Improving speech intelligibility in hearing aids. Part II: Quality Assessment

A. Padilla[1], G. Piñero[1], M. de Diego[1], M. Ferrer[1], A. González[1], D. Ayllón[2], R. Gil-Pita[2], M. Rosa-Zurera[2]

[1] Audio and Signal Processing Group
Inst. of Telecommunications and Multimedia Applications (iTEAM)
Universitat Politècnica de València

[2] Applied Signal Processing Group
Signal theory and communications department
University of Alcala

Abstract

Subjective tests are the most reliable methods for quantifying the perceived speech intelligibility, but the process to perform these tests usually is time consuming and cost expensive. For this reason, different objective measures have been proposed in the literature to evaluate the intelligibility and/or quality of speech in such a way that cooperation of human listeners is not necessary.

In this paper, we describe a wide range of subjective tests reported in the literature, focusing on those proposed to evaluate speech intelligibility of Spanish language, not only for normal hearing listeners, but for hearing impaired as well. Afterwards we summarize the most common objective measures of speech quality, and finally we perform a comparison between them and some subjective speech intelligibility tests. In the subjective tests, clean Spanish speech material has been contaminated with different real background noises: cafeteria and outside traffic noise. Results show that Short-Time Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ) indices present a better correlation and a lower mean square error when predicting intelligibility compared to other objective measures tested.

Keywords: Speech intelligibility, speech quality, objective measures, subjective speech intelligibility tests, speech Spanish corpus.

1. Introduction

Speech quality is related with two aspects: the perceived overall speech quality and the speech intelligibility. Perceived overall quality is the overall impression of the listener and is related to the quality of a reproduced speech signal with respect to the amount of audible distortions [1].

On the other hand, speech intelligibility is the proportion of speech items correctly repeated by a listener or a panel of listeners, for a given speech intelligibility test [2]. The type of speech material used can be diverse, consisting of short words, syllables (with or without meaning) or sentences [3].

In this article we review the most common subjective tests and objective indices proposed to assess the speech quality and speech intelligibility, especially for Spanish language. The remainder of the paper is organized as follows: in Section 2 subjective intelligibility tests together with a review of Spanish speech material are presented, and the most common objective measures are also described. The performed subjective intelligibility tests are explained in Section 3, whereas in Section 4 the most important results are reported. Finally the main conclusions are summarized in Section 5.

2. Speech intelligibility and quality assessment

In order to assess both aspects of speech quality mentioned above, there are two principal different assessment methods that may be applied: subjective assessment, where a panel of listeners is required, and objective assessment based on physical parameters of the speech transmission system [1].

Speech intelligibility is the proportion of speech items correctly repeated by a listener or a panel of listeners, for a given speech intelligibility test.

In the subjective assessment, a panel of subjects listens to some speech material disturbed with background noise or reverberation, and write down on paper or repeat (orally/verbally) what they have heard. The speech material employed in the speech intelligibility tests consists in: monosyllables words (meaningful or nonsense words), disyllabic words, sentences or numbers [3]. The result is expressed as the percentage of correctly items heard, and is highly dependent on factors such as the type of speech material employed and the familiarity of the listeners with the text, among others.

On the other hand, objective assessments are based on physical aspects, and quantify the effect on the speech signal and the related loss of intelligibility due to disturbances, such as: limited frequency transfer, masking noises with different spectra, reverberation and echoes, nonlinear transfer resulting from peak clipping and quantization, etc.

Speech intelligibility has been used to evaluate building or room acoustics [4, 5], hearing aid performance [6], speech synthesis performance [7], and many others.

2.1 Subjective measurements

Speech audiometry is useful in order to measure the ability of a patient to perceive speech signals, which is not possible with tonal audiometry only. Speech material (i.e., a set of speech items like words or sentences) for speech audiometry has been massively developed for English language during the last half century; indeed, there are standardized tests [8, 9]. Although a similar standardization for Spanish language is not available, there are some research works in the literature for Spanish speech discrimination threshold and word discrimination: test for speech discrimination threshold [10], word lists for Speech Reception Threshold (SRT) [11] nonsense materials for speech discrimination testing [12], lists for speech discrimination testing [13], and tests for the intelligibility of speech with synthetic sentences [14]. Most of the speech material used in these works corresponds to the Spanish language spoken in the following countries: Argentina [10, 15], Spain [16, 17], Mexico [13, 18], and Chile [19].

The next section is a summary and explanation of previous research efforts in this regard.

2.1.1 Speech Spanish corpus

Tato [10] was the pioneer in the development of Spanish speech material. Tato et al., [10,15] developed twelve lists of 25 phonetically balanced[1] (PB) trochaic words (one long syllable followed by one short syllable, e.g.: mesa), five lists of 15 trochaic, disyllabic words each, and three lists of 50 monosyllabic words each, none of the last two lists were PB. The speech material was selected from newspaper articles, classic and modern novel, etc., and was tested in 5 normal-hearing listeners.

There are some criticisms made to Tato's work: Rosas [22] pointed out that there is no clear specification of the clinical use of the material; Quirós [23] and Cárdenas and Marrero [16] pointed out that written language is different from spoken language, concluding that Tato's lists are not representative of the spoken Spanish.

Cancel and Ferrer [11] developed 7 lists of 6 words each. They worked with 19 subjects from 19 Latin-American countries. For intensities from 0 to 39 dB Hearing Level (HL) in 5 dB steps, they measured performance of the lists. The carrier phrase was attenuated 5 dB below for the test word. They concluded that the word list were adequate to find the hearing thresholds for listeners from the 19 countries sampled in their research. After this speech material was recorded and employed in subjective intelligibility tests.

Ferrer [12] developed four lists of nonsense monosyllables words considering phonetic composition representative of the Spanish language, and equal phonetic composition among all lists. The material was presented to eleven native Spanish-speaking subjects at different sound pressure levels (SPL) from 60 to 20 dB SPL in steps of 10 dB SPL. Each participant listened and responded to the lists 16 items. In this study, Ferrer concluded that "the nonsense syllable lists proved to be more difficult material than the disyllabic PB lists made by Tato [10, 15]", also he considered that this material could be useful in order to distinguish between a conductive and a non-conductive hearing loss.

Berruecos and Rodriguez [13] developed four lists of 25 PB words each. The words were taken from newspapers, widely read books, songbooks, words recorded in a conversation and from the Linguaphone Method for teaching Spanish. From these materials, 954 trochaic words were selected.

Benitez and Speaks [14] worked with sentences and continuous speech instead of monosyllabic or disyllabic words, since they provide a more realistic assessment of speech understanding. Unlike traditional methods, where a listener had to repeat (or write) the word that he or she heard, in this procedure, the subject had to identify a sentence from a set of alternatives. Another difference was the use of artificial or synthetic sentences, instead of real ones.

[1] *In the phonetically balanced (PB) lists, the test words are chosen such that the relative frequency of phoneme occurrence in the entire set approximates that of the language [20, 21].*

Cancel [24] created twenty lists of 50 disyllables each. Words were selected from newspapers since they were the most common reading material, at that time. Words were of paroxytone type (with an accent on the next to last syllable of the word), because they are the most common type of disyllables in Spanish [15] and most closely approximate the English spondee words. The lists were recorded by ten Spanish-Americans students; in the recording, words were preceded by a carrier sentences in Spanish. For the subjective tests, sixty-five Spanish-American subjects listened between 8 and 20 lists in noisy and quiet environments. The three most common error-responses to 1000 test items were retained, and a multiple-choice intelligibility list was developed in order to measure speech discrimination.

Zubick et al., [25] developed nine lists of 50 disyllabic words and eight lists of 50 trisyllabic words, none of them were phonetically balanced. The material was denominated Boston College (BC) Auditory Test and was designed based on previous test designs. Criteria for the inclusion of words into the lists were as follows: most frequent stress model in Spanish, word familiarity, phonetic dissimilarity, homogeneity of basic audibility, equal average difficulty and equal range of difficulty. The words that make up these lists were taken from the Frequency Dictionary of Spanish Words [26], and were recorded by a native Spanish-speaking male and presented to ten normal-hearing native Spanish-speaking subjects. The use of these lists is only for adults.

Weisleder and Hodgson [27] pointed out some drawbacks in previous research. Firstly, although Berruecos and Rodriguez [13] reported that they compiled lists of Spanish spondee words, however there are no true spondaic words in the Spanish language. According to Tato [10], the most frequent accent model in Spanish is the paroxytone type; the predominant words type are disyllabic and tetraphonemic. Secondly, in [24, 28, 29] some of the lists were not recorded in a professional recording laboratory, and different talkers recorded different versions of the test. Finally, in [24, 29] subjective tests use only one arbitrarily predetermined presentation level. Weisleder and Hodgson [27] assessed the commercially available word recognition lists from Auditec of St. Louis. The speech material was evaluated in terms of inter-list equivalence, word difficulty, intelligibility of the talker, and slope of the performance/intensity (PI) function. Four lists were tested in 16 native Spanish-speaking subjects, whose countries of origin and number of subjects per country were: Mexico, 9; Panama, 2; Venezuela, 2; Spain, 1; Honduras, 1; Colombia, 1. Subjects listened to the four lists at four different presentation levels: 8, 16, 24 and 32 dB HL. Their results show that at the highest presentation level (32 dB HL) the best scores were obtained, and the talker's speech intelligibility was also judged to be very clear by all subjects at that level. Mean intelligibility scores were poorest for list three at almost all presentation levels, and its intelligibility was significantly different from the other lists.

Castañeda et al., [18] developed four lists of 50 disyllables and four of 50 nonsense monosyllables. The words were taken from radio and television interviews. They analysed the percentage of occurrence of the phonemes in the Spanish language spoken in Mexico and made a review of the phonetic analysis between different published lists: Tato [10,15], Berruecos [13] and Weisleder [27]. Authors also provided a detailed comparison between their lists and other published lists. Their results showed that the phonetic balance of speech material was very similar to Tato and Berruecos's lists, despite both the difference in each methodology and the dates on which the different studies were developed.

Cárdenas and Marrero [16, 17] created two lists of 24 polysyllable words each in order to assess the STR, another twenty lists of 25 disyllables for word discrimination test, and two lists of 58 words each designated "Test de rasgos distintivos" equivalent to the Diagnostic Rhyme Test (DRT) in English language, in which listeners were shown a word pair, and then asked to identify which word is presented by the talker. They also developed speech material for children between 6 and 12 years old, all these lists are still employed in clinical practice.

Sommerhoff and Rosas [19] developed a corpus of 1000 logatoms[2], grouped in 10 lists of 100 words each. Nevertheless, it has been shown that monosyllables' tests give lower intelligibility scores [10, 12, 24].

Other research activities are especially addressed to users of hearing aid devices [30, 31, 32, 33, 34]; nevertheless, the tests proposed and their results are quite similar to those presented above.

2.2 Objective measurements

Since subjective tests for speech quality evaluation are usually time consuming and cost expensive, many researchers have developed objectives measures where the cooperation of human listeners is not needed. Generally speaking, objective measurements are calculated from the comparison between a distorted speech signal and the corresponding clean speech signal using some mathematical formula or algorithm. Although good estimators of subjective quality have been developed, there are still situations where all estimations fail, thus the need to find robust and reliable methods of evaluating the perceived speech quality. This section describes the most commonly used objective quality measures.

Articulation Index (AI)

This index was proposed by French and Steinberg [35] and is based on the idea that intelligibility can be calculated by the sum of the individual contributions extracted

[2] *A logatom is a nonsense monosyllabic word with a CVC (consonant-vowel-consonant) structure.*

from the frequency decomposition of the speech signal into twenty bands, having the frequency limits between 250 and 7000 Hz (see Table III of [35]). Articulation Index is obtained by calculating the Signal- to-Noise Ratios (SNR) for each band, and averaging them. The values of the articulation index range from 0 (no intelligibility) to 1 (perfect intelligibility). This index launched a fruitful research on the development and application of objective measures for predicting speech intelligibility in different transmission systems [36].

Speech Intelligibility Index (SII)

The SII can be described as an updated and expanded version of AI [37]. Some parameters that have been updated include: spread of masking, standard speech spectrum level, and relative importance of the individual bands [38].

Speech Transmission Index (STI)

STI was developed in the early 1970's and is a widely accepted objective measure that can estimate speech intelligibility for a broad range of environments (e.g. reverberant environment) [39, 40, 41]. In order to carry out the measurement, an artificial speech-like input signal is used, which is a spectral-shaped noise that has a long-term spectrum envelope identical to speech. Speech can be regarded as an amplitude-modulated signal, where the modulation contains useful information. After transmission over the channel under test, noise and/or reverberation can be added; the extent of modulation in the signal will be affected. The loss in the modulation is calculated in seven octave bands, centred at 125 Hz to 8 kHz, each modulated by 14 frequencies at 1/3-octave intervals ranging from 0.63 Hz up to 12.5 Hz. The depth of modulation of the speech signal is compared with the output signal in a full set of frequencies (7 carrier frequencies and 14 modulations frequencies). Finally, a weighted averaged is calculated and a single value is obtained, varying from 0 (completely unintelligible) to 1 (perfect intelligibility). Figure 1 shows a simplified block diagram of the STI measurement.

Some researchers [40, 42] have established a qualitative intelligibility scale and its relationships to objective index and intelligibility percentages. Relationship between the STI and different types of subjective intelligibility tests (monosyllabic words, short phrases, PB word lists, numbers, etc.) are depicted in Figure 2.

RApid Speech Transmission Index (RASTI)

In order to reduce the measurement time of STI, other parameter was developed as a simpler alternative, called RASTI. In contrast to STI, RASTI measures only the output of two octave bands centred at 500 Hz and 2 kHz, and four and five modulation frequencies respectively. It uses

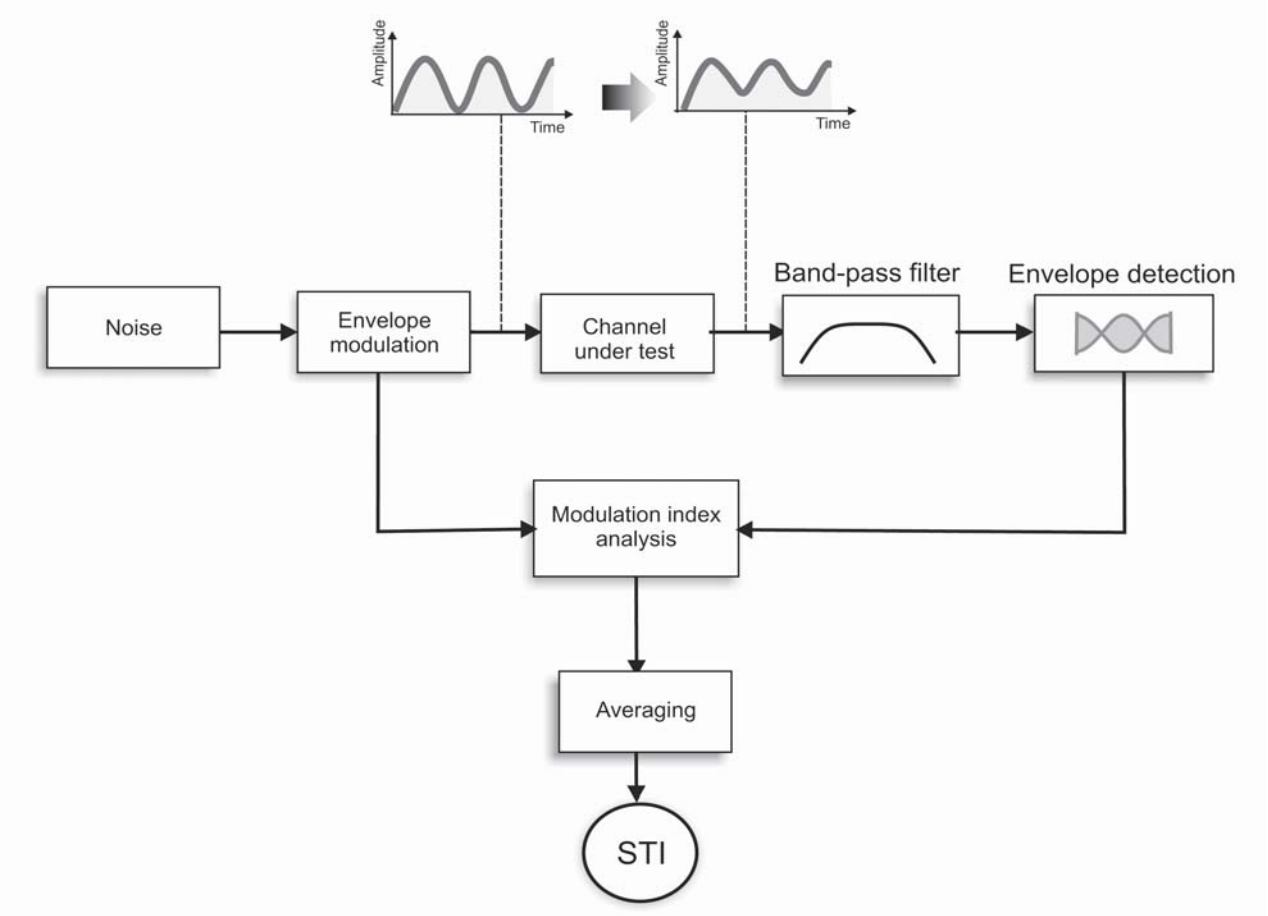

Figure 1. Simplified block diagram measurement of the STI measuring setup from [39].

a speech-like excitation signal and correlates reductions in modulation depth to loss of intelligibility [40].

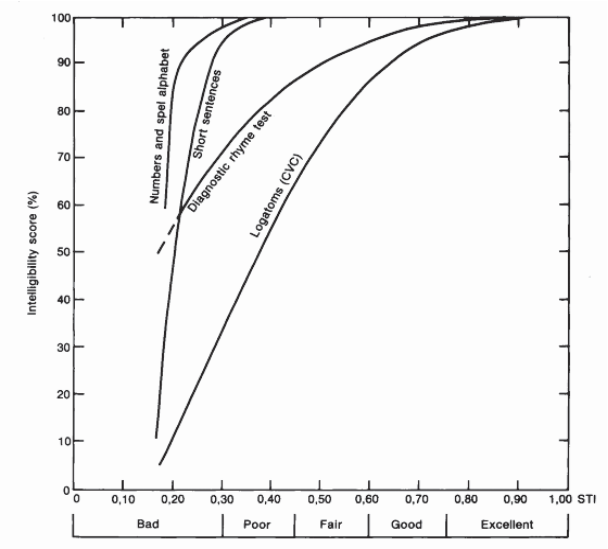

Figure 2. Relationship between the STI and different types of subjective intelligibility tests (monosyllabic words, short phrases, numbers, etc.) from [42].

Signal-to-Noise Ratio (SNR)

An objective measure widely used in order to assess speech quality is SNR. From the computational point of view it is very easy to calculate, but requires the distorted and corresponding undistorted (clean) speech samples. Assuming discrete signals of length N, the SNR calculates the ratio between the energy of the clean signal $x(n)$ and the distorted signal $y(n)$, n is the sample index, as follows [1]:

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_{n=1}^{N}x^2(n)}{\sum_{n=1}^{N}\{x(n)-y(n)\}^2} \quad (1)$$

The SNR measure is highly dependent on the time-alignment between the clean and degraded speech signals. For that reason, several variations to the traditional SNR exist, showing much higher correlation with subjective quality. Indeed, in [43, 44] researchers demonstrated that SNR measurement is a very poor predictor of speech quality.

Segmental SNR (segSNR)

One main drawback of averaging the SNR over the entire signal is that sections where the speech energy is small and the noise level is high may bury sections where the speech energy is large and that of the noise is low. Thus, an alternative to solve this problem is calculating the SNR over short frames and then average it; this measure is called segmental SNR, and is defined as:

$$\mathrm{SNR}_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\sum_{i=Lm}^{Lm+L-1}\left(\frac{x^2(n)}{(x(n)-y(n))^2}\right) \quad (2)$$

There are two main methods to assess speech quality: subjective assessment, where a panel of listeners is required, and objective assessment.

where $x(n)$ is the clean signal, $y(n)$ is the distorted signal, L is the frame length in number of samples, and M is the number of frames in the signal. The length of segments is typically 15 to 20 ms. The segSNR is also reported to be a poor predictor of speech quality [45, 46].

Frequency-weighed SNR (fwSNRseg)

Another variation to the SNR is the frequency-weighed SNR (fwSNRseg). This is essentially a weighted SNRseg within frequency bands proportional to the critical band. The fwSNRseg is defined as follows [1, 46]:

$$fwSNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=0}^{K-1}W(j,m)\log_{10}\frac{X(j,m)^2}{(X(j,m)-Y(j,m))^2}}{\sum_{j=0}^{K-1}W(j,m)} \quad (3)$$

where $W(j,m)$ [47] is the weight on the jth subband in the mth frame, K is the number of subbands, $X(j,m)$ is the spectrum magnitude of the jth subband in the mth frame, and $Y(j,m)$ is the distorted spectrum magnitude.

Weighted-Slope Spectral Distance (WSS)

The WSS distance is a direct spectral distance measure [1, 46]. It is based on the comparison of the smoothed spectra from the clean and distorted speech samples. The smoothed spectra can be obtained from either LP analysis, Cepstrum filtering or filter bank analysis. One implementation of WSS can be defined as follows,

$$d_{WSS} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K-1}W(j,m)(S_x(j,m)-S_y(j,m))^2}{\sum_{j=1}^{K-1}W(j,m)} \quad (4)$$

where K is the number of bands, M is the total number of frames and $S_x(j,m)$ and $S_y(j,m)$ are the spectral slopes of the jth band in the mth frame from clean and distorted speech, respectively. Spectra slope $S_x(j,m)$ is defined as the difference between (j+1)th band and jth band energies. $M(j,m)$ is the weight applied to the corresponding band and frame [47].

Linear Prediction Based Measures

The speech production process can be modelled efficiently by a linear prediction (LP) model. There are a number of objective measures that use the distance between two sets of linear prediction coefficients (LPC) calculated on the clean and the distorted speech respectively. Only three of them are mentioned.

- **Log-Likelihood Ratio (LLR): It is calculated as:**

$$d_{LLR}(\boldsymbol{a}_d,\boldsymbol{a}_c) = \log\left(\frac{\boldsymbol{a}_d\boldsymbol{R}_c\boldsymbol{a}_d^T}{\boldsymbol{a}_c\boldsymbol{R}_c\boldsymbol{a}_c^T}\right) \quad (5)$$

where $\boldsymbol{a}_c$ and $\boldsymbol{a}_d$ are the LPC vectors for the clean and distorted speech, respectively. $\boldsymbol{a}^T$ is the transpose of $\boldsymbol{a}$, and $\boldsymbol{R}_c$ is the autocorrelation matrix of the clean signal.

- **Itakura-Saito (IS) distance: It is given by:**

$$d_{IS}(a_d, a_c) = \left[\frac{\sigma_c^2}{\sigma_d^2}\right]\left[\frac{a_d R_c a_d^T}{a_c R_c a_c^T}\right] + \log\left(\frac{\sigma_c^2}{\sigma_d^2}\right) - 1 \quad (6)$$

where σ_c^2 and σ_d^2 are the all-pole gains extracted from the LPC analysis for the clean and degraded speech respectively.

- **Cepstrum Distance (CD)**

CD is an estimate of the log spectral distance between clean and distorted speech. Cepstrum is calculated by taking the logarithm of the spectrum and transforming back to the time-domain. Cepstrum can also be calculated from LPC parameters using the following expression [46]:

$$c(m) = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c(k) a_{m-k} \qquad 1 \le m \le p \quad (7)$$

where p is the order of the LPC analysis. Cesptral Distance can be calculated as follows [1, 46]:

$$d_{CEP}(c_c, c_d) = \frac{10}{\log 10}\sqrt{2\sum_{k=1}^{P}\{c_c(k) - c_d(k)\}^2} \quad (8)$$

where $\boldsymbol{c}_c$ and $\boldsymbol{c}_d$ are the Cepstrum vectors for clean and distorted speech respectively, and P is the order.

Perceptual Evaluation of Speech Quality (PESQ)

The PESQ index is an international standard measure to evaluate the speech quality of handset telephony and narrowband speech codecs [48]. The PESQ algorithm compares a reference signal with a degraded signal that is the result of passing the reference signal through the system under test. The output of PESQ is considered a prediction of the perceived quality that would be obtained by the degraded signal in a subjective listening test. Several works report high correlation between PESQ and subjective listening tests [46, 49, 50], which demonstrates that the PESQ score is also a good indicator of speech intelligibility. PESQ is regarded as one of the most sophisticated and accurate estimation methods available today.

Short-Time Objective Intelligibility (STOI)

The STOI index is a method recently developed by Taal et al. [51, 52]. The model decomposes signals into time-frequency sections, followed by energy clipping and normalisation. Intelligibility predictions are based on mean cross-correlations between processed and clean signals across time-frequency regions. STOI is designed for a sample rate of 10 kHz in order to capture a relevant frequency range for speech intelligibility, although the method can be easily extended to other sample rates. Some researchers have demonstrated that STOI shows better correlation with speech intelligibility compared to other reference objective intelligibility models [51, 52, 53].

Hearing-Aid Speech Quality Index (HASQI)

The procedures described above are intended for normal-hearing listeners. Nevertheless there is a recently proposed index specifically developed for hearing-impaired listeners: the Hearing-Aid Speech Quality Index (HASQI) by Kates and Arehart [54].

HASQI predicts the quality of a speech processed through a simulated hearing aid, while considering a wide variety of distortions commonly found in these devices. Furthermore, HASQI is based on a cochlear model that incorporates elements of impaired hearing. A new version of the originally proposed HASQI is available in [55]. HASQI was compared with other indices such as PESQ, segSNR, fwSNR and IS [56]. The results show that a trained version of HASQI predicts speech quality quite well and achieves performance comparable to PESQ and other commonly used measures; however these results are validated only for normal-hearing listeners.

3. Subjective intelligibility tests

In the following we describe the main settings of the experiment carried out at the Institute of Telecommunications and Multimedia Applications of Universitat Politècnica de València (UPV). We have performed several subjective tests with a panel of human listeners in order to obtain a speech intelligibility measure and compare it to objective quality indices.

3.1 Participants

The panel consisted of eight subjects (6 males, 2 females) of ages from 21 to 35. All of them were Spanish native speakers and all of them reported to present a normal hearing. None of the participants were familiar with the lists of words used in the study.

3.2 Speech material

The speech material consisted of eight different lists of 25 meaningful disyllabic words in Spanish. All lists were phonetically balanced (See Appendix). The material was taken from [16, 17] (from list 5 to list 12) and is commercially available in a CD. All the speech material included in the CD was recorded by a professional announcer, native Spanish-speaking female, at a professional recording studio. The speech stimuli were recorded at 44.1 kHz sampling rate.

Since the speech material was designed for audiometry tests, the CD presents the speech signal only at the right channel, while a masking noise signal is emitted by the left channel [57]. For our experiment, the original speech material of the CD was processed to remove the masking noise signal and present the speech signal on both channels.

Furthermore, speech material was contaminated with two different background noises: cafeteria and outside traffic street noise. The recordings were taken from background noise database[3] [58], where files are in wav format, have a length of 30 seconds and a sampling rate of 48 kHz. The noise signals were downsampled to 44.1 kHz in order to add them to the clean speech stimuli and obtain noisy speech signals. For each type of noise, cafeteria and outside traffic, four different signals were generated, keeping the speech energy within a comfortable auditory level, and varying the noise level to cover a wide range of the intelligibility percentage. For this purpose, some preliminary tests were carried out to different subjects that were discarded afterwards.

3.3 Procedure

Intelligibility tests were carried out inside the listening room available at the Laboratory of Signal Processing for Audio and Communications[4] of the Institute of Telecommunications and Multimedia Applications of UPV. The subject listened via headphones (Sennheiser eH 250) one of the lists from the Appendix, contaminated with a particular noise at a particular level. Once a word was presented, it was followed by a silence to allow the subject to repeat in loud voice the word that he or she had just listened. The subject's responses were recorded for a following checking step. The test took typically about 15 minutes for each subject.

Speech intelligibility score was calculated for every subject and every list by multiplying by four the number of words correctly repeated, in order to obtain a percentage (25 correct words over 25 corresponds to a 100% speech intelligibility).

Finally different objective measures presented in section 2.2 were also computed: PESQ, segSNR, fwsegSNR, WSS, LLR, CEP and STOI. Most of the objective measures' algorithms were implemented in MATLAB by Hu and Loizou [46][5], whereas STOI algorithm was implemented by Taal et al. [51][6]. Both noisy and clean speech files employed in the objective measures were previously downsampled to 16 kHz in order to capture the relevant range of the speech.

4. Results

Due to the large variations of the scales amongst the objective scores studied, the first result in Fig. 3 shows the Pearson's correlation coefficient (see eq. (1) of [46]) between the objective measure and the subjective score. The Pearson's correlation coefficient, denoted by r, measures the linear dependence between subjective speech intelligibility and the corresponding objective index. As a second result, the mean squared error (MSE) was calculated as the difference between the real subjective score and the predicted scores obtained by a least-squares linear fitting to the objective values. The MSE values are plotted in Fig.4.

Results showed that STOI and PESQ indices present better correlation and lower MSE compared to the rest of the objective measures, thus confirming their ability to predict speech intelligibility for a variety of ambient noises.

In order to determine how significant the Pearson's correlation coefficient is, the p-value was calculated for all indices and a 95% significance level was considered. The p-value for all indices is shown in the Fig. 3.

Fig.3 shows that the STOI measure yields the highest correlation with the subjective score (r = 0.84), followed by the PESQ index (r = 0.73) and the segSNR measure (r = 0.68). The lowest correlation (r = 0.05) was obtained for the SNRfwseg measure. According to the results shown in Fig. 4, the STOI index also yields the smallest MSE (MSE = 5.25), followed by the PESQ (MSE = 8.32). The highest MSE corresponds to the SNRfwseg measure (MSE = 18.33).

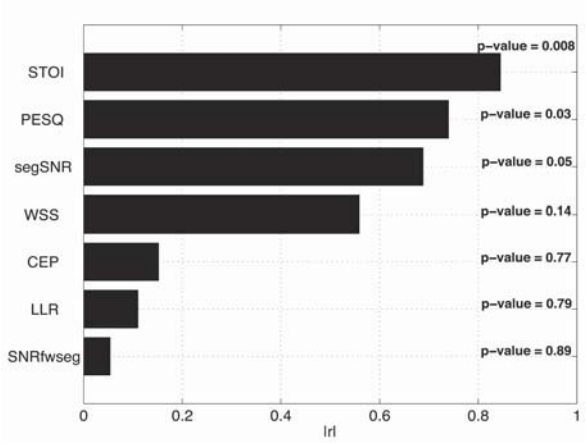

Figure 3. Absolute value of the correlation between objective and subjective scores plotted for all noise conditions. The p-value is shown for all indices.

Fig. 5 plots the mean intelligibility score achieved with each of the noisy signals in the subjective tests versus STOI and re-scaled PESQ values. STOI scores range between 0 (completely unintelligibility) and 1 (perfect intelligibility), whereas PESQ ranges between 1 and 4.5.

[3] *Available online: http://docbox.etsi.org/stq/Open/EG%20202%20396-1%20Background%20noise%20database/*

[4] *http://www.iteam.upv.es/group/gtac.html*

[5] *Available on line: http://ecs.utdallas.edu/loizou/speech/software.htm*

[6] *Available on line: http://siplab.tudelft.nl/*

Therefore, in order to compare PESQ and STOI values together, a re-scaled PESQ has been calculated as rPESQ=(PESQ -1)/3.5, resulting in a new range from 0 to 1. Regarding Fig. 5, the black and blue lines are the resulting linear and exponential curve fit to the data, respectively. The exponential curve fit was modelled by the following expression:

$$Y = 100 * \{1 - \exp[-\alpha * (X - \beta)]\} \quad (9)$$

Where α and β are the fitting parameters.
The MSE values were now calculated for the exponential curve fit, showing a relevant improvement for STOI index (MSE=1.06), and a slight decrease for PESQ index (MSE=7.99). It has to be noticed that the segSNR value was accordingly rescaled to the STOI range as $segSNR$=($segSNR$+10)/45, since it had also obtained good performance for both correlation and MSE measures. However, the rescaled $segSNR$ covered a tiny range from 0.042 to 0.052, which means that segSNR values cannot reliably describe the intelligibility scores.

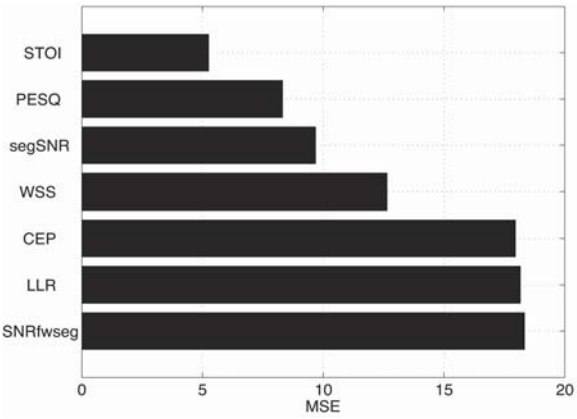

Figure 4. Mean Square Error (MSE) between objective measures and subjective measure for all noise conditions.

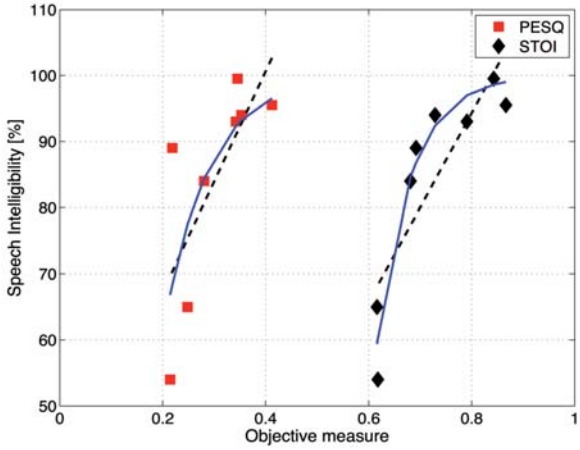

Figure 5. Subjective speech intelligibility score versus (rescaled) PESQ and STOI indices (markers), best-fit first-order polynomial (dashed black line), and exponential curve fit (blue line).

5. Conclusions

A subjective test has been run to assess the intelligibility of Spanish speech contaminated with two common ambient noises such as cafeteria and traffic noises. Test scores have been compared to the most common objective measures proposed in the literature to predict the speech quality perceived by humans. Results showed that STOI and PESQ indices presented better correlation and lower MSE compared to the rest of the objective measures, thus confirming the ability of STOI and PESQ to predict speech intelligibility for languages different from English and for a variety of ambient noises.

6. Acknowledgements

This work has been supported by European Union ERDF and Spanish Government through TEC2012-38142-C04 project, and Generalitat Valenciana through PROMETEOII/2014/003 project. Participation of author A. Padilla has been supported by a postdoctoral fellowship from Conacyt (Mexico). The authors wish to acknowledge Prof. Felipe Orduña for his insightful comments that contributed to improve the manuscript, and to everyone who participated in the listening tests.

Appendix

Table 1. Lists of words used in the subjective speech intelligibility tests.

List 5	List 6	List 7	List 8	List 9	List 10	List 11	List 12
día	noche	alzar	moza	leyes	dice	eres	muela
uvas	montón	leyes	veo	ese	alzar	tiempo	fuego
tiempo	tiempo	hacha	lado	cine	techo	tiño	tela
tiño	cada	ese	osa	conde	hotel	frío	reza
tima	coche	fuente	usen	una	coger	melón	limón
pista	saca	pintor	orden	madre	mimas	cena	este
pierna	fleco	mesa	lengua	saco	medios	raíz	ajo
venas	sartén	justa	fresa	papel	duque	tengo	tierno
regla	perros	hijas	copias	padre	pegues	oso	quema
nunca	mantel	cinco	callos	tiendas	ida	crema	huerto
lloras	hierba	brisa	gaita	hábil	renta	seca	doble
mudo	curas	torres	riña	actor	viñas	tambor	caro
creo	bajo	nubes	bedel	pecho	sola	plata	pierna
cebra	tía	terca	tecla	anchos	paso	haya	días
anda	llaves	borde	pleno	santa	gente	dame	abre
seas	cientos	sueño	mote	fundes	crean	calle	cunas
leche	vuelas	pila	laven	lejos	basta	limas	bichos
amén	ruegas	mero	finos	filo	hielos	esas	sueño
velo	pelas	humo	cine	cierta	vienen	chisme	primas
refrán	luces	dejo	arme	amor	unos	yodo	higo
nidos	guapa	choca	verdad	tío	sello	sudar	dedos
ligo	crema	bondad	puerta	guías	paran	pedal	campo
gases	cedo	tiendo	fiesta	urna	litro	culpa	nieves
corren	anís	lunes	cobre	cuatro	fuera	besa	llenos
cartel	tardes	alga	techo	rubios	clase	kilo	hasta

References

[1] K. Kondo, Subjective quality measurement of speech, Springer, 2012.

[2] T. Brand, "Speech Intelligibility", in Handbook of signal processing in Acoustics, vol. I, New York, Springer, 2008.

[3] I. Hirsh, H. Davis, S.R. Silverman, E.G. Reynolds, E. Eldert, R.W. Benson, "Development of materials for speech audiometry", Journal of Speech and Hearing Disorders, vol. 17, pp. 321-337, 1952.

[4] T. Horrall, T. Jacobsen, "RASTI Measurements: Demonstration of different applications", Brüel & Kjaer, Application Note, BO 0123-11.

[5] J. Anderson, T. Jacobsen, "RASTI Measurements in St. Paul's Cathedral, London", Brüel & Kjaer, Application Note, BO 0116-11, 1985.

[6] J.M. Kates, K.H. Arehart, "Coherence and the speech intelligibility index", Journal of the Acoustical Society of America, vol. 117, no. 4, pp. 2224-2237, 2005.

[7] A. Kain, M.W. Macon, "Spectral voice conversion for text-to-speech synthesis". In IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 285-288, 1998.

[8] ISO/TR 4870:1991 Acoustics - The construction and calibration of speech intelligibility tests.

[9] ANSI/ASA S3.2-2009 Method for Measuring the Intelligibility of Speech over Communication System.

[10] J.M. Tato (1949). Lecciones de audiometría (El ateneo, Buenos Aires).

[11] C.A. Cancel, O. Ferrer, "Pruebas auditivas para pueblos de habla Española", Otolaryngologica, vol. 3, pp. 40-74, 1952.

[12] O. Ferrer, "Speech audiometry: A discrimination test for Spanish language". Laryngoscope, vol. 70, pp. 1541-1551, 1960.

[13] P. Berruecos, J.L. Rodriguez, "Determination of the phonetic percent in the Spanish language spoken in México City, and formation of P.B. lists of trochaic words", International Journal of Audiology, vol. 6, no. 2, pp. 211-216, 1967.

[14] L. Benitez, C. Speaks, "A test of speech intelligibility in the Spanish language", International Journal of Audiology, vol. 7, no. 1, pp. 16-22, 1968.

[15] J.M. Tato, F. Lorento, J.A. Bello, "Características acústicas de nuestro idioma". Revista Otolaringologica, Vol. I, 1948.

[16] M.R. Cárdenas, V. Marrero, "Cuaderno de Logoaudiometría", Madrid, Universidad Nacional de Educación a Distancia (UNED), 1994.

[17] E. Salesa, E. Perelló, A. Bonavisa, Tratado de audiología. Elsevier Masson, 2nd. Edition, 2013.

[18] R. Castañeda, S.J. Pérez, "Análisis fonético de las listas de palabras de uso más extendido en logoaudiometría", Anales de la Sociedad Mexicana de Otorrinolaringología. No. 1, pp. 23-30, 1991.

[19] J. Sommerhoff, C. Rosas, "Logatom corpus for the assessment of the intelligibility in Spanish speaking environments and its relation with STI measurements", Applied Acoustics, vol. 73, pp. 1190-1200, 2012.

[20] M. Goldstein, "Classification of methods used for assessment of text-to-speech systems according to the demands placed on the listener, Speech Communication, vol. 16, pp. 225-244, 1995.

[21] J. Logan, B. Greene, D. Pisoni, "Segmental intelligibility of synthetic speech produced by rule", Journal of the Acoustical Society of America, vol. 86, no. 2, pp. 566-581, 1989.

[22] B.M. Rosas, "Speech audiometry in English, Portuguese and Spanish. Independent Studies and Capstones", paper 394. Program in Audiology and Communications Sciences, Washington University School of Medicine, 1958.

[23] J.B. Quirós, "Aspectos formales de la presentación de las listas logoaudiométricas", Revista Otolaringológica, vol. 9, no. 3, pp. 304-308. 1969.

[24] C.A. Cancel, "Multiple-choice intelligibility lists for Spanish speech audiometry", International Audiology, vol. 4, no. 2, pp. 91-93, 1965.

[25] H. H. Zubick, L.M. Irizarry, L. Rosen, P. Feudo, J. H. Kelly, M. Strome, "Development of speech-audiometric materials for native Spanish-speaking adults", Audiology, vol. 22, pp. 88-102, 1983.

[26] A. Juilland, E. Chang-Rodríguez, (1964). Frequency dictionary of Spanish words, The Hague: Mouton.

[27] P. Weisleder, W.R. Hodgson, "Evaluation of four Spanish word-recognition ability lists". Ear and Hearing, vol. 10, pp. 387-392, 1989.

[28] P.A. Connery, "Spanish language word lists for speech discrimination assessment", Hearing Aid Journal, vol. 6, pp. 13-41, 1977.

[29] O.E. Tosi, "Estudio experimental sobre la inteligibilidad de un test de múltiple elección en idioma español", Fonoaud, vol. 15, pp. 28-35, 1969.

[30] S. Mastroianni, L. Aronson, S. Arauz, "Batería de Habilidad Auditiva (BATHA): programa de selección y rehabilitación para pacientes con implante coclear". Otolaringológica, vol. XV, pp. 13-56, 1988.

[31] H.M. Furmanski, MC. Flandin, MI. Howlin, ML. Sterin, S. Yebra, "P.I.P. Pruebas de identificación de palabras", Fonoaudiológica, vol. 43, no. 2, 1997.

[32] H.M. Furmanski, C. Berneker, MA. Levato, M. Oderigo, "PIP-S Prueba de identificación de palabras a través de suprasegmentos", Fonoaudiológica, vol. 45, no 2, pp. 14-24, 1999.

[33] H.M. Furmanski, S. Yebra (2003), "PIP-V Pruebas de identificación de palabras a través de vocales", Fonoaudiológica, vol. 49, no. 2, pp. 54-57, 2003.

[34] L. Aronson, D. Milone, C. Martínez, P. Estienne, D. Tomassi, H.L. Rufiner, M.E. Torres, "Batería para la evaluación del reconocimiento del habla en pacientes con prótesis auditiva", Revista FASO, No. 1, pp. 17-24, 2007.

[35] N.R. French, J.C. Steinberg, "Factors governing the intelligibility of speech sounds", Journal of the Acoustical Society of America, vol. 19, pp. 90-119, 1947.

[36] K.D. Kryter, "Methods for the calculation and use of the articulation index", Journal of the Acoustical Society of America, vol. 34, pp. 1689 – 1697, 1962.

[37] ANSI S3.5-1969: Method for the calculation of the Articulation Index. New York, 1969.

[38] ANSI S3.5-1997: Method for the calculation of the Speech Intelligibility Index. New York, 1997.

[39] H.J.M. Steeneken, T. Houtgast, "A physical method for measuring speech-transmission quality" Journal of the Acoustical Society of America, vol. 67, pp. 318 – 326, pp. 1980.

[40] T. Houtgast, HJM. Steeneken, "A multi-language evaluation of the RASTI-Method for estimating speech intelligibility in auditoria", Acustica, vol. 54, pp. 185 – 199, 1984.
[41] T. Houtgast, HJM. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria", Journal of the Acoustical Society of America, vol. 77, pp. 1069 – 1077, 1985a.
[42] T. Houtgast, HJM. Steeneken, "Technical Review No. 3, The Modulation Transfer Function in room acoustics", Marlborough, MA: Brüel & Kjaer Instruments, pp. 1-44. (1985b)
[43] B.J. McDermott, C. Scaglia, D.J. Goodman, "Perceptual and objective evaluation of speech processed by adaptive differential PCM", IEEE ICASSP, Tulsa, pp. 581-585, Apr. 1978.
[44] J.M. Tribolet, P. Noll, B.J. McDermott, R.E. Crochiere, "A study of complexity and quality of speech waveforms coders", IEEE ICASSP, Tulsa, pp.586-590, Apr. 1978.
[45] Y. Hu, P.C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms", Speech communication, vol. 49, no. 7, pp. 588–601, 2007.
[46] Y. Hu, P.C. Loizou, "Evaluation of objective quality measures for speech enhancement", IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, pp. 229–238, 2008.
[47] D. Klatt, "Prediction of perceived phonetic distance from critical band spectra", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 7, pp. 1278-1281, 1982.
[48] ITU-T Recommendation P. 862 (2001): Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs.
[49] A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra. "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs". In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 2, pp. 749-752, 2001.
[50] J. Ma, Y. Hu, P.C. Loizou. "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions", Journal of the Acoustical Society of America, vol. 125, no. 5, 2009.
[51] C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech", IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7 pp. 2125–2136, 2011.
[52] C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech", in Proc. ICASSP, pp. 4214-4217, 2010.
[53] J. Jensen, C.H. Taal, "Speech intelligibility prediction based on mutual information", IEEE Transactions on Audio, Speech, and Language Processing, vol. 22, no. 2 pp. 430–440, 2014.
[54] J.M. Kates, K.H. Arehart, "The hearing-aid speech quality index (HASQI)", Journal Audio Engineering Society, vol. 58, no. 5, pp.363-381, 2010.
[55] J.M. Kates, K.H. Arehart, "The hearing-aid speech quality index (HASQI) v2", Journal Audio Engineering Society, vol. 58, no. 5, pp.363-381, 2014.
[56] A. Kressner, D. Anderson, C. Rozell, "Evaluating the generalization of the Hearing Aid Speech Quality Index (HASQI)", IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 2 pp. 407–415, 2013.
[57] ISO 8253-3:2012 Acoustics – Audiometric test methods – Part 3: Speech audiometry.
[58] ETSI EG 202 396-1 V1.2.2 (2008-09). Speech processing, transmission and quality aspects (STQ); Speech quality performance in the presence of background noise; Part 1: Background noise simulation technique and background noise database.

Biographies

Ana Padilla has a BSc in Communications and Electronics Engineering (Instituto Politécnico Nacional, México, 2003), as well as MSc and PhD degrees in Electrical Engineering (UNAM, México, 2007 and 2012). She spent three Student Internships (2007, 2008, 2010) at the Research Laboratories of Intel Corp. in Guadalajara, México. Her research deals with subjective and objective methods for measuring speech intelligibility, currently focusing on binaural speech intelligibility, with a broader interest on binaural sound technologies and acoustic instrumentation in general. Her work has been presented and published in full-text proceedings of meetings and symposia organized by the Institute of Noise Control Engineering-USA (Noise-Con 2007), Acoustical Society of America, Federación Iberoamericana de Acústica (ASA-FIA 2010), and Sociedad Mexicana de Instrumentación, (SOMI 2007, 2008, 2009).

Gema Piñero was born in Madrid, Spain, in 1965. She received the Ms. in Telecommunication Engineering from the Universidad Politécnica de Madrid in 1990, and the Ph.D. degree from the Universidad Politecnica de Valencia in 1997, where she is currently working as an Associate Professor in digital signal processing. She has been involved in different research projects including array signal processing, active noise control, psychoacoustics and wireless communications in the Audio and Communications Signal Processing (GTAC) group of the Institute of Telecommunications and Multimedia Applications (iTEAM) of Valencia. She has leaded several projects on sound quality evaluation for the toy industry, and she has also been involved in several projects on 3G communications supported by the Spanish Government and big industries as Telefonica. She has published more than 70 contributions in journals and conferences. Her current research interests include wireless acoustic sensor networks, interference management in communications, and array and distributed signal processing.

Maria de Diego was born in Valencia, Spain, in 1970. She received the Telecommunication Engineering degree from the Universidad Politecnica de Valencia (UPV) in 1994, and the Ph.D degree from the same University in 2003. Her dissertation was on active noise conformation of enclosed acoustic fields. She is currently working as Associate Professor in digital signal processing and communications. Dr. de Diego has been involved in different research projects including active noise control, fast adaptive filtering algorithms, sound quality evaluation, and 3-D sound reproduction, in the Institute of Telecommunications and Multimedia Applications (iTEAM) of Valencia. She has published more than 40 papers in journals and conferences about signal processing and applied acoustics. Her current research interests include multichannel signal processing and sound quality improvement.

Miguel Ferrer was born in Almería, Spain. He received the Ingeniero de Telecomunicacion degree from the Universidad Politécnica de Valencia (UPV) in 2000, and the Ph.D degree from the same University in 2008. In 2000, he spent six months at the Institute of applied research of automobile in Tarragona (Spain) where he was involved in research on Active noise control applied to interior noise cars and subjective evaluation by means of psychoacoustics study. In 2001 he began to work in the Audio and Communications Signal Processing Research Group (GTAC) of the Institute of Telecommunications and Multimedia Applications. He is currently working as assistant professor in digital signal processing in the Communications Department of UPV. His area of interests includes efficient adaptive algorithms and digital audio processing.

Alberto Gonzalez was born in Valencia, Spain, in 1968. He received the Ingeniero de Telecomunicacion degree from the Universidad Politecnica de Catalonia, Spain in 1992, and the Ph.D degree from de Universidad Politecnica de Valencia (UPV), Spain in 1997. His dissertation was on adaptive filtering for active control applications. From January 1995, he visited the Institute of Sound and Vibration Research, University of Southampton, UK, where he was involved in research on digital signal processing for active control. He is currently heading the Audio and Communications Signal Processing Research Group (www.gtac.upv.es) of the Institute of Telecommunications and Multimedia Applications (iTEAM). Dr. Gonzalez serves as Professor in digital signal processing and communications at UPV where he heads the College of Telecommunication Engineering (www.etsit.upv.es) since July 2012. He has published more than 80 papers in journals and conferences on signal processing and applied acoustics. His current research interests include fast adaptive filtering algorithms and multichannel signal processing for communications, 3D sound reproduction and MIMO wireless systems.

David Ayllón was born in Valladolid, Spain, in 1984. He received his BSc. in Telecommunication Engineering at the University of Valladolid (Spain) in 2006, his MSc. in Information and Communications Technologies by University of Alcalá (Madrid) in 2009, his MSc in Biomedical Engineering by University of Boras (Sweden) in 2009, and his Ph.D in Information and Communications Technologies by University of Alcala (Madrid) in 2013. His dissertation was on speech enhancement algorithms for audiological applications. His current research is concerned with speech and biomedical signal processing in distributed sensor networks.

Roberto Gil-Pita received the M.Eng. degree in telecommunication engineering and the Ph.D. degree (with hons.) in electrical engineering from the University of Alcalá, Madrid, Spain, in 2001 and 2006, respectively. From 2001, he has worked at the Signal Theory and Communications Department in the University of Alcalá, in the Applied Signal Processing research group. His research interests include pattern recognition and audio signal processing, focusing on sound source separation, hearing aids, and emotional speech. In these fields, he is author of more than 20 journal papers included in the Journal Citation Report, and around 70 conference papers. He is also project manager of several projects with public and private fundings, including the 2-year ATREC project for the real-time analysis of combat stress, funded by the spanish Ministry of Defense.

Manuel Rosa Zurera received the B.Eng. degree (with hons.) in technical telecommunication engineering from the University of Alcalá, Madrid, Spain, in 1990, the M.Eng. degree in telecommunication engineering from the Technical University of Madrid, Spain, in 1995, and the Ph.D. degree (with hons.) from the University of Alcalá, Madrid, Spain, in 1998. From 1991 to 1997, he worked as a Researcher and Lecturer at the Circuits and Systems Engineering Department in the Technical University of Madrid. Since 1997, he has worked at the Signal Theory and Communications Department in the University of Alcalá, where he has been head of the department from 2004 to 2010. He is Full Professor at the same department and Dean of the Polytechnic School of the University of Alcalá since 2010. His research interests include statistical signal processing, signal models, source coding, speech and audio signal processing, and radar systems, areas in which he has been involved in many research projects, and has published more than 50 papers in international journals.