w_x *il bagno che esce.*

where w_x can be selected between a list of acoustical similarly words, e.g.:

$w_x \in$ (*beffa* | *beffe* | *biffa* | *bissa* | *biscia*)

4. WEB BASED INTELLIGIBILITY EVALUATION TOOL

Another issue related to the organization of subjective tests is recruiting of the adequate number of subjects to reach a statistically valuable result. Nowadays, thanks to the massive diffusion of Internet, it is possible to get in touch with a lot potential subjects and allowing them to perform the test online. Compared to the experiments carried out in the laboratory, the online test is deprived from the opportunity of supervising the subjects when they are performing the experiment, but it has the great advantage of being able to reach more people, leaving them the freedom to carry out the test when and where they want only using a *web browser*.

For these reasons we spent efforts on the development of a web-based evaluation tool: we implemented additional features into an open source web application for listening tests called BeaqleJS (Kraft & Zölzer, 2014).

Figure 3 shows the instructions for the experiment available when the subject begins the test. The instructions are configurable, and they should be decided depending on the particular context. An example of instructions for the subjects can be:

- listen to the sentence as many times you want;

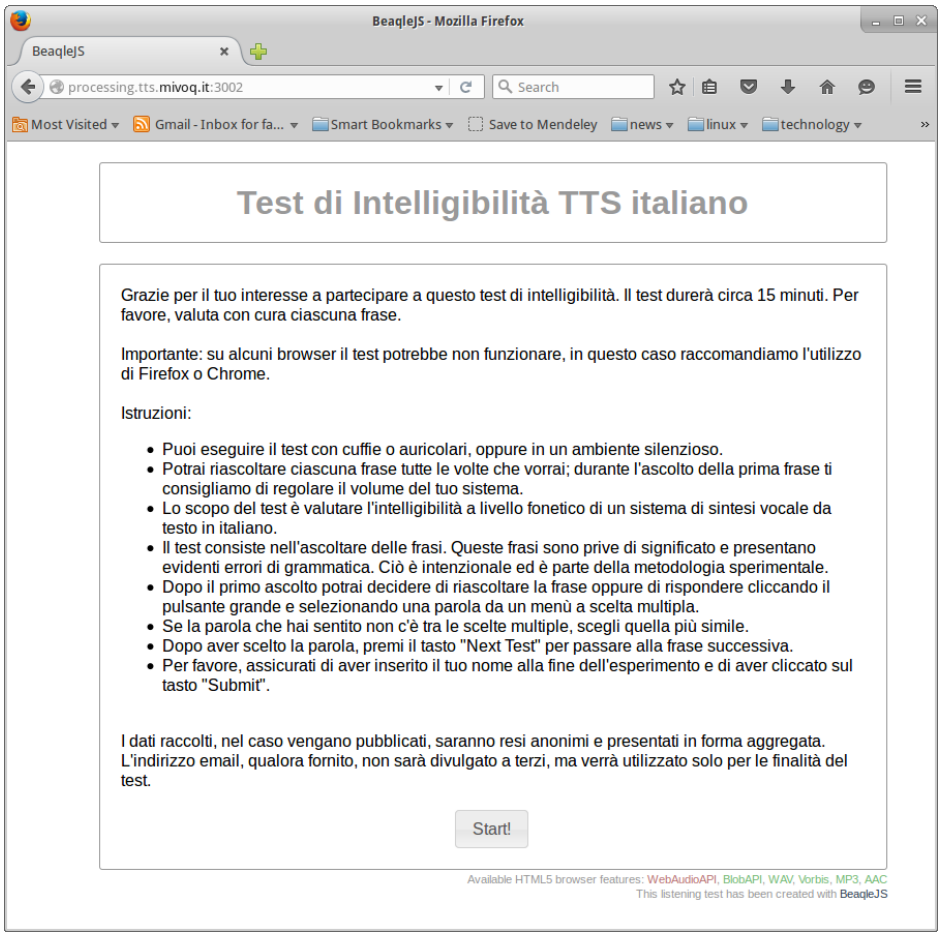

Figure 3: Intelligibility web evaluation tool: initial page (instructions).

- when you are ready to give the answer, the text of the sentence will be displayed on the screen with the exception of one word: the missing word;
- with reference to the missing word, please select the one that you have understood by choosing from a list of five words in a multiple choice menu.

Figures 1, 4 and 5 show some screen-shots of the web tool during different phases of line test.

5. DATA ANALYSIS

The web application has been designed to store the responses of each subject and also additional information such as the number of listenings (i.e., the number of clicks on the "play" button) and the time the user took to give each answer.

The data are stored in JSON format and collected by a dedicated server. It is easy to write some scripts to analyse the data and compute indexes related to intelligibility.

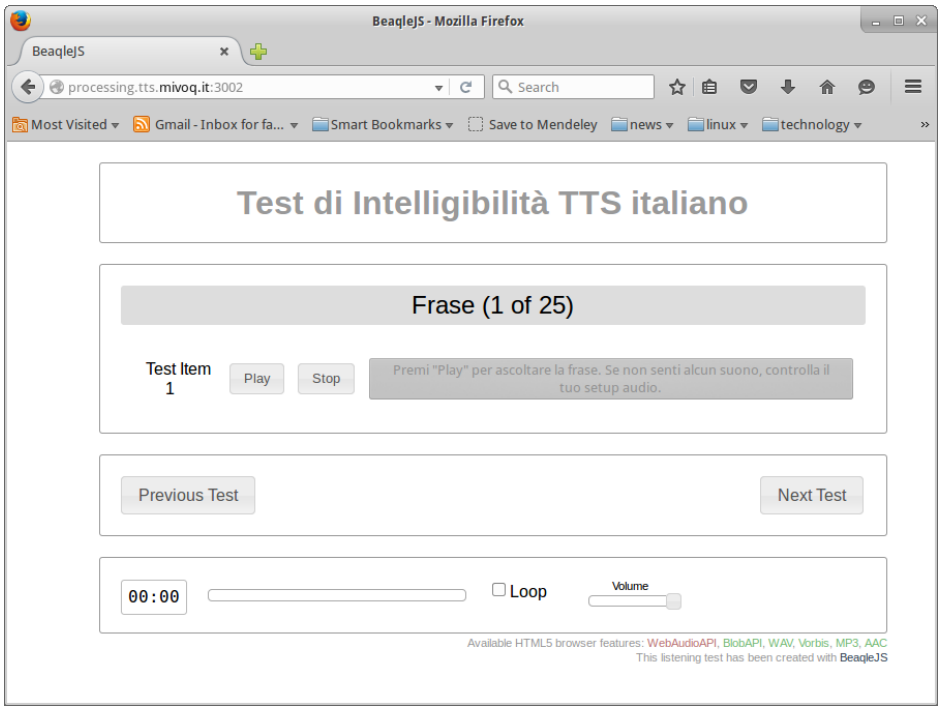

Figure 4: Intelligibility web evaluation tool: click to play.

The typical index that measures intelligibility is the Word Correct Rate (WCR), i.e. the ratio between the number of words correctly identified and the total number of words:

$$WCR = \frac{Number\,of\,Correct\,Words}{Total\,Number\,of\,Words} \tag{1}$$

Other interesting indexes that could be computed with this method are:

- number of listenings for the correct words;
- number of listenings of wrong words;
- time spent on correct words;
- time spent on wrong words.

Finally, a phonetic analysis of the errors is possible with this method. In fact, with the phonetic transcriptions of the reference word and the chosen word it is possible to analyse which phonemes have caused more misunderstandings. This is possible because when a user chooses a word different from the reference word, it is possible to see how the transcripts of the two words differ. In particular, it is possible to investigate the frequency of events such as: phoneme replaced with another, phoneme inserted or phoneme deleted.

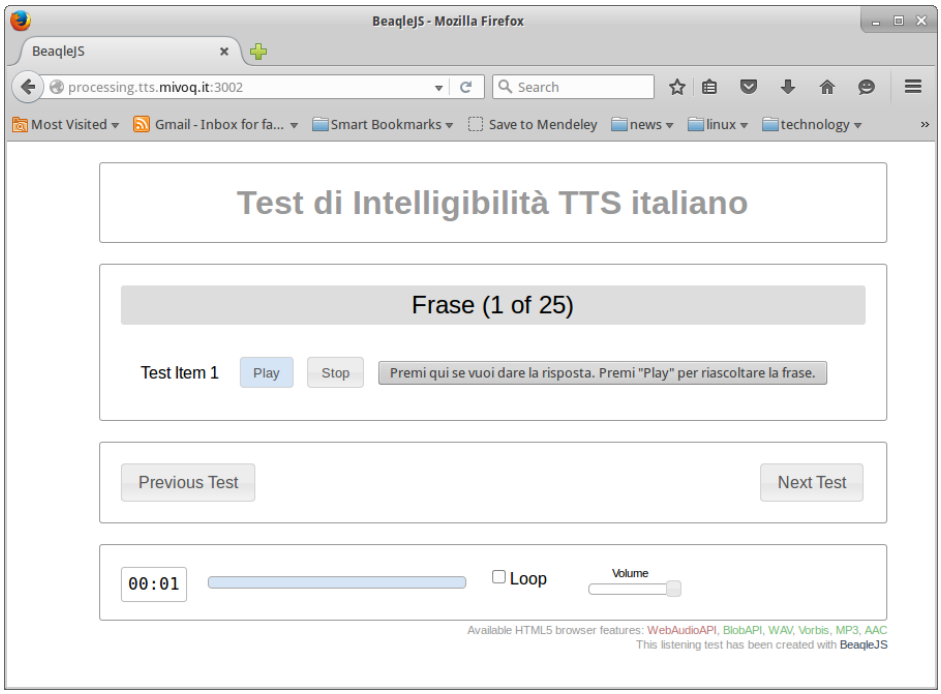

Figure 5: Intelligibility web evaluation tool: click to answer.

6. EXPERIMENT

In order to prove the whole methodology, we ran an intelligibility experiment to evaluate two Italian voices offered by FA-TTS[1], an open TTS system released in the context of the European project FI-Content 2[2]. The two voices, a female one (*istc-lucia*) and a male one (*istc-speaker_internazionale*) were built with Statistical Parametric Speech Synthesis technology (Zen et al., 2009).

Since in this experiment we evaluated Italian voices, we have sent requests to several Italian mailing lists to recruit subjects. The only required constraint for the target user was being Italian mother tongue. We did not make difference between people expert on speech technology or phonetics/phonology. The targeted end users were Italian people familiar with email and internet.

Regarding the *conditions* upon which our intelligibility test was based, we were interested in ensuring that the sentences were understandable with every device (headphones, headsets, desktop speakers, hi-fi systems); for this reason we did not give any indications about what device to use.

Also, in order to simulate real conditions, it was decided to leave the user the freedom to listen to the test where and how she/he preferred. In this way we can receive feedback from people who are running the test in front of a Desktop PC or using a smart-phone and

[1] `http://lab.mediafi.org/discover-flexibleandadaptivetexttospeech-overview.html`
[2] `http://mediafi.org/`

earphones. Everything necessary to run the test was: a device with a browser and audio output and a working network connection.

For this experiment the following design choices were carried out:

- we made use of the freely available wikipedia corpus to extract the words of the Italian language;
- starting from these, we selected the $N = 87$ words that maximize the phonetic coverage, using the procedure described in section 3.1;
- for each of these words we computed the $R = 5$ more similar words, using the algorithm described in section 3.2;
- the SUS generation algorithm, described in section 3.3, was used to generate 5 grammatically different kinds of sentences for each voice and taking into consideration the 87 missing words: a total of $87 \cdot 5 \cdot 2 = 870$ sentence was generated;

For each of these sentences it was generated the corresponding audio file using the related TTS voice. Regarding the missing word, the text passed to the synthesizer was randomly selected between the 5 acoustically similar words.

The web service for delivering this experiment and collecting the data has been running from October 9th to October 13th 2015. For each user, the test session consisted of a set of 25 randomly selected sentences among the 870 ones that have been automatically generated by the system described above. The estimated time necessary to execute the experiment was 10-15 minutes. A total number of 146 subjects participated to the experiment.

Table 1 shows the Word Correct Rate computed from the analysed data; the global WCR of the Italian voices is 87.7%, and there is not a significant difference between the female voice and the male one.

FA-TTS voices	**WCR** (%)
istc-lucia	87.2
istc-speaker_internazionale	88.1
Aggregate	**87.7**

Table 1: Word Correct Rate of the two FA-TTS voices, *istc-lucia* is the female voice, and *istc-speaker_internazionale* is the male voice.

Figure 6 shows the Word Correct Rate by the number of subjects. From this figure we can see that 12 subjects out of 146 have correctly recognized all the missing words (WCR = 100%). Most participants have reached a WCR of 92 %.

	Number
Average number of listenings for the correct words	1.89
Average number of listenings of wrong words	1.66

Table 2: Average number of listenings for the correct and wrong answers.

Other interesting figures are shown on Tables 2 and 3. Table 2 shows that the average number of listenings when the user guessed right the missing word is not so different from

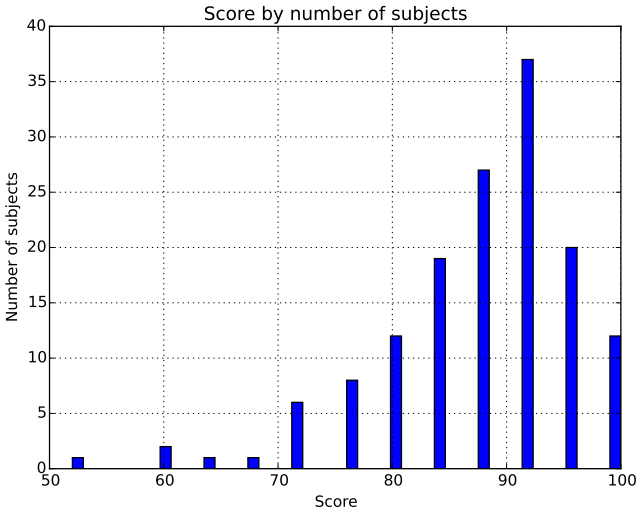

Figure 6: Word Correct Rate by number of subjects.

	Time (s)
Average time spent on correct words	17.72
Average time spent on wrong words	22.55

Table 3: Average time spent on correct and wrong answers.

the number of listenings when the the answer was wrong. On the contrary, Table 3 shows that the time spent on wrong words is, in average, almost 5 seconds longer than the time spent on correctly guessed words.

To analyse more in detail the intelligibility of the TTS system, the analysis of phonetic errors done during the experiment was carried out. The error matrix shown in Figure 7 illustrates the phoneme errors: the colour of each element of the matrix represents the frequency of the event "phoneme in the abscissa was mistaken for the one in the ordinate".

From the error matrix it is possible to notice that the most common mistake has been to misunderstand very similar vowels: /i/ and /e/ or /e/ and /a/. Another interesting and frequent error was confusing the geminates with its non-geminate counterpart. Also liquid consonants /l/ and /r/ have been misunderstood quite frequently.

7. CONCLUSION AND FUTURE WORKS

A new method to design and perform subjective tests of intelligibility has been presented and described. This method has the advantage of being easily replicable for each language as long as phonetic information, a Letter To Sound module, a SUS NLG tool and a big textual database are available for the concerned language.

Analysis of the data does not require manual effort to transcribe or correct user answers because the test is based on multiple choice answers.

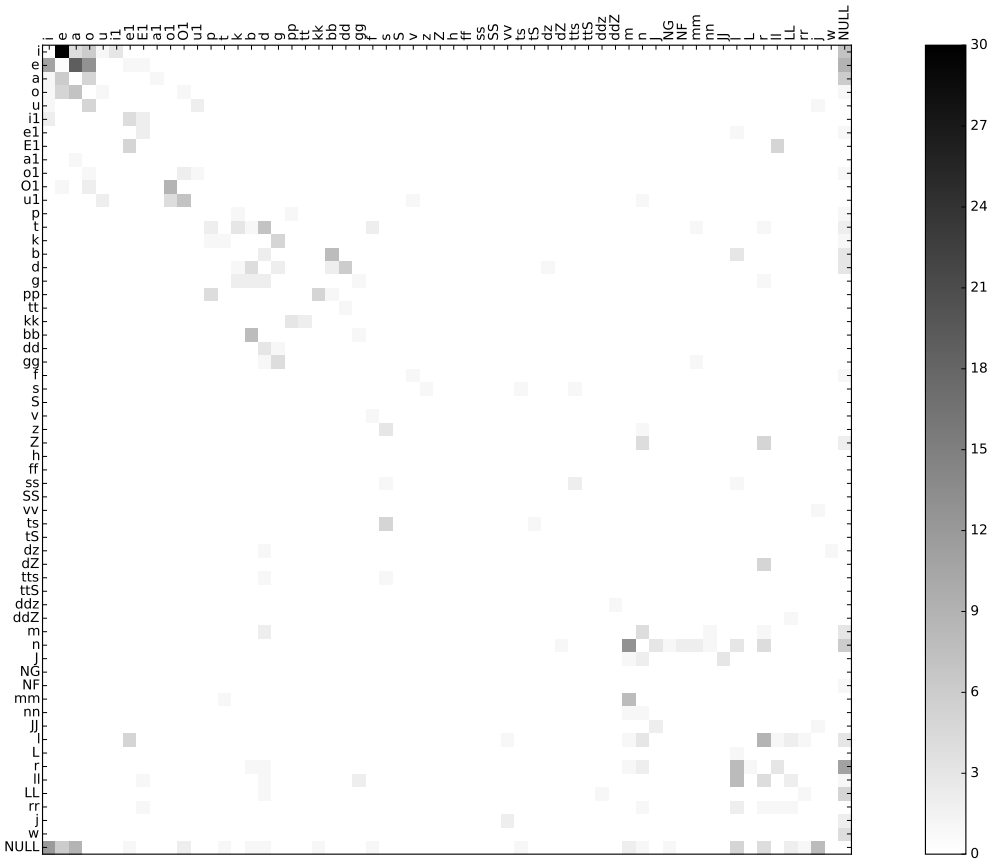

Figure 7: Error matrix of intelligibility for the two Italian FA-TTS voices: the colour of each element of the matrix represents the frequency of the event "phoneme in the abscissa was mistaken for the one in the ordinate"; white colour means zero errors, black colour means 30 errors. NULL is used to represent the case when a phoneme has been inserted or deleted.

An experiment on the Italian language to test the intelligibility of two TTS voices was designed, executed and tested with success. We aim to investigate this methodology further. The next steps to take, in order to validate deeply this new method comprises: a) the addition of sentences from a stable reference such as a real voice or a different TTS with a known intelligibility; b) adding different levels and kind noise to the speech signal, in order to test controlled speech-in-noise intelligibility; c) make a comparison with the traditional intelligibility test methodology. Finally, we are evaluating the possibility of using also *non-words*, to check if this could this improve the phonetic similarity among the words in the multiple choice menu.

8. ACKNOWLEDGMENT

The authors want to thank all the participants in the subjective test. This work was supported by the EU FP7 "FI-Content 2" project (grant number 603662).

REFERENCES

Benoît, C., Grice, M., & Hazan, V. (1996), The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences, Speech Communication, Vol. 18, no. 4, 381–392.

Cooke, M., Mayo, C., & Valentini-Botinhao, C. (2013), Intelligibility-enhancing speech modifications: the Hurricane Challenge, in Interspeech, 3552–3556.

Egan, J. P. (1948), Articulation testing methods, The Laryngoscope, Vol. 58, no. 9, 955–991.

Erro, D., Zorila, T.-C., & Stylianou, Y. (2014), Enhancing the Intelligibility of Statistically Generated Synthetic Speech by Means of Noise-Independent Modifications, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 22, no. 12, 2101–2111.

Godoy, E., Koutsogiannaki, M., & Stylianou, Y. (2013), Assessing the Intelligibility Impact of Vowel Space Expansion via Clear Speech-Inspired Frequency Warping, in Interspeech, Vol. 8, 1169–1173.

Goldstein, M. (1995), Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener, Speech Communication, Vol. 16, no. 3, 225–244.

King, S., & Karaiskos, V. (2010), The Blizzard Challenge 2010, in Proc. Blizzard Challenge Workshop, Kyoto, Japan.

Kraft, S., & Zölzer, U. (2014), BeaqleJS : HTML5 and JavaScript based Framework for the Subjective Evaluation of Audio Quality, in Linux Audio Conference (LAC-2014).

Logan, J. S., Greene, B. G., & Pisoni, D. B. (1989), Segmental intelligibility of synthetic speech produced by rule, The Journal of the Acoustical Society of America, Vol. 86, no. 2, 566.

Needleman, S. B., & Wunsch, C. D. (1970), A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology, Vol. 48, no. 3, 443–453.

Nicolao, M., Tesser, F., & Moore, R. K. (2013), A phonetic-contrast motivated adaptation to control the degree-of-articulation on Italian HMM-based synthetic voices, in 8th ISCA Workshop on Speech Synthesis, Barcelona, Spain, 107–112.

Pammi, S., Charfuelan, M., & Schröder, M. (2005), Multilingual voice creation toolkit for the MARY TTS platform, in Proc. Int. Conf. Language Resources and Evaluation, Valleta, Malta.

Valentini-Botinhao, C., Yamagishi, J., & King, S. (2011, may), Evaluation of objective measures for intelligibility prediction of HMM-based synthetic speech in noise, in 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 5112–5115.

Valentini-Botinhao, C., Yamagishi, J., King, S., & Maia, R. (2014, mar), Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion, Computer Speech & Language, Vol. 28, no. 2, 665–686.

Venkatagiri, H. S. (2003), Segmental intelligibility of four currently used text-to-speech synthesis methods, The Journal of the Acoustical Society of America, Vol. 113, no. 4, 2095.

Venkatagiri, H. S. (2005), Phoneme Intelligibility of Four Text-to-Speech Products to Non-native Speakers of English in Noise, International Journal of Speech Technology, Vol. 8, no. 4, 313–321.

Yu, Z., Yue, D., Zu, Y., & Chen, G. (2010), Word intelligibility testing and TTS system improvement, in ICASP 201, IEEE, 593–596.

Zen, H., Tokuda, K., & Black, A. W. (2009), Statistical parametric speech synthesis, Speech Communication, Vol. 51, no. 11, 1039 - 1064.

Zorila, T., Kandia, V., & Stylianou, Y. (2012), Speech-in-noise intelligibility improvement based on power recovery and dynamic range compression, in Eusipco, 2075–2079.