D2.2 - Objective quality criteria for vocal announces and for alarms

Project arconym:	I‘CityForAll
Project name:	Age Sensitive ICT Systems for Intelligible City For All
Strategic Objective:	Socio-acusis ICT solutions for a better social well-being of Elderly People
Project number:	**AAL 2011-4-056**
Project Duration:	July, 1st 2012 – Jun, 30th 2015 (36 months)
Co-ordinator:	CEA : Commissariat à l’Energie Atomique
Partners:	UPD :Université Paris Descartes ENEA : Italian National Agency for New Technologies, Energy, an Sustainable Economic Development TUM : Technische Universität München CRF : Centro Richerche FIAT CENTICH : Centre d’Expertise National des Technologies de l’Information et de la Communication pour l’autonomie Active Audio, EPFL : Ecole Polytechnique de Lausanne – Lab. D’Electromagnétisme et d’Acoustique

D2.2

Version:	2.00
Date:	7-26-2013
Author:	Nader MECHERGUI Tifanie BOUCHARA Gael MAHE Sonia LARBI Meriem JAIDANE

Dissemination status: PU

This project is co-funded by the Ambient Assisted Living (AAL) Joint programme, by the German BMBF, by the l’Agence Nationale de la Recherche - ANR, by the Ministero dell‟Istruzione dell„Università e della Ricerca – MIUR, and by Federal Office for Professional Education and Technology OPET

Once completed please e-mail to WP leader with a copy to
sylvie.ghalila@cea.fr

D2.2	Executive Summary
This report concerns objective criteria that measure: global quality, intelligibility, saliency and sharpness of sounds. First, these algorithms are resumed in identification sheets that help to identify the best algorithm for I'City For All project. We focus on existing criteria of global quality and intelligibility which can be adapted for public environments and presbycusis persons. The selected algorithms are based on different models. Some of them are based on auditory perception, like PESQ which predicts effectively global quality. These types of algorithms are more likely to be fit for presbycusis problems. We selected also algorithms that are based on acoustic parameters which predict intelligibility regarding to reverberation and loudspeaker effects due to the target environment of I'City For All. We also propose new criteria to measure intelligibility: the sharpness index inspired from image processing. This is a new measure of audio clarity that can be adapted for presbycusis problems thanks to its range of sensibility regarding noise and reverberation. Besides, we suggest to measure auditory saliency to predict the attractive power and the ease of recognition of vocal announces and car alarms for our target audience. Keywords: intelligibility, global quality, saliency, sharpness, clarity.	

Dissemination Level of this deliverable (*Source: I'CityForAll Technical Annex p20 & 22***)**	
PU	Public
Nature of this deliverable (*Source: I'CityForAll Technical Annex p20 & 22***)**	
R	Report.

Due date of deliverable	2013/06/30
Actual submission date	2013/07/26
Evidence of delivery	

Authorisation			
No.	**Action**	**Company/Name**	**Date**
1	Prepared	CEA-Linklab/UPD	2013/07/10
2	Revision	CEA-Linklab/UPD	2013/07/26
3	Approved		
4	Released		

Disclaimer: The information in this document is subject to change without notice. Company or product names mentioned in this document may be trademarks or registered trademarks of their respective companies.

General introduction

The objective of this report is to give an overview of the existing objective speech quality and intelligibility assessment algorithms. We focus however on the assessment methods **open to be fit to the different requirements of the I'City for All project**, namely the assessment of:

- intelligibility and clarity of vocal announces for all
- global listening quality and comfort for all
- saliency of vocal announcements and jingles for all

This report is organized in two parts. The first part gives **identification sheets** for each assessment algorithm, each of them being detailed in the second part of the report.

The second part is structured in two categories of assessment criteria:

- Classical and standardized quality and intelligibility assessment algorithms: the selected assessment methods presented in this part are based on perceptual aspects and frequency octave/bark band analysis, thus allowing for fitting to take into account presbycusis.
- Recently proposed for **AAL I'City for All project** audio saliency and sharpness assessment methods: both inspired from image processing and adapted here to measure the saliency and the sharpness of sounds.

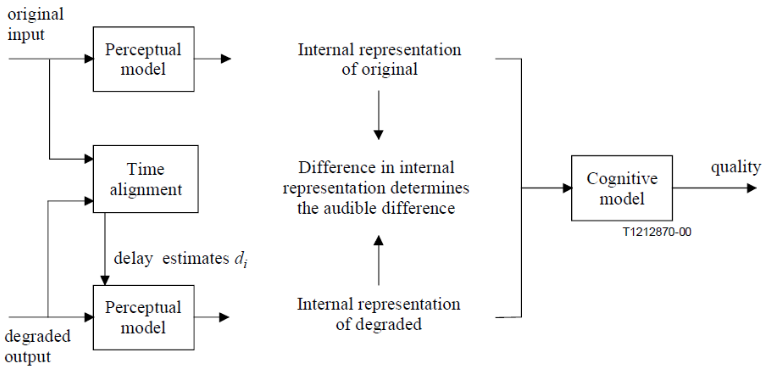
PART I

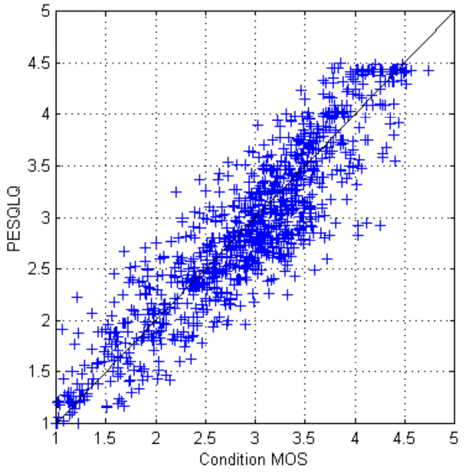
Identification sheets

of objective assessment algorithms and criteria

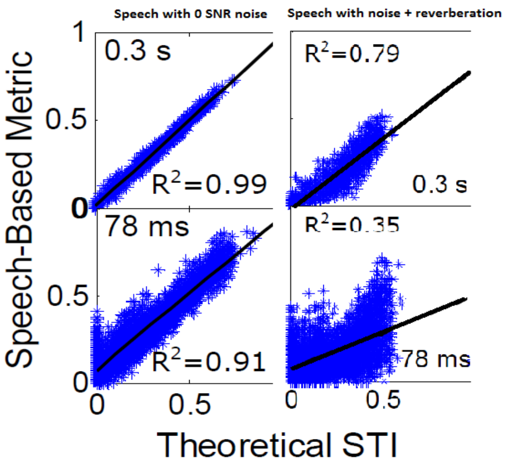
Measure of	Quality
Name	**Frequency weighted variant signal to noise ratio (fwvarSNR)**
Applications	Used to test noise reduction algorithm [9].
Model	The frequency weighted variant signal to noise ratio is computed in the frequency domain and expressed as follows: $$\boldsymbol{fwvarSNR} = \boldsymbol{W_0} + \sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{W_j} \left[\frac{\boldsymbol{1}}{\boldsymbol{M}} \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \boldsymbol{10\ \log_{10}} \left[\boldsymbol{F^2(m,j)} \Big/ \boldsymbol{(F(m,j) - \widehat{F}(m,j)^2)} \right] \right]$$ Where $\boldsymbol{M}$ is the number of frames, $\boldsymbol{K}$ the number of filters in the filter bank, $\boldsymbol{W_j}$ the weight of the j th frequency band, $\boldsymbol{F(m,j)}$ the amplitude of the jth frequency band of clean signal, $\boldsymbol{\widehat{F}(m,j)}$ the amplitude of jth frequency band of noisy signal.
Inputs	Intrusive measure that take as input : ➔ clean and degraded speech
Outputs	This criterion gives an SNR measure in dB.
Complexity	Computed frame by frame, the **fwSNR** can be running in real time with parallel processing for each band. An overall SNR can be computed for each sentence.
Known limits	The criterion was not tested for long reverberation time and it doesn’t takes into account non-linear distortion.
Known performances and conditions of evaluation	In [9], the **fwSNR** was tested with subjective overall quality measure. With 25 frequency bands, the fwSNR predicted score reaches up to 0.8 correlation with subjective score and an error standard deviation of 0.36.
Interests for the I’City For All Project	The **fwSNR** is simple to use and can be adapted for different subjective measure and population.

Possible adaptation for presbycusis	The **fwSNR** can be adapted for presbacusis with two methods : - Adapt the frequency band weight in order to correlate the **fwSNR** with presbycusis subjective measure. - Consider the hearing loss of presbycusis as an internal noise for each frequency band. This noise can be added as an SNR for each band depending on the audiogram of the subject.
Observations	This measure is based on the original frequency weighted SNR [5]. Algorithm available in Matlab.
References	[5] [6] [9]

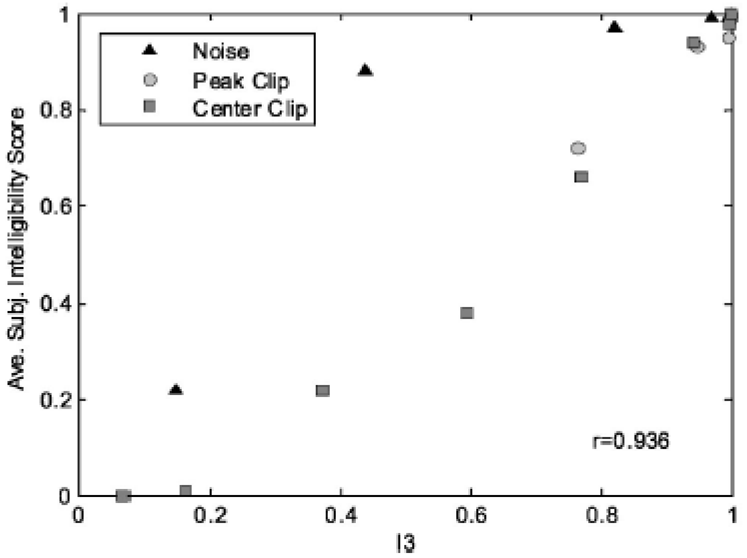
Measure of	Global quality (Mean Opinion Score)
Name	**Perceptual Evaluation of Speech Quality (PESQ)**
Applications	Used to test speech quality in telephony system. PESQ takes into account, noise, codec degradation, packet loss…
Model	**PESQ** is a standard objective algorithm that is based on psychoacoustic model. Perceptual internal signal representation is computed using auditory model following the steps illustrated below :  original input Perceptual model Internal representation of original Time alignment Difference in internal representation determines the audible difference delay estimates d_i Perceptual model Internal representation of degraded degraded output Cognitive model quality T1212870-00 **Descriptive schema of PESQ model**
Inputs	Intrusive measure that take as input : ➔ clean and degraded speech
Outputs	Mean Opinion Score (MOS) between 1 and 4.5. 1 for bad quality and 4.5 for excellent quality.
Complexity	Even if PESQ is based on frame by frame processing, we cannot run it in real time because 'time alignment' requires the entire sentence to find the best alignment between degraded and clean signals.
Known limits	The criterion was not tested for room reverberation degradation and public address systems.
Known performances and conditions of evaluation	For 22 known ITU benchmark experiments, the average correlation was 0.935. The figure below gives mapping between subjective score and PESQ score.

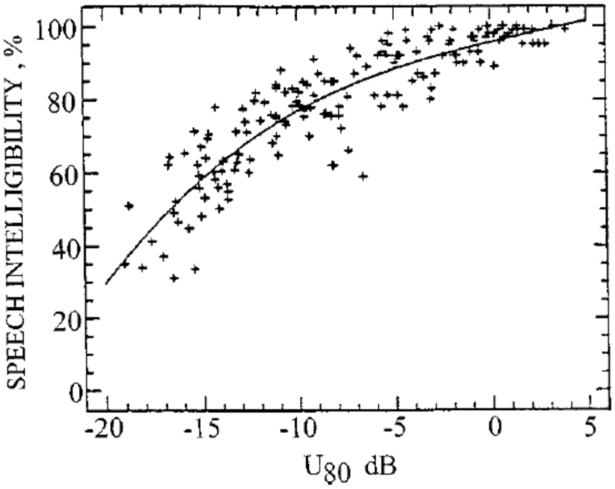
	 **Mapping between subjective score (abscissa) and PESQ score (ordinate)**
Interests for the **I'City For All Project**	PESQ algorithm takes into account perceptual hearing features that can be used to simulate the perceived quality of announces.
Possible adaptation for presbycusis	The psychoacoustics models implemented in PESQ can be modified to reflect the hearing loss of presbycusis. In fact the absolute hearing level can be modified together with masking effect.
Observations	PESQ was extended to a new objective algorithm named Perceptual Objective Listening Quality Assessment (POLQA) that takes into account super wideband for speech communication and reverberation. Algorithm is available in Matlab and C.
References	ITU-T recommendations P.862/P.862.1/ P.862.2/ P.862.3

Measure of	Intelligibility
Name	**Speech-based Speech Transmission Index (Speech-based STI)**
Applications	Used in public address communication systems.
Model	The speech-based STI is an intelligibility measure that is based on speech modulation frequency to compute an intelligibility score. It's derived from STI measure that uses synthetic signal to compute intelligibility score. The speech-based STI uses real sentences to extract Modulation Transfer Function (MTF). One of methods that computes the MTF from speech is the envelope regression method and it is expressed as follows : $$m_k = \frac{\mu_{xk}}{\mu_{yk}} \frac{E\{(x_k(t)-\mu_{xk})(y_k(t)-\mu_{yk})\}}{E\{(x_k(t)-\mu_{xk})^2\}}$$ Where μ_{xk} and μ_{yk} are the temporal mean of $x_k(t)$ and $y_k(t)$. $x_k(t)$ and $y_k(t)$ are the temporal envelopes of speech filtered by kth octave band filter.
Inputs	Intrusive measure that takes as input : ➔ clean and degraded speech
Outputs	Intelligibility score correlated with the standard STI score.
Complexity	**Speech-based STI** can be used in real time situation but with frame size above 0.3s to keep good correlation with **STI**.
Known limits	Not effective with short frame size.

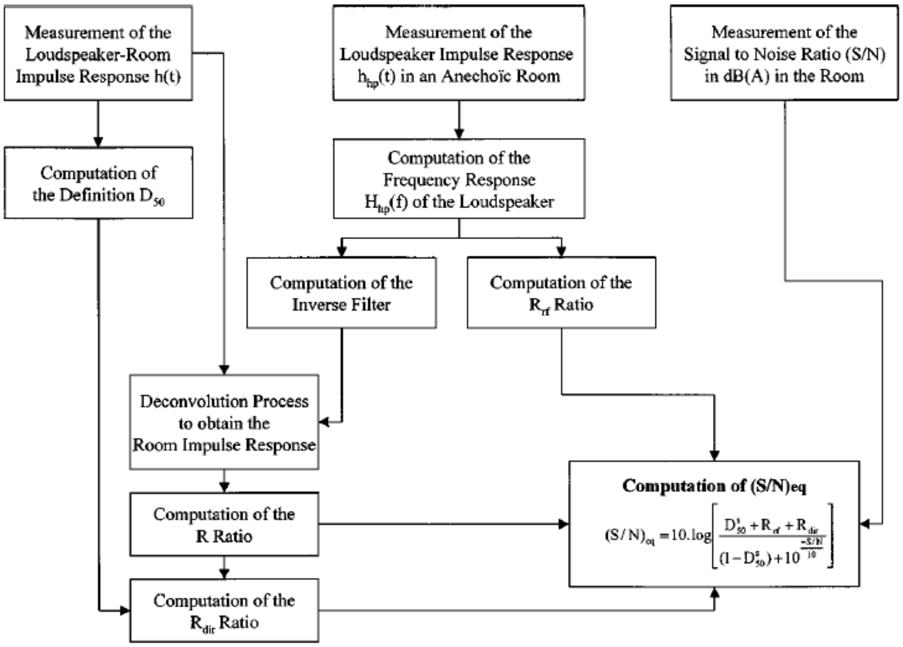
Known performances and conditions of evaluation	 **Metric computed from ER with noise in left column and ER with noise+reverberation vs. Theoretical STI using 0.3 s windows in top and 78ms windows in bottom. The solid lines represent best linear fits to the data. [19]**
Interests for the **I'City For All Project**	Takes into consideration noise and reverberation degradation. Less restrictive for real live test in railway station.
Possible adaptation for presbycusis	The same adaptation can be done for STI and speech based STI in two different ways : - Adapt the model by for example varying the masking effect in function of age. - Adapt the scale of STI scores to reflect the perceived intelligibility as it was done in IEC standard of STI.
Observations	Other method was developed [17] to compute speech-based STI like Normalized Correlation (NC) method and real cross-power spectrum method. A CEA Linklab implementation of the algorithm is available in Matlab.
References	[12][13][16][17][19]

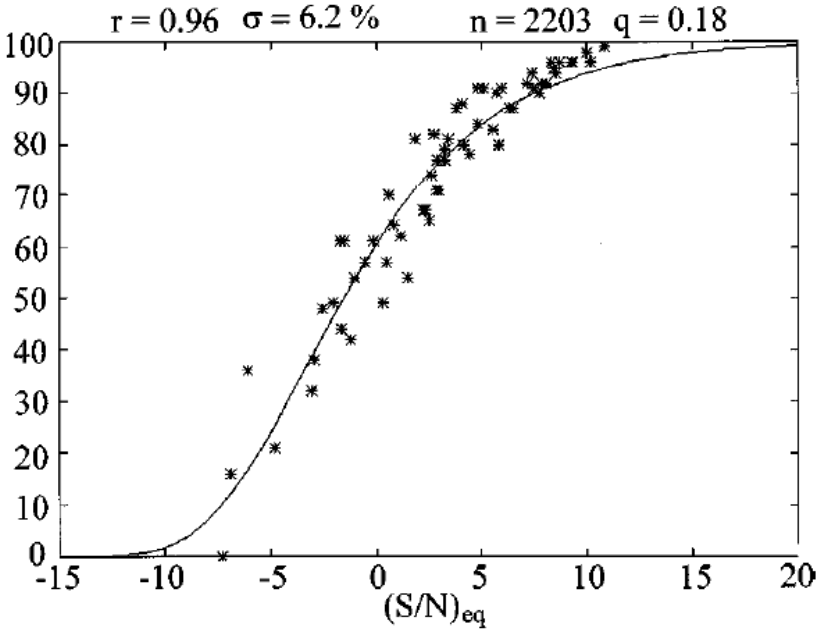
Measure of	Intelligibility
Name	**Coherence Speech Intelligibility Index (CSII)**
Applications	Used for hearing aid evaluation~~s~~.
Model	The CSII is an extension of SII ANSI standard to cover the nonlinear distortion introduced by enhancement algorithm. The model is based on the coherence measure to predict effective noise from speech signal. The SNR becomes then a Signal-to-noise and Distortion Ratio (SDR) and it is computed as follows: $SDR(j) = \frac{\sum_{k=0}^{K} W_j(k)\hat{P}(k)}{\sum_{k=0}^{K} W_j(k)\hat{N}(k)}$ $\hat{P}(k) = \lvert\gamma(k)\rvert^2 S_{yy}(k)$ $\hat{N}(k) = [1 - \lvert\gamma(k)\rvert^2]S_{yy}(k)$ $\lvert\gamma(k)\rvert^2 = \frac{\lvert\sum_{m=0}^{M-1} X_m(k)Y_m^*(k)\rvert^2}{\sum_{m=0}^{M-1}\lvert X_m(k)\rvert^2 \sum_{m=0}^{M-1}\lvert Y_m(k)\rvert^2}$ • $\hat{P}(k)$ and $\hat{N}(k)$ predicted speech and noise power spectra • $\lvert\gamma(k)\rvert^2$ coherence measure • $S_{yy}(k)$ auto-spectral density • $X_m(k)$ and $Y_m(k)$ are the spectra of mth window of clean and degraded speech The **CSII** is computed in three amplitude regions of speech envelope and we obtain the intelligibility score as follows : $c = -3.47 + 1.84CSII_{low} + 9.99CSII_{Mid} + 0.0CSII_{High}$ $I_3 = \frac{1}{1 + e^{-c}}$
Inputs	Intrusive measure that takes as input : ➔ clean and degraded speech

Outputs	**CSII** score between 0 and 1.
Complexity	The criterion is computed frame by frame but not in real time.
Known limits	Not effective for reverberation.
Known performances and conditions of evaluation	 **Proportion of the HINT sentences indentified correctly plotted versus the three-level CSII intelligibility predictions I_3 for the normal-hearing subjects**
Interests for the **I'City For All Project**	Takes into consideration additive noise like (railway station noise) and nonlinear noise introduced by enhancement algorithm.
Possible adaptation for presbycusis	Considers the hearing loss of presbycusis as an internal noise for each frequency band. This noise can be added as an SDR for each band depending on the audiogram of the subject.
Observations	The CSII is based on SII standard and differs only in computing the effective SNR. We use the same weights for each band as for SII. The algorithm is not available.
References	[21]

Measure of	Intelligibility
Name	**Useful-to-detrimental ratio**
Applications	Room acoustic quality.
Model	The useful-to-detrimental ratio is expressed as follows [23] : $$U_{te} = 10\,log\left[\frac{R_{te}}{(1-R_{te}) + 10^{(-S/N)/10}}\right]$$ S/N is the signal to noise ratio R_{te} is the ratio between early and total energy: $R_{te} = E_e/(E_e + E_l)$ 'te' is the time limit between late sound arrival and early time arrival
Inputs	Non-intrusive measure that takes as input : ➔ Room impulse response
Outputs	**SI** score between 0% and 100% of speech recognition.
Complexity	Easy to compute. The intelligibility can be predicted quickly if we know the room impulse response.
Known limits	Does not take into account Non-Linear degradation and speech enhancement algorithms.
Known performances and conditions of evaluation	The Speech Intelligibility (SI) is predicted for 'te' 80ms as follow : $$SI = 95.65 + 1.219\,U_{80} - 0.02466\,U_{80}^2 + 0.00295\,U_{80}^3$$ 

	Measured speech intelligibility scores versus 1kHz U_{80} values and 3rd order polynomial best fit with STD error 7.5%
Interests for the **I'City For All Project**	Takes into consideration additive noise and room reverberation.
Possible adaptation for presbycusis	The **Useful-to-detrimental ratio** is computed for different reflected frequencies. The idea is to find the frequencies that represent the presbycusis person.
Observations	This measure is a variant of clarity measure [22] that doesn't take into account noise degradation. Algorithm is not available.
References	[23] [22]

Measure of	Intelligibility
Name	**Equivalent Signal to Noise ratio (SNeq)**
Applications	Room acoustic quality
Model	 Measurement of the Loudspeaker-Room Impulse Response h(t) Measurement of the Loudspeaker Impulse Response $h_{hp}(t)$ in an Anechoïc Room Measurement of the Signal to Noise Ratio (S/N) in dB(A) in the Room Computation of the Definition D_{50} Computation of the Frequency Response $H_{hp}(f)$ of the Loudspeaker Computation of the Inverse Filter Computation of the R_{rf} Ratio Deconvolution Process to obtain the Room Impulse Response Computation of the R Ratio Computation of the R_{dir} Ratio Computation of (S/N)eq $(S/N)_{eq} = 10.\log\left[\frac{D_{50}^{s}+R_{rf}+R_{dir}}{(1-D_{50}^{s})+10^{\frac{-S/N}{10}}}\right]$ Method of (S/N)eq computation
Inputs	Non-intrusive measure that takes as input : ➔ Room impulse response ➔ Loudspeaker impulse response ➔ Signal-to-Noise Ratio of the room
Outputs	Signal to effective noise ratio in dB.
Complexity	The required inputs add complexity to the algorithms. This measure can't be done online.
Known limits	Tested for one position of transmitter/receiver. Use single loudspeaker. Don't take into account enhancement algorithm.

Known performances and conditions of evaluation	The regression line with intelligibility score is obtained as follows : $I(\%) = 100(1 - 10^{-[(S/N)_{eq}+40]/(60\times0.18)})^{2203}$  Measured speech intelligibility scores versus (S/N)eq predictor corresponding values and best least-squares fit
Interests for the I'City For All Project	Measure of speech intelligibility including room, loudspeaker and background noise influence.
Possible adaptation for presbycusis	As for **Useful-to-detrimental ratio** we can find the frequencies that are significant for presbycusis person and adapt the measure in this perspective.
Observations	Algorithm is not available.
References	[26]

Applying auditory saliency in the context of the I’City For All Project

Several urban places dedicated to public transport (airports, train stations...) use vocal announces to communicate information to passengers. One aspect of the AAL I'City for All project is to enhance the intelligibility of such announces so that every passenger could understand the messages despite the degraded listening conditions in those environments (e.g. ambiant noise due to the crowd and the traffic, reverberation, poor loudspeakers quality) aggravated by impaired hearing. Some of us are working on objective measures to predict the intelligibility of a speech signal, measures that would take into account all degradation types including presbycusia.

However, even if a vocal announce is intelligible, i.e. if the entire message of the nnounce is actually understandable, it does not mean that passengers will listen to it. Indeed, during the diffusion of the announce, users could be engaged in another task in parallel (like phone call, reading, video game...). It is therefore very important that announces attract the attention of the concerned users. The ability for a sound to attract attention is referred as **auditory saliency**.

Furthermore, listeners suffering from presbycusia generally report some difficulties to segregate the different sources of a complex acoustic scene and to focus on one specific sound of this noisy environment. Despite these difficulties, they can achieve the same speech recognition performances than normal listeners at the cost of high mental effort leading to auditory fatigue [12]. Now, previous studies on visual and auditory perception have demonstrated that the perceptual processing of a salient object, either a sound or a picture, requires very few cognitive resources compared to the processing of a non salient object. Therefore, increasing the saliency of vocal announces will reduce the auditory fatigue “for all” passengers.

Our goal is therefore to establish some objective measures to predict the auditory saliency of a vocal announce, depending on different acoustic parameters like signal-on-noise ratio, signal spectrum, voice type or even intelligibility. The final objective is to define some guidelines to conceive and produce salient intelligible announces, as well as enhancement algorithms that would correct the different degradation applied on the signal in terms of saliency and intelligibility.

Overview of auditory saliency measures

While a lot of visual saliency models have been investigated, the idea of modeling auditory attention is relatively new and very few auditory saliency models are available. Auditory saliency models should be able to detect sounds and predict which ones should be treated first by the auditory system. We would consider as salient, sounds that can be noticed without attention or that can capture the listeners' attention and cause them to shift their attention from the currently attended task.

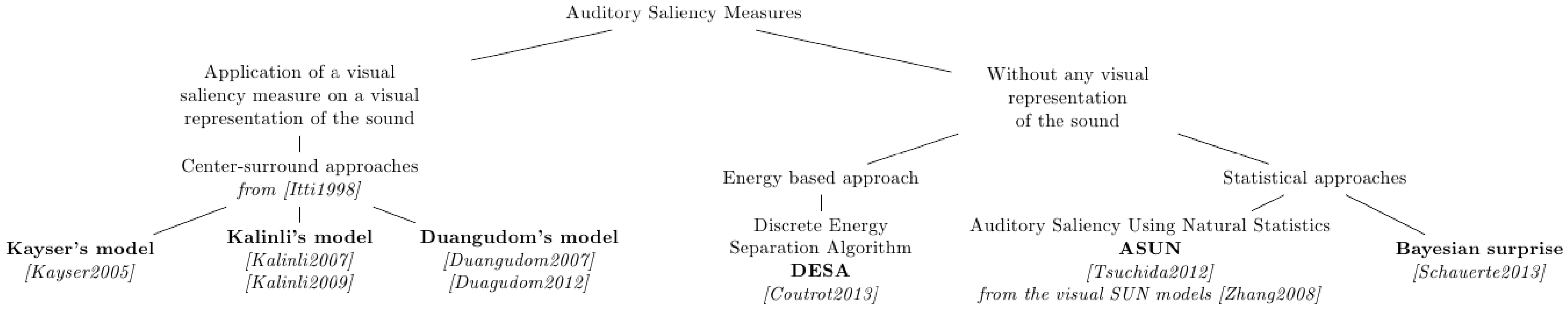
Several issues make the modeling of auditory saliency a challenging task. First, even if auditory and visual systems are similar in many ways, they differ in the features used to analyze complex scenes. In vision, the basic features of early processing have been extensively studied since the Feature Integration Theory (see Annex 1). The conception of visual saliency models is made easy through the analysis of, for example, color, luminance, orientation, shape or contrast. On the contrary, very few primitives have been determined in audio. It is therefore more complex to define the appropriate feature set of an auditory saliency model. Basically, auditory models presented in the current section rely on intensity and temporal or spectral modulations.

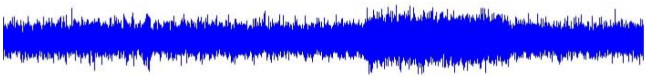
A second issue in defining auditory saliency models is the fact that audio has a temporal component. The auditory scene is constantly changing over time. On the contrary, the visual saliency models work on still images not varying in time. Furthermore, in hearing there is an effect of temporal masking that can be either backward or forward masking. Therefore both past and future sounds are important to predict what is heard and salient. Time should be treated carefully.

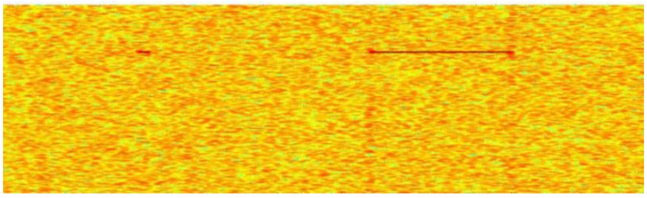
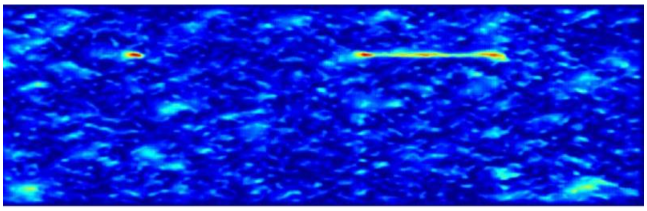
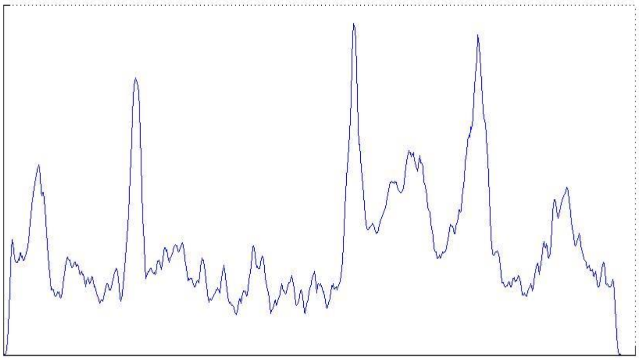
Finally, research on auditory saliency could have been reduced because of the difficulties in evaluating the models. Visual saliency models are evaluated by comparing the predicted salient regions with the areas actually looked at by participants during eye-tracking experiments. Since the auditory system does not have any physical correlate that can be easily measured, i.e. it is not possible to directly measure what is actually listen to, it is difficult to evaluate the auditory saliency models. Annex 5 discusses the experimental protocols that have been suggested for auditory saliency evaluation.

Despite all these difficulties, six models have been developed to measure auditory saliency. As summarized in **Error! Reference source not found.**, the three first methods are biologically inspired and simply rely on the application of a well-known visual saliency measure [14] on the spectrogram or cochleogram of the sound. On the contrary, the three other methods do not mimic the auditory system and do not require, in a first step, to transform the sound in a picture. The rest of this section presents a summary of each method. A more detailed description is available in the annexes.

Classification of auditory saliency models


Auditory Saliency Measures
Application of a visual saliency measure on a visual representation of the sound
Center-surround approaches *from [Itti1998]*
Kayser's model *[Kayser2005]*
Kalinli's model *[Kalinli2007]* *[Kalinli2009]*
Duangudom's model *[Duangudom2007]* *[Duagudom2012]*
Without any visual representation of the sound
Energy based approach
Discrete Energy Separation Algorithm **DESA** *[Coutrot2013]*
Statistical approaches
Auditory Saliency Using Natural Statistics **ASUN** *[Tsuchida2012]* *from the visual SUN models [Zhang2008]*
Bayesian surprise *[Schauerte2013]*

Measure of	Auditory Saliency
Name	**Kayser's model**
Model	The principle is to apply a visual saliency model to a visual representation of the sound. Biologically inspired, it mimics the auditory perceptual system. The model is decomposed in several steps: 1) **Basic spectrogram:** Decompose the signal in a visual representation of frequency over time 2) **Extracting features**, i.e. applying a visual saliency model on the spectrogram. The spectrogram is analyzed at different scales (gaussian pyramid) through gabor filters. Then the different scales are compared through a center-surround mechanism to obtain three feature maps: - intensity - frequency modulations (vertical variations) - temporal modulations (horizontal variations) **3)** **Inhibition stage** = normalization of each maps to promote or suppress some of the feature **4)** Association (weighted averaging) of the 3 normalized maps to obtain a **saliency map** 5) OUR IMPROVEMENT: averaging over each frequency bands to obtain a **saliency curve** The complete algorithm is in Annex 2.
Inputs	Recording of the degraded signal
Outputs	A **saliency map** (to super-imposed to the spectrogram) indicating which zone of the spectrogram is more salient. A **saliency score** is suggested as the peak level of the saliency map. We suggest transforming the saliency map into a **saliency curve** indicating the evolution of saliency level over time.
Example	Basic example obtained with a short tone followed by a long tone in a white Gaussian noise. Waveform 

	Spectrogram  Saliency map  Saliency curve (not in the original paper)  One can observe three highest peaks. The first peak indicates the position of the first tone then the two last peaks indicate the beginning and the end of the long tone.
Complexity	Visual representation requires a frame by frame analysis. Complexity also increases with the resolution of the spectrogram (i.e. the number of pixel of the visual representation).
Applications	Originally it was only evaluated in laboratory conditions to test the correlation between the model estimation and the perceived saliency. It could be used applied to vocal announces as well as sound alarms in car to predict the detection of such messages.
Interests for the I'City For All Project	It is the first so the must known auditory saliency measure. Not intrusive
Known limits	Used future sample to compute the normalized features so is not usable for real time processing. Takes very few parameters into account. Other acoustic parameters

	that could be important for saliency measure are not known yet. Experimental procedures are limited both due to stimuli limits (no evaluation on speech) and experimental tasks (detection level is not exactly saliency level, defining saliency to participants is difficult)
Known performances and conditions of evaluation	1) Reproduces basic properties of auditory scene perception as demonstrated with basic examples (long tone more salient than short ones, modulated tones more salient than stationary tones, the second of a sequential pair of tones is less salient, missing parts in a broad spectrum are salient) 2) Well correlated to human performances according to laboratory tests on environmental sound snippets: - pairwise comparison (2AFC test) where the task is to choose the most salient sound between two possibilities (significant correlation of 0.47±0.1, p<0.05) - detection task: salient sounds are detected more often (81% *versus* 71%) than less salient sounds even if their intensity is low compared to the background noise level (spearman rank correlation r=0.56, p<0.01) 3) Well correlated to macaque monkey behavior (they turn their head more in the direction of a sound if it is salient)
Possible adaptation for presbycusis	During the frequency analysis used for processing the spectrogram, it is possible to mimic the loss of high frequency hearing by applying a frequency weighting. The normalization step takes forward masking into account. It is maybe possible to modify this step to take other masking effects into account. Using a different weighting to associate the three features is also a way to adapt the measure to the elderly as they may rely more on one of the features.
References	[16]
Related methods	The Kayser's model relies on the Itti & Koch's model dedicated to visual saliency measurement [14]. Models from Duangudom [8],[7], Kalinli [15] and de Coensel are some extensions of the Kayser's model.

Measure of	Auditory Saliency
Name	**Kalinli's model**
Model	Equivalent to Kayser's model except first and second stages 1) **Auditory spectrogram** (modelling early auditory processing, equivalent to cochleogram) 2) Feature extraction. Same procedure but more features are extracted (intensity, frequency modulations, temporal modulations + orientation of pitch variations) 3) Normalization 4) Association of the different maps
Inputs	The degraded signal
Outputs	Saliency map
Example	Not implemented
Complexity	Almost the same as in Kayser's model
Applications	Used to determine accent in prosody Recently used for phoneme separation and speech recognition
Interests for the I'City For All Project	Reveal that different visual representations can be used. It could be an interesting parameter to vary for a presbycusis adaptation. More complete than Kayser's model as more features are taken into account. Not intrusive.
Known limits	No real-time. No evaluation was carried out to compare subjective performances of naïve listeners to objective measures. The pitch feature finally causes performance degradation.
Known performances and conditions of evaluation	Model prediction was compared to expert annotation of prominent syllables
Possible adaptation for presbycusis	Similar as those proposed for the Kayser's model.

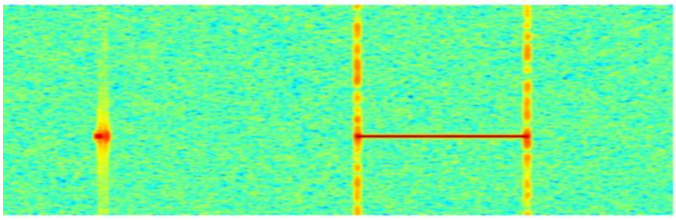
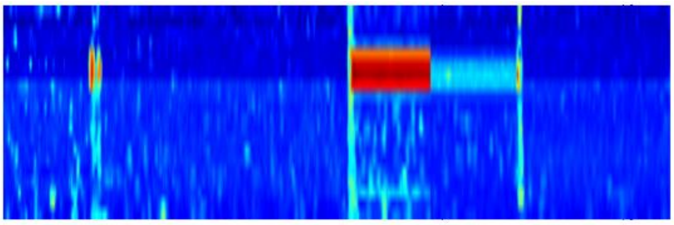
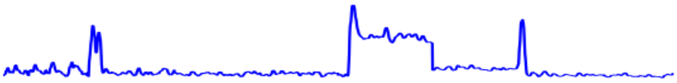
References	[15]
Related methods	Based on the visual saliency map model of Itti [14] and the extended the auditory saliency map model of Kayser.

Measure of	Auditory Saliency
Name	**Duangudom's model**
Model	Biologically inspired. Equivalent to Kalinli's model except second stage (so equivalent to Kayser's model except 1st and 2nd stages) 1) Auditory spectrogram 2) Feature extraction = overall energy distribution + temporal modulations + frequency modulations + areas with simultaneous temporal & frequency modulations 1) Normalization 2) Saliency map obtained by association of the different feature maps
Inputs	Degraded signal
Outputs	Saliency map or saliency curve
Example	Not implemented as it is too close from Kayser and Kalinli's models
Complexity	Almost the same as in Kayser's model
Applications	Used to find acoustic parameters that influence auditory saliency
Interests for the I'City For All Project	The main interest resides in the evaluation protocols used to evaluate this model. Otherwise the model itself is too close from Kalinli's and Kayser's models. Not intrusive.
Known limits	Tested with various stimuli but never with the same protocol so it is not possible to assure the efficiency of this model for all kind of stimuli especially for voice.
Known performances and conditions of evaluation	1) Reproduces basic properties of auditory scene perception as demonstrated with basic examples. 2) Well correlated to human performances in three experiments: - Pairwise sound comparison (average correlation between participants and model responses = 0.47, std = 0.22) - Comparison of five 1 second movie segments of a 5 seconds

	extract (mean correlation between participants and model responses = 0.48, std= 0.11, p=0.003) - Dual task experiment with laboratory stimuli (pure tones): primary task = counting low tones in a sequence, secondary task = detection of a modulated noise. Increasing saliency of the modulated tones improved performances of both primary and secondary tasks.
Possible adaptation for presbycusis	Similar than those proposed for Kayser's model
References	[7], [8]
Related methods	Based on the visual saliency map model of Itti [14] and extended the Kayser's model.

Measure of	Auditory Saliency
Name	**Discrete Energy Separation Algorithm (DESA)**
Model	The model is based on the Teager-Kaiser energy used for detecting amplitude and frequency modulations in AM-FM signals. FOR EACH TIME FRAME: Step 1: **Multiband demodulation analysis** (Gabor filtering) FOR EACH SAMPLE OF THE FRAME: Step 2: Computation of the **Teager-Kaiser energy** for each sample Step 3: Choosing the frequency band of step 1 that maximize the Teager Kaiser energy Step 4: Compute the **instant frequency** and the **instant amplitude** Step 5: **averaging** the Teager-Kaiser energy, the instant amplitude and the instant frequency over all the samples of the frame and normalized each feature Step 6: combining the three averaged and normalized features to obtain the **saliency score of the frame**. Step 7: **thresholding** to detect salient events
Inputs	The degraded signal
Outputs	A saliency curve indicating the evolution of saliency level over time + time of salient events.
Complexity	Frame by frame filtering Only 6 frequency bands (compared to the 256 frequency bands of the Kaiser's spectrogram)
Applications	Used by Evangelopoulos *et al* for video summarization and speech detection in noise [9] [10]. Used by Coutrot *et al* to predict saccades in eye movements (muti-sensory perception) [5]
Interests for the I'City For All Project	Computation times are reduced compared to those of previous models. The number of samples from the future required to compute the DESA measure is very limited so it can be more easily adapted to

	real time. Not intrusive.
Known limits	No real-time Not really evaluated as a predictor of human behaviour.
Known performances and conditions of evaluation	Evaluating by comparing annotations on movies from expert annotators.
Possible adaptation for presbycusis	Possible to process instant amplitude and instant energy not only in the frequency bands that maximize the Taiger-Kaiser energy but in all frequency bands and then ponderate the contribution of each band.
References	[9], [10], [5]
Related methods	Based on studies about detection of modulations in AM-FM signals with Taiger-Kaiser [Kaiser1990].

Measure of	Auditory Saliency
Name	**Auditory Saliency Using Natural statistics (ASUN)**
Model	The principle is to compute the difference between the signal at time k and the expected signal at the same time knowing the past samples. The difference measure is computed over different features directly obtained through a Principal Components Analysis on the past samples.
Inputs	Degraded signal
Outputs	A saliency map + a saliency curve
Example	With the same example as in Kayser's model, i.e. a short plus a long tone (example from [25]): Cochleogram  Saliency map  Saliency curve 
Applications	Not indicated.
Interests for the I'City For All Project	Does not require any sample from the future. Not intrusive.
Known limits	The past samples and PCA measures are updated only every 250 ms due to computional limits. Optimization is required.
Known performances and conditions of	Pairwise comparison with participants having to choose the most

evaluation	"interesting sound". Pearson correlation between ratings of participants and predictions by the model are equal in mean equal to 0.3262 (Standard Deviation = 0.0635) and is higher with urban and animal sounds than with other environmental sounds.
Possible adaptation for presbycusis	The use of cochleogram instead of a simple spectrogram confirm the hypothesis that it is possible to use a visual representation of the sound that take into account a model of hearing loss.
References	[25]
Related methods	Based on the visual saliency SUN model [27].

Measure of	Auditory saliency
Name	**Bayesian surprise**
Model	It relies on a probabilistic model of the signal’s frequency distribution applied on the spectrogram of the sound.
Inputs	speech/synthetic signal/impulse response
Outputs	Saliency curve, i.e. saliency score S(t) for each time t
Example	A matlab implementation of the algorithm is available online at: http://www.mathworks.com/matlabcentral/fileexchange/33573-gaussian-surprise-and-running-windowed-mean-variance A demonstration is included with the sound file downloadable at: https://cvhci.anthropomatik.kit.edu/~bschauer/code/data/surprise_demo.22k.wav
Complexity	Low compared to center-surround approaches. Authors give an estimation of 1.5 sec to process 1 min of sound
Applications	Control of computational resources of humanoid robots (control sensor orientation in direction of salient sounds to optimize the scene analysis). Association with visual saliency measures for multimodal attention modeling.
Interests for the I’City For All Project	Low run-time so is more appropriate for real-time measurements. Not intrusive.
Known limits	Algorithm parameters substantially influences the performances and run-time so it will needs some tests to adjust the parameters to our application.
Known performances and conditions of evaluation	Measures of precision and recall in a detection task of salient acoustic events previously annotated by one expert. The database is the CLEAR2007 database composed of recordings of meetings.
Possible adaptation for presbycusis	As in Kayser’s model, the importance of each frequency band contribution can be ponderated to be adjusted to the perception of presbycusis listeners.
References	[21]
Related methods	Extended previous works of the authors [22]

Conclusions and future works

We presented six models of auditory saliency measures and observed that they were never compared in any reviewing paper neither in a comparison study. It is thereby difficult to predict which of these models will be the more adapted for our own project.

Moreover, except for one study, they were only validated through laboratory conditions (detection tasks, pairwise comparisons, pure tones or isolated environmental sounds). The only attempt of an ecological validation was presented in [6] that confirmed that salient sounds of traffic transports are more disturbing. The measure referred in this paper was limited to subjective ratings of comfort. No voice stimulus was used in this experiment neither any measure of mental effort. The other experimentation used to validate auditory saliency models were also dedicated to very specific application without any of them being reusable for our own project concerning vocal announces.

We suggest first to develop an experimental protocol adapted to the saliency estimation of vocal announces.

We also suggest some improvements of the saliency models. Actually several aspects of these models can be modified.

First of all, a lot of visual saliency measures have been proposed in the last decades (18 papers just in the 2012 European conference on Computer Vision). The SUN model and the Center Surround approach of Itti&Koch was already extended to audio. We assume that other methods can be used to analyse an auditory spectrogram. For example, the measures of visual saliency from image histograms [17] and spectral residual ([13] implementation in Annex 6) have been proved to be extremely efficient in terms of computational costs making them ideal for real-time processing.

A second possibility would be to combine several of the already available auditory saliency measures for example mixing results from a statistical approach and a biologically inspired method to analyze more acoustic features and enlarge the number of acoustic conditions that can be treated efficiently.

None of the measures described above have been tested on presbycusis participants. It would be very interesting for our project to either **determine the minimum level of saliency to achieve** so every announces and alarms will be salient enough **to be attractive for all** listeners or to adjust the saliency measures themselves so they can predict the behaviour of all listeners.

Finally another issue would be to determine how saliency is influenced by the natural degradations also modifying the intelligibility of announces like, reverberations, non-linearities, kind of background noise. Furthermore the link between saliency and intelligibility was never studied.

Sharpness Index measure:

Measure of	Clarity
Name	**Audio Sharpness Index**
Model	The principle is to measure the sensitivity of the total variation of a signal (actually any regularity measure) to the convolution of the signal by a white gaussian noise.
Inputs	Distorted speech
Outputs	score
Example	
Complexity	The SI is computed on long frames of signal (1 to several seconds). Its complexity is of order N.log2(N), where N is the number of samples of the frame Although its complexity makes it relevant for real time, the lengths of the frames of analysis dedicate it to low-reactivity real-time.
Applications	Originally dedicated to image sharpness evaluation, it could be used to measure the clarity of any sound having undergone any impairment
Interests for the I'City For All Project	Low complexity Non-intrusive, which avoids synchronization between test and reference signals
Known limits	Today not validated as a clarity measure. The SI were never applied to sound till now.

Known performances and conditions of evaluation	According to preliminary experiments : – When speech is corrupted by white noise, the SI has the same variations as the STI, but in a different range of SNRs (10 to 40dB instead of -15 to 15) – In the case of reverberation, the SI is a decreasing function of the reverberation time in a similar manner as the STI, though decreasing faster.
Possible adaptation for presbycusis	Since the STI is more sensitive to noise and reverberation than the STI, it could be relevant as a clarity index for hearing-impaired people, whereas its variations according to conjugated noise and reverberation seem to make it unrelevant for normal-hearing people
References	no
Related methods	Based on [Blanchet2012]

PART II

Detailed description

of objective assessment algorithms and criteria

Classical and standardized assessment methods

Table of Contents

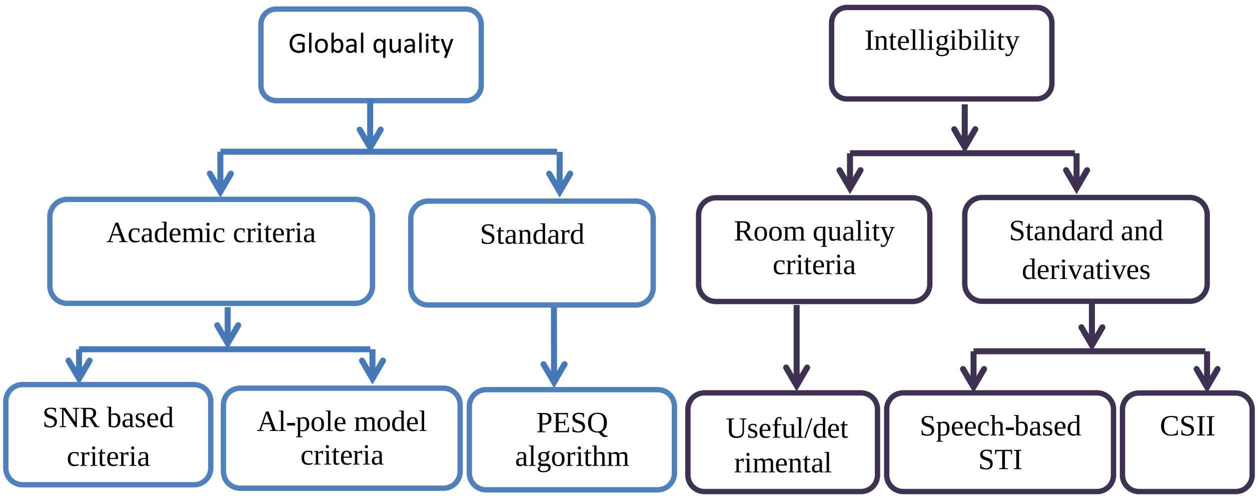
Introduction

The scope of this deliverable is to analyse existing objective measures of quality that cover a very large spectrum of applications and identify the ones that can be used reliably in the analysis and optimization of I'CityForAll algorithms. Indeed, the quality measure should reflect the improvement of I'CityForAll algorithms and should cover the different degradations for "all" population.

Speech quality can be observed from different angles. Indeed, many perceptual attributes can describe the speech quality as for example the most known: global quality and intelligibility. Global quality is a multi-dimensional attribute and can include various types of other attributes like "naturalness", "scratching", "noisy", etc… this is why during the subjective evaluation of the global quality, a description of the observed attributes must be given to the test subject. On the other side, intelligibility of speech can be easily quantified by counting the number of phonemes, syllables or words identified by the test subject.

Global quality	**Intelligibility**
Multi-dimensional attribute	Uni-dimensional attribute
Highly subjective	Easy to quantify Percentage of phonemes or words recognition by list
Cultural dependence	Cognitive dependence

In the first section of this part, we highlight the kind of quality we are interested in: global quality, intelligibility and comfort… In the second section, we describe the objective measure of quality that could be used in I'CityForAll project for enhancement filter optimization and intelligibility measure "for all". In this part of the report, the objective measure is organized in three categories: academic mathematical criteria, standards criteria and their derivatives and finally room acoustics criteria as described by the organization chart below.



Organization chart 1: objective measure of speech quality

1. Subjective speech quality measure

The quality is by definition subjective. It depends on linguistic, cultural and cognitive aspects. We describe in this section some subjective tests that quantify the global quality and intelligibility.

Subjective measure of intelligibility

The test of intelligibility may be carried out with different types of phonetic databases: syllables, words or sentences. These databases should have:

- ✓ Phonetically balanced content to represent the distribution of phonemes commonly used in the language under test,
- ✓ The same level of difficulty,
- ✓ A controlled contextual information.

We describe below 3 types of intelligibility tests:

- **/V-C-V/:** The database of this test consists of different nonsense syllables presented in the format /V-C-V/, where V and C refer respectively to vowel and consonant. One vowel is fixed to the entire database, the most used in the language, like /a/ or /e/, and the consonants are chosen to also cover the most frequently used in the language. The Consonants are altered and corrupted and then presented to a group of listeners for identification. The percentage of identification of consonant per list is the intelligibility score [1].

- **DRT:** The Diagnostic Rhyme Test is composed of lists of rhyming word couples (veal-feel/bean-peen/dense-tense/vast-fast/…) that have the same phonetic feature (Voicing, Nasality, Graveness…). This type of list helps to localize the most affected feature due to the degradation. The subjects are asked to underline the heard word. The percentage of words identification per list yields the intelligibility score [2].
- **HINT:** The Hearing In Noise Test is composed of lists of phonetic balanced sentences. These sentences are diffused at a specific noise level and the subject is asked to repeat what he heard. The percentage of words identification per sentence yields the intelligibility score [3].

For the purpose of I'CityForAll project, a modified HINT test was proposed for a better intelligibility evaluation depending on noise and reverberation. It was also be motivated by a new approach of ecological tests to help the patient "to be aware" of his deficiency. This test is described in appendix A.

Subjective measure of global quality

The subjective evaluation of global quality of speech is normalized by the I-TUT P.800 recommendation [4]. This standard is applied to evaluate the transmission quality of speech. The databases are composed of phonetically balanced sentences and the test consists in asking the subject to assess the quality of these sentences by giving a score of quality between 1 and 5, where 1 is bad and 5 is excellent.

No subjective tests for global quality measure have been planned in I'CityForAll project because we are more concerned with the intelligibility for all than with the global quality.

Besides, we propose to focus also on the degree of "comfort" of speech because it is necessary to improve the intelligibility for elderly without troubling the speech comfort for the normal hearing person. However, "speech comfort" as attribute must be defined rigorously to avoid bias in subjective evaluation. It is clear that "speech comfort" is related to "loudness comfort" and "acoustic comfort" but are there other parameters that contribute to "speech comfort" variation? An investigation should be done to lighten this attribute.

2. Objective speech quality measure

It is well known that subjective measures of quality are financially and time consuming. Such subjective tests are not planned within the scope of I'CityForAll. To avoid this constraint, the quality will be measured with objective assessment methods. However, there is a large spectrum of objective measures of speech quality, our goal is to identify the most efficient "for all" situations.

In this section, we firstly discuss some academic objective quality measures that are easy to compute. Those measures can be used quickly to evaluate the digital filter for speech enhancement. We then describe the standardized objective measures and we focus on its derivatives which could be used "for all". Finally, a complementary objective measure based on an acoustics approach is presented as a criterion that can be used to develop a more global objective criterion of intelligibility "for all" that covers most of degradations.

2.1. Academic mathematical criteria of speech quality

Signal to Noise Ratio and derivatives: the Signal to Noise Ratio (SNR) is the most used criterion to measure sound quality not for its accuracy but thanks to its simplicity. The overall SNR is measured by equation 1.

$$\boldsymbol{SNR = 10\ \log_{10}\left(\frac{\sum_{n=1}^{N} x_n^2}{\sum_{n=1}^{N}(x_n - \hat{x}_n)^2}\right)} \qquad (\mathbf{1})$$

Where $\boldsymbol{x_n}$ is the clean signal, $\boldsymbol{\hat{x}_n}$ the noisy signal and N the length of the signals (in samples).

A derived measure from the SNR is the segmental SNR (SNRseg) which corresponds to the geometric mean of the SNR of each frame of the signal. The SNRseg is defined in equation 2. In practice, the SNRseg can get large negative values due to the silent frames. To resolve this, the log function is shifted by 1 to make the SNRseg positive.

$$\boldsymbol{SNRseg_R = \frac{10}{M}\sum_{m=1}^{M} \log_{10}\left(1 + \frac{\sum_{n=Nm}^{Nm+N-1} x_n^2}{\sum_{n=Nm}^{Nm+N-1}(x_n - \hat{x}_n)^2}\right)} \qquad (\mathbf{2})$$

Where $\boldsymbol{M}$ is the number of frames. Note that it is important to align the clean and noisy signals on the time axis.

A SNRseg extension was developed in [5] based on the frequency domain which consists in measuring the spectral SNRseg with a weighted filter bank. The fwSNRseg is described by equation 3.

$$\boldsymbol{fwSNRseg = \frac{10}{M}\sum_{m=1}^{M} \frac{\sum_{j=1}^{K} W_j\ \log_{10}\left[F^2(m,j) \Big/ (F(m,j) - \hat{F}(m,j)^2)\right]}{\sum_{j=1}^{K} W_j}} \qquad (\mathbf{3})$$

Where $\boldsymbol{M}$ is the number of frames, $\boldsymbol{K}$ the number of filters in the filter bank, $\boldsymbol{W_j}$ the weight of the jth frequency band, $\boldsymbol{F(m,j)}$ the amplitude of the jth frequency band of clean signal, $\boldsymbol{\widehat{F}(m,j)}$ the amplitude of jth frequency band of noisy signal. $\boldsymbol{W_j}$**,** the frequency weight, can be adjusted to have optimal correlation with subjective tests.

To maximize correlation between subjective and objective measures, Barwell [6] formulates differently the fwSNRseg by interchanging the summations between frequency and frame in order to compute a linear regression for frequency weights optimization. With this formulation we obtain the so called frequency-*variant* objective measures (equation 4):

$$\boldsymbol{fwvar = W_0 + \sum_{j=1}^{K} W_j \left[\frac{1}{M}\sum_{m=1}^{M} 10\ \log_{10}\left[F^2(m,j) \big/ (F(m,j) - \widehat{F}(m,j)^2)\right]\right]} \quad (4)$$

We note that the frequency-variant weighted SNR is very useful as it can be correlated with different objective scores of intelligibility or quality by just carrying on a linear regression. This can be done for hearing impaired population as well as for normal hearing population.

Quality measure based on all-pole models: One of speech modelling theories assumes that intervals of speech between 15-30 ms can be represented by an all-pole model with low $\boldsymbol{p}$ orders [7] as described by equation 5.

$$\boldsymbol{x_n = \sum_{i=1}^{p} (a_x(i).x_{n-i}) + \ G_x.e_n} \quad (5)$$

Where $\boldsymbol{p}$ is the model's order, $\boldsymbol{a_x(i)}$ are the coefficients of the all-pole filter, called also LPC coefficients, $\boldsymbol{G_x}$ the filter gain and $\boldsymbol{e_n}$ white filter excitation.

Based on this model, several distance measures between clean and noisy speech were derived. We describe in the following 3 major objective criteria that use different distances:

- Log Likelihood Ratio distance **(LLR)**
- Itakura Saito distance **(IS)**
- Cepstral distance **(CEP)**

The LLR is defined as a distance between LPC coefficients of clean and distorted speech.

$$d_{LLR}(\overrightarrow{\boldsymbol{a_x}}, \overrightarrow{\boldsymbol{a_{\hat{x}}}}) = \log \frac{\overrightarrow{\boldsymbol{a_{\hat{x}}}}.\boldsymbol{R_x}.\overrightarrow{\boldsymbol{a_{\hat{x}}}}^{\boldsymbol{T}}}{\overrightarrow{\boldsymbol{a_x}}.\boldsymbol{R_x}.\overrightarrow{\boldsymbol{a_x}}^{\boldsymbol{T}}} \quad (6)$$

Where $\overrightarrow{\boldsymbol{a}_x}, \overrightarrow{\boldsymbol{a}_{\hat{x}}}$ are respectively vectors containing the LPC coefficients of clean and distorted speech and $\boldsymbol{R}_x$ the autocorrelation matrix of the clean speech.

For the Itakura Saito measure, the filter gain G_x is introduced in the distance expression:

$$d_{IS}(\overrightarrow{\boldsymbol{a}_x}, \overrightarrow{\boldsymbol{a}_{\hat{x}}}) = \frac{G_x}{G_{\hat{x}}} \frac{\overrightarrow{\boldsymbol{a}_{\hat{x}}}.\boldsymbol{R}_x.\overrightarrow{\boldsymbol{a}_{\hat{x}}}^{\boldsymbol{T}}}{\overrightarrow{\boldsymbol{a}_x}.\boldsymbol{R}_x.\overrightarrow{\boldsymbol{a}_x}^{\boldsymbol{T}}} + \log\left(\frac{G_{\hat{x}}}{G_x}\right) - 1 \quad (7)$$

$$G_x = (r_x^T \overrightarrow{a_x})^{1/2} \quad (8)$$

Where r_x^T is the autocorrelation of the clean signal and “$.^T$” refers to vector transposition.

Note that IS measure gives importance to the overall spectral levels through the filter gain which is in contradiction with psychoacoustics studies [8] which state that changes in sound level have a minimal effect on quality.

A derived form of LPC coefficients provides a Cepstrum which is an estimation of smoothed speech spectrum as following:

$$\log\left(\frac{1}{\boldsymbol{A}_x\,(\boldsymbol{z})}\right) = \sum_{k=1}^{\infty} c(k).z^{-k} \qquad (9)$$

Where $c(k)$ denotes the Cepstral coefficients. We can obtain the Cepstral coefficients by a recursive computing as follows:

$$c(k) = a(k) + \sum_{i=1}^{k-1} \frac{i}{k} c(i) a_{k-i} \quad (10)$$

The Cepstral distance is obtained by:

$$d_{cep}(c_{x,}c_{\hat{x}}) = \frac{10}{\log_e 10} \sqrt{2 \sum_{i=1}^{p} [c_x(i) - c_{\hat{x}}(i)]^2} \qquad (11)$$

Weighted Spectral Slope distance measure (WSS)

The WSS is a spectral measure based on the distance between spectral slopes. This measure is motivated by the influence of formant frequency deviation on quality.

The spectral slope $S_x(k)$ is measured by the difference between the intensities of successive critical bands $(C_x(k+1), C_x(k))$ as follows:

$$S_x(k) = C_x(k+1) - C_x(k) \qquad (12)$$

$$S_y(k) = C_y(k+1) - C_y(k) \qquad (13)$$

The WSS is then measured by weighting the distance between the reference and the degraded spectral slopes as:

$$d_{WSS}\left(C_x(k), S_y(k)\right) = \sum_{k=1}^{K} W(k).\left(S_x(k) - S_y(k)\right)^2 \qquad (14)$$

The weight $W(k)$ can be adjusted to maximize correlation between the subjective and objective quality scores:

$$W(k) = \frac{K_{max}}{[K_{max} + C_{max} - C_x(k)]} \frac{K_{loc\,max}}{[K_{loc\,max} + C_{loc\,max} - C_x(k)]} \qquad (15)$$

Where K_{max} and $K_{loc\,max}$ are constants which can be used for correlation with subjective measure. C_{max} is the largest log-spectral magnitude for all bands and $C_{loc\,max}$ is the largest peak nearest band (k).

In [9], the correlation of all these academic objective measures with subjective global quality scores is investigated. The obtained correlation coefficients and standard deviations of prediction error are summarized in the table below.

Table 1 : correlation between objective measures and subjective global quality

Objective Measures	Correlation Coefficients	Standard deviations of error
WSS	0.53	0.52
LLR	0.63	0.47
IS	0.45	0.54
CEP	0.60	0.49
fwvarSNR (K=25)	0.81	0.36
fwSNR (k=25)	0.70	0.43
SegSNR	0.31	0.58

We note from the results above that it is very interesting to use the frequency-variant weighted SNR (fwvarSNR) because on one hand it presents a very good correlation factor (0.81) and on the other hand it can be adapted for different subjective measures.

2.2. Standard objective criteria and derivatives

We describe in this subsection standardized intrusive measures, which compare clean and distorted sentences to compute quality scores.

Perceptual Evaluation of Speech Quality (PESQ): Normalized by the ITU-T in the P.862 recommendation [10], PESQ is a perceptual evaluation based on psychoacoustic models. PESQ algorithm is composed by three main modules as illustrated in figure 1:

- Preprocessing: used to compensate the delay and weakness of the signal introduced by network transmission. Besides, PESQ includes an IRS filter (Intermediate Reference System) that models the telephone terminal.
- Perceptual model: converts the original and degraded signals to the frequency domain, then in perceptual loudness on Barks scale.
- Cognitive model: Different distance measures are used between original and degraded perceptual representations to compute the quality score.

Finally, a Mean Opinion Score (MOS) is deduced from the quality score that correlates with the subjective score of global quality.

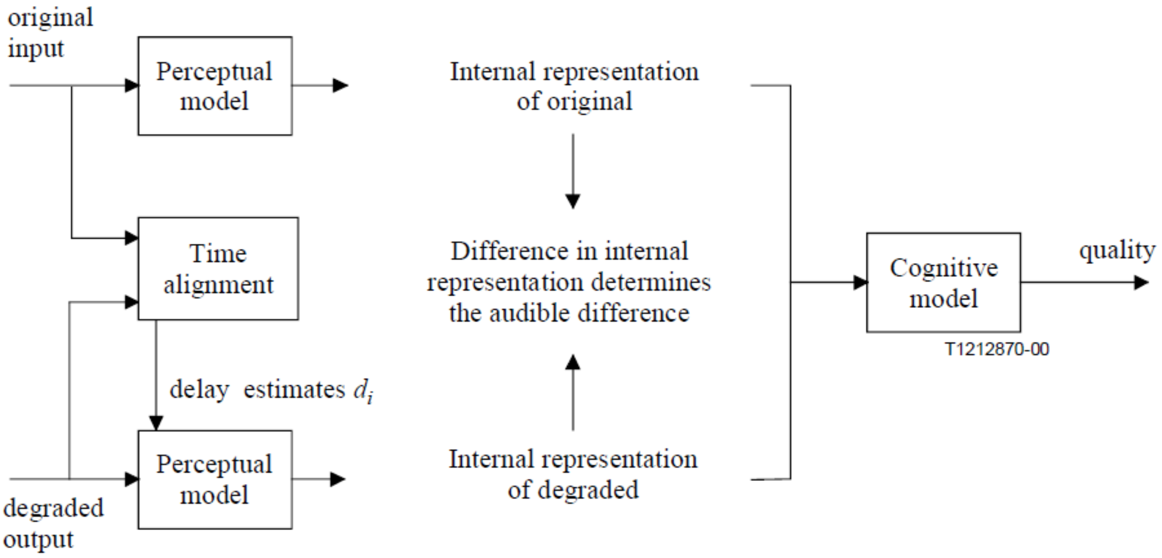

original input
Perceptual model
Internal representation of original
Time alignment
Difference in internal representation determines the audible difference
Cognitive model
quality
T1212870-00
delay estimates d_i
Perceptual model
Internal representation of degraded
degraded output

Figure 1 : Descriptive flow chart of PESQ [10]

For 22 known ITU benchmark experiments, the average correlation with subjective measures of global quality was 0.935 [11] for telephony transmission quality as illustrated in figure 2.

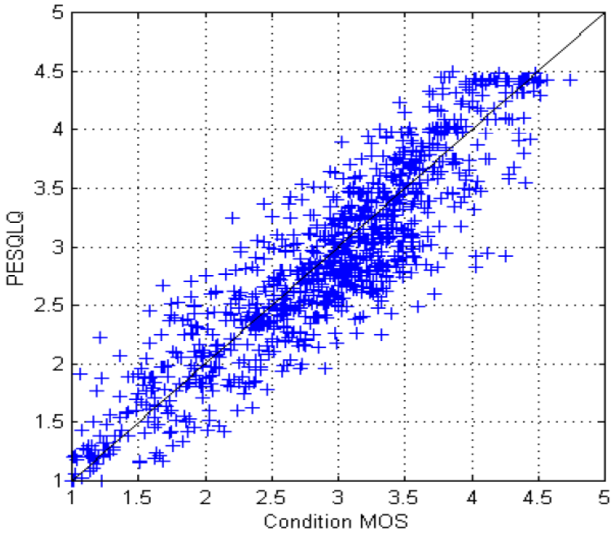
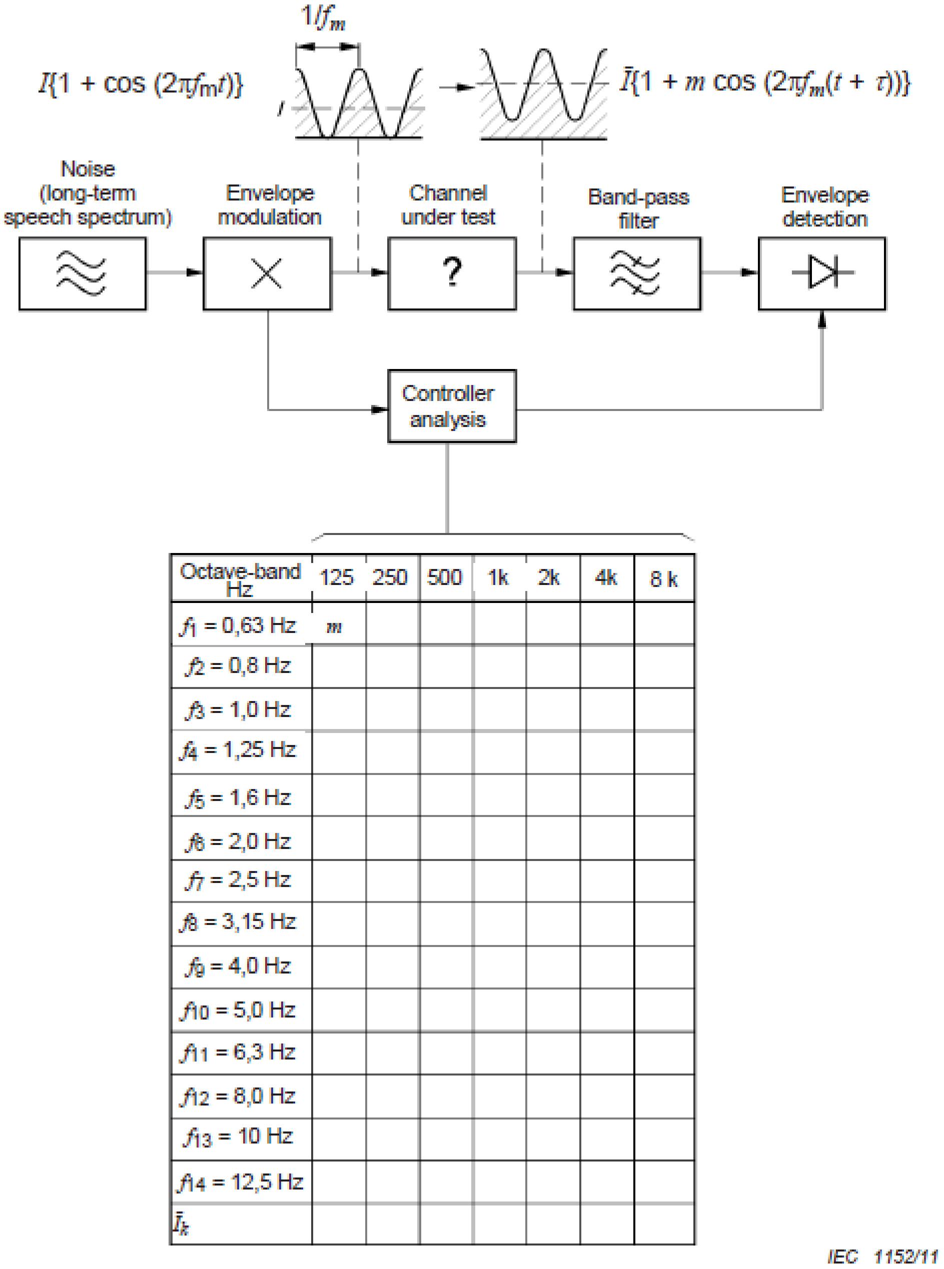


Figure 2 : Mapping of PESQ listening quality score vs. subjective mean opinion score of British sentences [11]

Speech Transmission Index (STI): developed by Houtgast and Steeneken [12] and normalized by IEC in part 16 of sound system equipment [13], the objective rating of speech intelligibility by Speech Transmission Index (STI) is a measure based on Modulation Transfer Function (MTF) in reverberant and noisy envirement.

The MTF can be measured with a speech, Room Impulse Response (RIR) or “modulated speech shaped noise”. The direct STI measure takes “modulated speech shaped noise” as input. The values of MTF are measured for 14 modulations frequency and 7 octave bands as illustrated in figure 3. The indirect STI measure takes RIR as input. The MTF of indirect method are measured with Schroeder method [14] derived by the equation 16.



Octave-band Hz	125	250	500	1k	2k	4k	8 k
f_1 = 0,63 Hz	m						
f_2 = 0,8 Hz							
f_3 = 1,0 Hz							
f_4 = 1,25 Hz							
f_5 = 1,6 Hz							
f_6 = 2,0 Hz							
f_7 = 2,5 Hz							
f_8 = 3,15 Hz							
f_9 = 4,0 Hz							
f_{10} = 5,0 Hz							
f_{11} = 6,3 Hz							
f_{12} = 8,0 Hz							
f_{13} = 10 Hz							
f_{14} = 12,5 Hz							
$\bar{I}_k$							

Figure 3 : Descriptive scheme of MTF computing with speech shaped noise probe [13]

$$m_k(f_m) = \frac{\left|\int_0^\infty h_k(t)e^{-j2\pi f_m t}dt\right|}{\int_0^\infty h_k(t)^2\,dt}.\left[1 + 10^{-SNR/10}\right]^{-1} \qquad (16)$$

Where k is the number of octave band, f_m the modulation frequency, h_k the impulse responses and SNR the signal to noise ratio.

After MTF computing, the procedure to reach the STI score is the same:

- Correction of the MTF using auditory masking: the STI 2010 revisions introduce the effect of the frequency masking in the MTF computing. This masking is modeled in STI algorithm as a noise addition to the octave band k depending on the intensity of octave band k-1.

$$\acute{m}_{k,f_m} = m_k(f_m) \times \frac{I_k}{I_k + I_{am,k} + I_{rt,k}} \qquad (17)$$

Where, $\acute{m}_{k,f_m}$is the modified MTF taking into account auditory masking, I_k is the intensity level of octave band k, $I_{am,k}$ the intensity of noise addition due to masking effect of octave band k and $I_{rt,k}$ the equivalent intensity of absolute threshold of octave band k.

- Computing the effective SNR: the $SNR_{eff\ k,fm}$ is SNR that takes into account the noise and the reverberation as effective noise. It's derived from the MTF as follows :

$$SNR_{eff\ k,fm} = 10 \times log10\left(\frac{\acute{m}_{k,f_m}}{1 - \acute{m}_{k,f_m}}\right) \quad (18)$$

- The transmission index (*TI*) are than computed from the effective SNR as follows :

$$TI_{k,fm} = \frac{SNR_{eff\ k,fm}+15}{30} \quad (19)$$

- The modulation transmission index are derived from the *TI* as an average over modulation frequencies (fm) :

$$MTI_k = \frac{1}{n}\sum_{m=1}^{n} TI_{k,fm} \qquad (20)$$

Where $'n'$ is the number of modulation frequencies.

- Finally the STI score is computed as follows :

$$STI = \sum_{k=1}^{7} \alpha_k \times MTI_k - \sum_{k=1}^{6} \beta_k \times \sqrt{MTI_k \times MTI_{k+1}} \qquad (21)$$

Where α_k and β_k are respectively the weight and redundancy factors for octave band k depending on the gender as detailed in the table below.

Table 2 : STI weights per gender

Octave band (Hz)		125	250	500	1000	2000	4000	8000
Males	α	0.085	0.127	0.230	0.233	0.309	0.224	0.173
	β	0.085	0.078	0.065	0.011	0.047	0.095	-
Females	α	-	0.117	0.223	0.216	0.328	0.250	0.194
	β	-	0.099	0.066	0.062	0.025	0.076	-

The STI predicted score of intelligibility has a good correlation with subjective C-V-C measure as illustrated in figure 4.

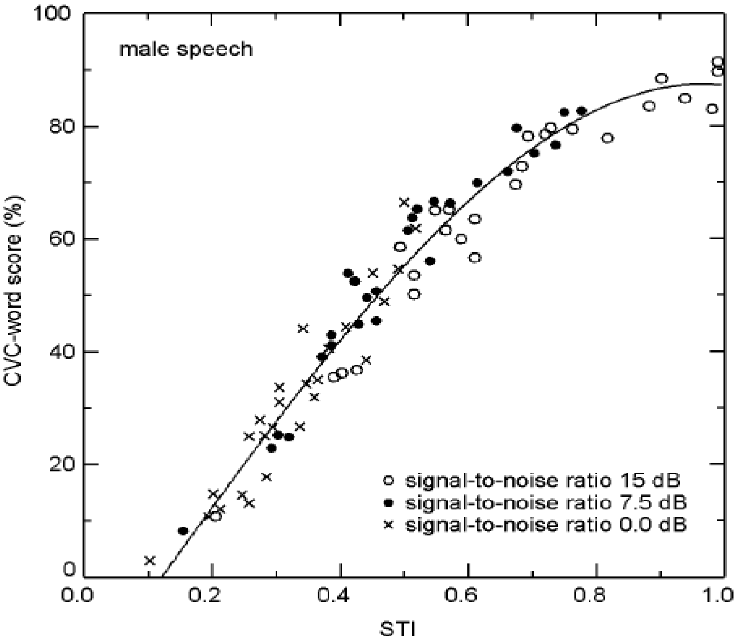


Figure 4: Relation between STI and CVC-word scores for 78 conditions involving MALE speech. The standard deviation, representing the vertical spread around the 3rd order polynomial best-fitting is s = 4.7% [15].

Speech based STI:

During the development of STI, Houtgast and Steeneken tried to use speech as input for computing MTF. They observe artefacts when measuring the envelope spectra of degraded speech [12, 16]. This artefact consists of increases in intensity envelope spectra when theory predicts decreases. This introduces bias in the calculation of the MTF.

The first work of Houtgast *et al.* proposes to compute the MTF of speech based STI as follow:

$$m_k(f_m) = \alpha \sqrt{\frac{S_{yy}(f_m)}{S_{xx}(f_m)}} \qquad (22)$$

Where $\alpha = E\{x(t)\}/E\{y(t)\}$. S_{yy} and S_{xx} are respectively power spectra of degraded and clean speech. To obtain the STI score we follow the same method as for STI direct measure.

To avoid artefacts, other methods can be used to calculate the MTF or the effective SNR directly. We can find a good resume of these methods in Ray L. Goldsworthy works [17]. In all those different methods, we are interested by Envelope Regression method (ER) proposed by Ludvigsen et al. in 1990 [18] and modified by Goldsworthy in 2004[17]. The method was tested by Payton and Mona in 2008 [19] in real time conditions with noise and reverberation degradation.

The ER method consists in computing the MTF as follows :

$$m_k = \frac{\mu_{xk}}{\mu_{yk}} \frac{E\{(x_k(t) - \mu_{xk})(y_k(t) - \mu_{yk})\}}{E\{(x_k(t) - \mu_{xk})^2\}} \qquad (23)$$

Where μ_{xk} and μ_{yk} are the mean of $x_k(t)$ and $y_k(t)$ which are the temporal envelopes of speech filtered by kth octave band filter.

We observe in figure 5 and 6 that the evolution of speech based STI with ER method converge to the theoretical STI when the frame size is greater than 0.3 second. In fact, we can obtain, in case of only noise degraded speech, a coefficient of correlation between 0.91 and 0.99 for frame size between 78ms and 0.3s.

However, when adding reverberation, the 78ms frame size becomes not valid and we need frame sizes greater than 0.3s to obtain at minimum a coefficient of correlation 0.79.

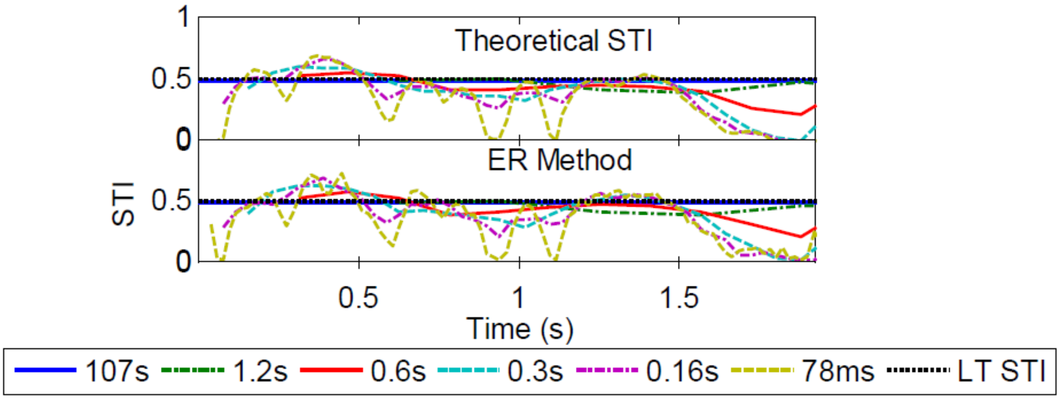


Figure 5 : metric results vs. Window length (top) theoretical STI (bottom) ER method for 0 dB SNR stationary speech-shaped noise condition. The black dotted line in each plot represents the long-term STI [19]

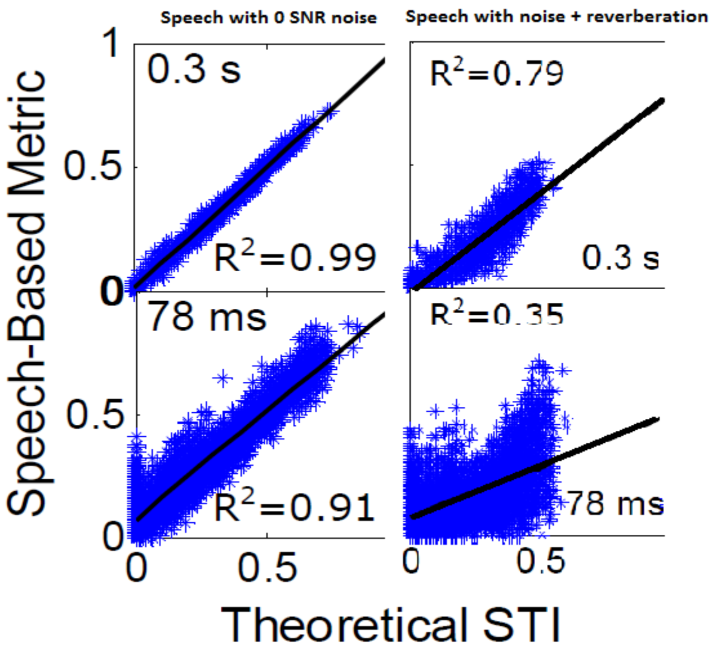


Figure 6 : Metric computed from ER with noise in left column and ER with noise+reverberation vs. Theoretical STI using 0.3 s windows in top and 78ms windows in bottom. The solid lines represent best linear fits to the data. [19]

This method is very interesting for the I'CityForAll project because firstly, speech-based STI can be used frequently in a public address system like a railway station, instead of the direct STI method that use a speech shaped noise to compute STI and thus needs for this an empty railway station. Secondly, the real time processing is very useful for a mobile application that computes the STI in different point of a public space. Finally, to adapt the STI "for all", it's more relevant to use real speech as probe than a synthetic signal because we can include psychoacoustic models in STI computing.

Speech Intelligibility Index (SII) and Coherence measure of SII (CSII):

SII, ANSI standard [20] (ANSI S3.5-1997), assumes that the speech and noise spectra have been measured separately. As for the frequency weighted SNR, the SNR is calculated in SII standard for each 1/3 octave, octave or critical bands as follow:

$$SNR(j) = \frac{\sum_{k=0}^{K} W_j(k)P(k)}{\sum_{k=0}^{K} W_j(k)N(k)} \qquad (24)$$

Where $P(k)$ and $N(k)$ are respectively the power spectrum of speech and noise. The $W_j(k)$ is computed as follow:

$$W_j(k) = \left(1 + p_j g\right) \exp\left(-p_j g\right) \quad (25)$$

$$p_j = \frac{4(1000 q_j)}{b_j} \quad (26)$$

$$g = \left|1 - f/q_j\right| \quad (27)$$

Where q_j the center frequency of band in kHz and b_j is the bandwidth of j^th^ band.

We note that SII does not take into consideration the distortion introduced by the communication system. Indeed, to compute intelligibility score we just need ambient noise power. This can constitute a problem for I'CityForAll enhancement algorithms as it can introduce non-linear distortion to enhanced speech. Kates and Arehart [21] extend SII criteria to include non-linear distortion introduced by hearing aids. They propose to compute the Signal-to-noise and Distortion Ratio (SDR) instead of the classic SNR. For this, they predict the speech and noise (ambient noise + distortion) power spectrums by the Magnitude Squared Coherence function (MSC) as follow:

$$\hat{P}(k) = |\gamma(k)|^2 S_{yy}(k) \quad (28)$$

$$\hat{N}(k) = [1 - |\gamma(k)|^2] S_{yy}(k) \quad (29)$$

Where $\hat{P}(k)$ and $\hat{N}(k)$ are the predicted speech and noise power spectra. $|\gamma(k)|^2$ is the MSC and typically estimated using Fourier Transform in each frame. The MSC is given by:

$$|\gamma(k)|^2 = \frac{|\sum_{m=0}^{M-1} X_m(k) Y_m^*(k)|^2}{\sum_{m=0}^{M-1} |X_m(k)|^2 \sum_{m=0}^{M-1} |Y_m(k)|^2} \quad (30)$$

Where $X_m(k)$ and $Y_m(k)$ are the spectra of m^th^ window of x(n) and y(n) the clean and degraded signal. Then, the SDR is given bys:

$$SDR(j) = \frac{\sum_{k=0}^{K} W_j(k) \hat{P}(k)}{\sum_{k=0}^{K} W_j(k) \hat{N}(k)} \quad (31)$$

Note that if only additive noise is present in the communication system, SDR should give the same result as SNR (figure 7), otherwise in presence of distortion like clipping, the SDR decreases due to MSC reduction.

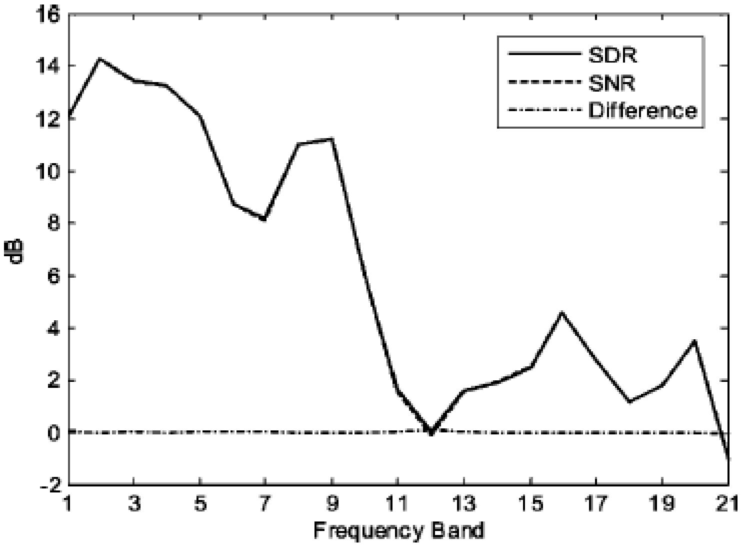


Figure 7 : SDR (solid line), SNR(dashed) and difference between them (dot-dashed) as a function of the frequency band number for the concatenated HINT sentences. Additive noise low-pass filtered at 900 Hz is present at an SNR of 10 dB [21].

However additive noise and distortion do not affect the speech signal in the same way. Kates [21] proposes to segment the speech signal envelope into three amplitude regions by computing the RMS level of each frame. The middle zone is between [10,30] dB, high zone upper 30 dB and low zone inferior to 10 dB. We obtain then, $\text{CSII}_{\text{low,}}$ CSII_{Mid} and $\text{CSII}_{\text{High}}$. In fact, the high-level segments will be most strongly affected by peak clipping, while the low-level segments will be affected by additive noise and center clipping [21]. Besides, the correlation between CSII_{low} and overall CSII is only 0.70 indicating that the low-level CSII provides information that is absent in overall CSII. Finally, with the help of non-linear minimization procedure we combine the three levels of CSII to obtain an intelligibility score (named I_3) as follow:

$$c = -3.47 + 1.84CSII_{low} + 9.99CSII_{Mid} + 0.0CSII_{High} \qquad (32)$$

$$I_3 = \frac{1}{1 + e^{-c}} \qquad (33)$$

I_3 model is shown in figure 8 and it predicts intelligibility with correlation coefficient of 0.94.

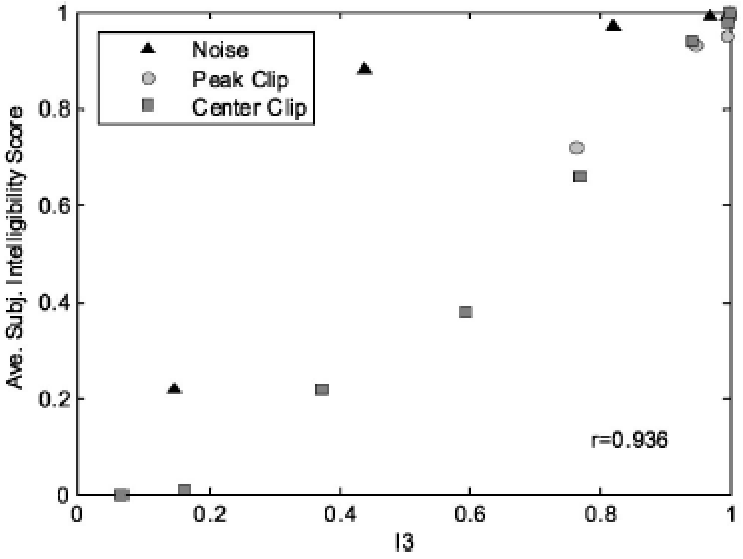


Figure 8 : proportion of the HINT sentences indentified correctly plotted versus the three-level CSII intelligibility predictions I_3 for the normal-hearing subjects [21].

We note that for SII and CSII measure for hearing-impaired person, Kates and ANSI standard consider the hearing loss as an internal noise source [21,20]. A frequency dependent SNR will be added to SII SNR or SDR CSII to simulate hearing-loss.

In addition to the STI, the criterion described previously does not take into account room reverberation. However, public space consideration in I'CityForAll project are often a very large rooms. The temporal distortions (reverberation, echo, and crosstalk) that occur in those places are very important and can't be neglected. We will describe in the next paragraph some criteria that take into consideration the acoustics features of large rooms to compute speech quality.

2.3. Room acoustics criterion

Those criteria are based on the separation of received energy signal on two parts:

- Useful energy: this part of signal is associated with the direct received sound with the first replicated part of signal.
- Delayed energy: this part of signal is associated with the late replicated part of signal, plus background noise arriving to the receiver.

Measure of speech intelligibility from Useful/Delayed energy:

Early/late sound ratio have relates to the degree of clarity for music and intelligibility for speech. Lochner and Burger [22] introduce the clarity measure and it's given by equation below:

$$C_{te} = 10 \log\left(\int_0^{te} h^2(t)dt \Big/ \int_{te}^{\infty} h^2(t)dt\right) \qquad (34)$$

Where 'te' is the time limit between late sound arriving and early time arriving,$h(t)$ is the room impulse response.

Bradley [23] develops after the concept of useful and detrimental sound energy that was used to predict speech intelligibility score. The useful-to-detrimental ratio is expressed as follow:

$$U_{te} = 10\ log\left[\frac{R_{te}}{(1 - R_{te}) + 10^{(-S/N)/10}}\right] \qquad (35)$$

Where S/N is the signal to noise ratio and R_{te} is the ratio between early and total energy: $R_{te} = E_e/(E_e + E_l)$.

Bradley use a te=80ms to predict speech intelligibility with 7.5% standard error using 1Khz octave band. The best-fit curve for predicting speech intelligibility is given by the third-order polynomial equation (36) and illustrated in figure 9.

$$SI = 95.65 + 1.219\ U_{80} - 0.02466\ U_{80}^2 + 0.00295\ U_{80}^3 \quad (36)$$

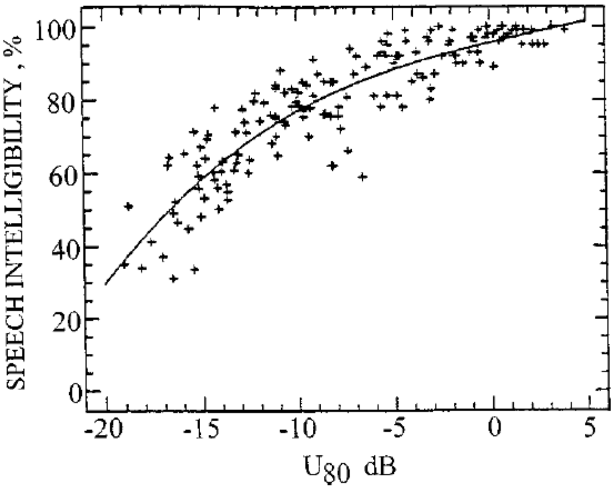


Figure 9 : measured speech intelligibility scores versus 1Khz U$_{80}$ values and 3rd order polynomial best fit [23]

Other methods used the concept of useful/detrimental sound energy [24,25] to compute the intelligibility score but we are going to focus in Faiget and Ruiz model [26] that includes a separation of room, loudspeaker and noise influence.

Measure of speech intelligibility including room, loudspeaker and background noise influence:

As for Clarity measure, the prediction of speech intelligibility comes from Room Impulse Response (RIR) measure. The idea of Faiget and Ruiz model [26] is to compute a deconvolution of RIR to extract room and loudspeaker effects separately. Take the following formulation of RIR:

$$h(t) = h_{hp}(t) * h_s(t) + n(t) \qquad (37)$$

Where, $h_{hp}(t)$ is the impulse response of loudspeaker, $h_s(t)$ is the impulse response of the room and $n(t)$ is the noises. If we de-noise the RIR response, we can find the inverse filter $f(t)$ which verifies $f(t) * h(t) = h_s(t)$. We measure $h_{hp}(t)$ in anechoic chamber and we deduce the $f(t)$ as the inverse of the loudspeaker IR. Finally, via the Fourier Transform we can found $h_s(t)$.

Than the useful/detrimental energies is measure for room and loudspeaker. We compute the ratio for room as follow:

$$D_{50}^{s} = \frac{\int_0^{50} h_s^2(t)dt}{\int_0^{T} h_s^2(t)dt} \qquad (38)$$

Where T is the total time of $h(t)$ and D_{50}^{s} is the influence of the room.

The influence of the loudspeaker is measured as follow:

$$R_{dir} = \frac{D_{50}}{D_{50}^{s}} = \frac{\int_0^{50} h^2(t)dt \int_0^{T} h_s^2(t)dt}{\int_0^{T} h^2(t)dt \int_0^{50} h_s^2(t)dt} \qquad (39)$$

Besides of loudspeaker IR influence, the distortion in frequency response between 100hz and 4khz are taken into account. In fact, the tolerance in deviation is fixed at $\mp 1.5\ dB$ to be more restrictive. The criteria which take into account frequency fluctuation is computed as follows:

$$R_{rf} = \frac{E_{hp} - E_{n,hp}}{E_{hp}} = 1 - \frac{E_{n,hp}}{E_{hp}} \qquad (40)$$

Where, E_{hp} is the energy of the frequency response in the band 100-4000 hz and $E_{n,hp}$ is the energy above and under the tolerance $\mp 1.5\ dB$ as illustrated in figure 10.

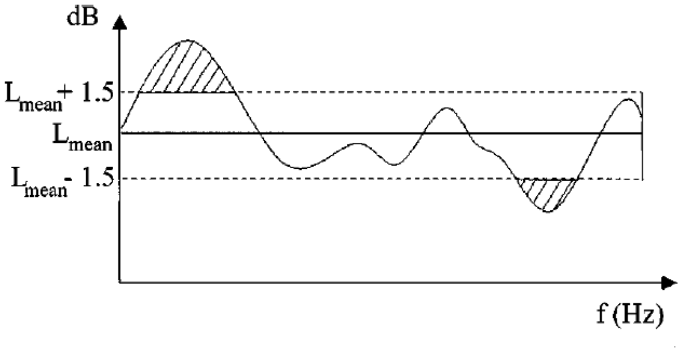


Figure 10 : simulation of a loudspeaker frequency response with upper and lower R_{rf} limits [26].

The final model is given by the equivalent signal-to-noise ratio as follow:

$$(S/N)_{eq} = 10log\left(\frac{D_{50}^{s} + R_{dir} + R_{rf}}{\left(1 - D_{50}^{s}\right) + 10^{(-S/N)/10}}\right) \quad (40)$$

To resume, we need loudspeaker IR measure in anechoic room, RIR and Signal to Noise Ratio to compute the finally equivalent (S/N)$_{eq}$ as illustrate in figure 11.

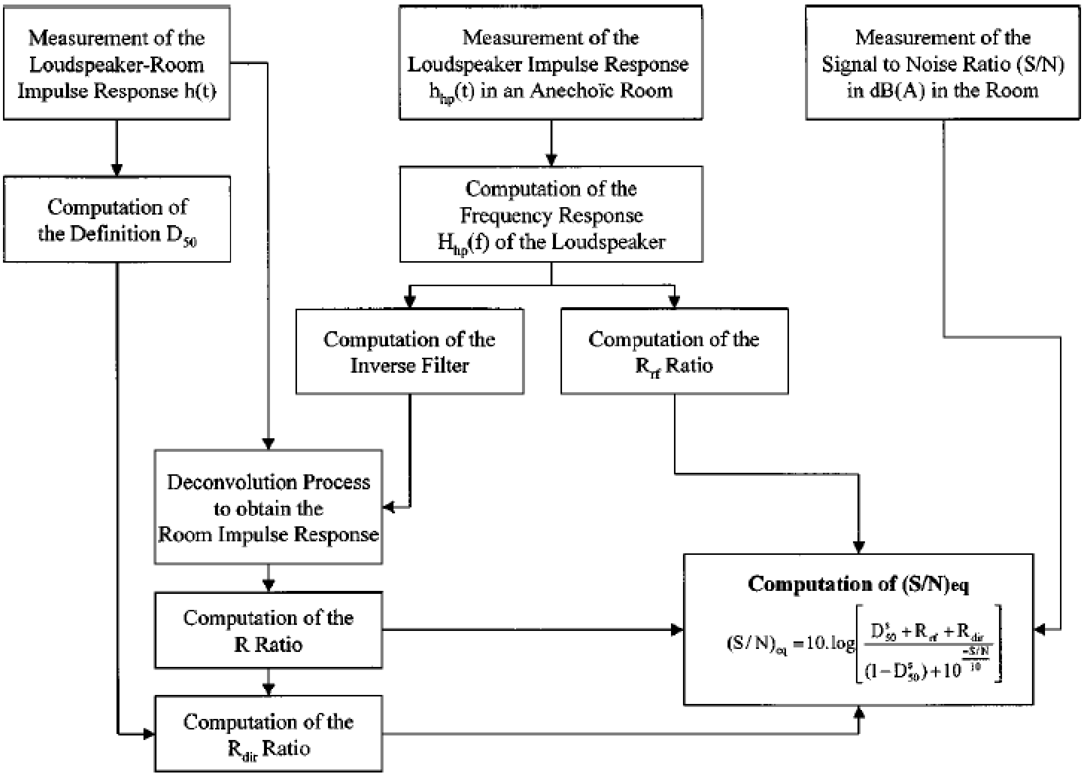

Measurement of the Loudspeaker-Room Impulse Response h(t)
Measurement of the Loudspeaker Impulse Response $h_{hp}(t)$ in an Anechoïc Room
Measurement of the Signal to Noise Ratio (S/N) in dB(A) in the Room
Computation of the Definition D_{50}
Computation of the Frequency Response $H_{hp}(f)$ of the Loudspeaker
Computation of the Inverse Filter
Computation of the R_{rf} Ratio
Deconvolution Process to obtain the Room Impulse Response
Computation of (S/N)eq
$(S/N)_{eq} = 10.\log\left[\frac{D^s_{50}+R_{rf}+R_{dir}}{(1-D^s_{50})+10^{\frac{-S/N}{10}}}\right]$
Computation of the R Ratio
Computation of the R_{dir} Ratio

Figure 11 : Method of (S/N)eq computation [26]

We obtain a correlation coefficient of 0.96 and a standard variation equal to 6.2% with 61 intelligibility measure score as illustrated in figure 12. The regression line is obtained according to the equation (41).

$$I(\%) = 100(1-10^{-[(S/N)_{eq}+40]/(60\times 0.18)})^{2203} \qquad (41)$$

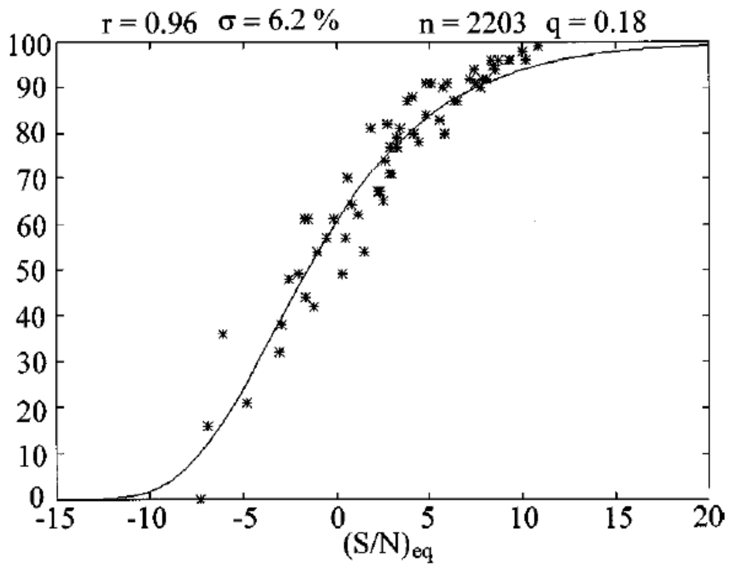


Figure 12 : measured speech intelligibility scores versus (S/N)eq predictor corresponding values and best least-squares fit [26]

This type of objective intelligibility measure is useful for I'CityForAll project view we work on loudspeakers enhancement will be needed to separate the effect of the loudspeaker with the room. Besides, it's important to combine different type of criteria that can be complementary to cover all the possible effect of degradation in the communication system.

Conclusion

The aim of the task 2.3 of I'CityForAll project is to provide firstly a means of optimizing and testing digital filters that are used to enhance sound for normal hearing persons and impaired hearing persons. We focused in this deliverable in the existing objective criteria of quality that can help to evaluate and improve quality "for all".

In the first part we focused on academic criteria that can be used quickly to optimize digital filters. We recommend using the so called frequency-variant weighted SNR that can be used to test digital filter for different kinds of quality, global quality or intelligibility, and for different population, normal hearing or impaired hearing.

In the second part of objective quality criteria, we focused on the standardized quality criteria that are accepted in the community. In fact, we are more interested by the derivatives of these standards like Speech Based STI and the Three Level Coherence SII. The Speech Based STI is the best suited to be adapted for real time computing and hearing impaired persons. Besides CSII can be used to take into consideration the nonlinear distortion added by the enhancement algorithm developed in this project.

Finally, we extend our investigation to the room acoustics criteria that is based on Early/Late received sound energy. We found that we can predict accurately the intelligibility using the RIR function. Indeed, we are interested in the separation of room and loudspeaker effects on intelligibility with help of equivalent signal to noise ratio $(S/N)_{eq}$ that predicts intelligibility with 0.96 correlation coefficient.

In addition, we propose to provide in the next deliverable a new global criterion that can predict intelligibility "for all" and takes into consideration the degradation of I'CityForAll algorithm, loudspeaker, room reverberation and background noise. This objective measure can be an improvement of standardized criteria plus a combination with room acoustics criteria.

References

[1] G. Miller and P. Nicely. An analysis of perceptual confusions among some english consonants. JASA, 27(2):338_352, 1955.

[2] W.D. Voiers. Evaluation processed speech using the diagnostic rhyme test. Speech Technol., January-February:30_39, 1983.

[3] Soli S. Nilsson, M. and J. Sullivan. Development of hearing in noise test for the measurement of speech reception thresholds in quiet and noise. J. Acoust, Soc. Am., 95(2):1085_1099, 1994.

[4] UIT-T Rec.P.800(1996), « Méthode de l'évaluation subjective de la qualité de transmission ».

[5] Tribolet, J., Noll, P. , et al, (1978), A study of complexity and quality of speech waveform coders, Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, pp. 586-590.

[6] T. P. Barnwell and W. D. Voiers, An analysis of objective measures for user acceptance of voice communication systems, final report, September 1979.

[7] A. Gray, Jr. and J.D. Markel, Distance measures for speech processing, IEEE Trans. Acoust., Speech and Signal Processing, vol. 24, pp. 380-91, October 1976.

[8] Klatt, D. (1982), Prediction of perceived phonetic distance from critical band spectra, Proc. IEEE Int. Conf. Acoust. Speech Signal Processing., vol 7, pp. 1278-1281.

[9] Philipos C. Loizou, Speech enhancement: theory and practice, CRC Press edition 2007, pp. 568-575.

[10] UIT-T Rec. P.862 (2001), « Evaluation de la qualité vocale perçue : méthode objective de l'évaluation de la qualité vocale ».

[11] A. W. Rix, Comparison between subjective listening quality and P.862 PESQ score, Psytechnics white paper, September 2003.

[12] H.J.M. Steeneken & T. Houtgast, « A physical method for measuring speech transmission quality », J. Acoust. Soc. Am. 67 (1), 1980.

[13] Norme NF EN 60268-16 « Équipements pour systèmes électroacoustiques. Partie 16 : évaluation objective de l'intelligibilité de la parole au moyen de l'indice de transmission de la parole », 1998.

[14] Schroeder, M, Modulation Transfer Function: Definition and Measurement, Acoustica, 49, 1981.

[15] T. Houtgast, H. Steeneken , et al, Past, present and future of the Speech Transmission Index, 2002.

[16] Payton KL, Braida LD (1999). A method to determine the speech transmission index from speech waveforms. J Acoust Soc Am 106: 3637-3648.

[17] Goldsworthy RL, Greenberg JE (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J Acoust Soc Am 116: 3679-3689.

[18] Ludvigsen C, Elberling C, Keidser G, Poulsen T (1990). Prediction of intelligibility of non-linearly processed speech. Acta Otolaryngol Suppl 469: 190-195.

[19] Karen L. Payton, Mona Shrestha, Evaluation of short-time speech-based intelligibility metrics, Communication Int. Cong. on Noise as Public Health Problem (ICBEN), 2008.

[20] ANSI S3.5-1997, Methods for calculation of the speech intelligibility index.

[21] James M. Kates, Kathryn H. Arehart, Coherence and the speech intelligibility index, JASA, V 117(4), pp 2224-2237, April 2005.

[22] J. P. A. Lochner and J. F. Burger, The influence of reflections on auditorium acoustics, J. Sound Vib. 1, 426–454 ~1964.

[23] S. Bradley, Predictors of speech intelligibility in rooms, JASA. 80, 837–845 ~1986.

[24] H. Fletcher and R. H. Galt, The perception of speech and its relation to telephony, JASA. 22, 89–151 ~1950.

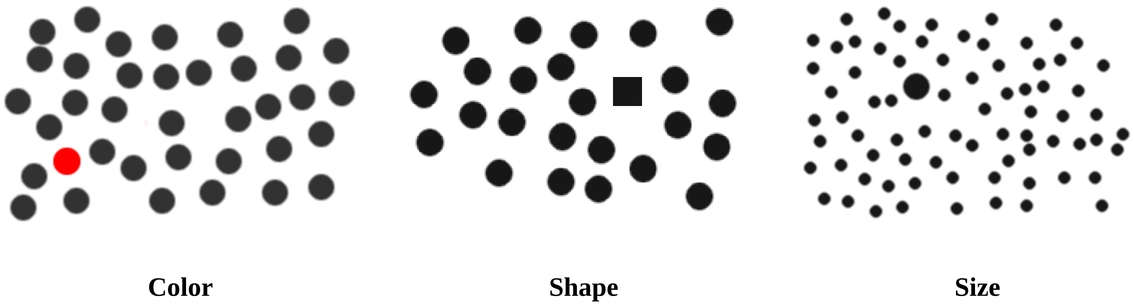
[25] D. Dirks, T. S. Bell, R. N. Rossman, and G. E. Kincaid, Articulation index predictions of contextually dependent words, JASA 80, 82–92 ~1986.

[26] L. Faiget & R. Ruiz, Speech intelligibility model including room and loudspeaker influences, JASA. 105 (6), 1999.

APPENDIX 1: ITTI AND KOCH VISUAL SALIENCY MODEL

One of the first method have been proposed by Itti and colleagues in 1998 [14]. This biologically-based model mimics the cortex behavior to analyse the picture through a set of three visual attributes: intensity (also called luminance), orientation and color.

Indeed, previous studies on visual perception had demonstrated that some stimuli will automatically and involuntarily attract our attention in a given context. For example, a red dinner jacket among black tuxedos in an official dinner will pop-out of the visual scene. During the last three decades, several studies have determined the different **pre-attentive visual features** that are responsible for this saliency effect (see [26] for a review). For instance, shape, luminance, color, orientation are some of these basic visual parameters and, as presented in Figure 13, a square among circles, a big circle among small ones or a vertical object among horizontal ones are all very salient objects. The Feature Integration Theory presented in [24] argues that all these basic features come before perception (early features), and are registered automatically in parallel across the visual fields. The main idea of computational modeling of visual saliency is to mimic this parallel processing by analyzing the visual scene in different feature dimensions.



Color **Shape** **Size**

Figure 13 Examples of pre-attentive visual features~: distinct objects automatically attract attention.

In the Itti and Koch model, the **features are extracted** in parallel at **various scales**.

The different scales are obtained through a dyadic gaussian pyramid (lowpass filtering and subsampling, see Figure 14) ranging from scale 0 (original image P_0, high resolution) to 8 (low resolution). The layer $P_(i+1)$ in the Gaussian pyramid is obtained by:

1. convolving layer P_i by a gaussian kernel like $\frac{1}{16}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix}$,
2. then subsampling it by a factor of 2.

The different scales are then compared using a **center-surround mechanism** to obtain six maps for each parameter. The "center" is a pixel at scale c $\in\{2,3,4\}$ while the "surround" is the same pixel at a coarser scale s = c + δ, δ $\in\{3,4\}$. This multi-resolution process is very close to Difference of Gaussian used to detect edges in a picture [18]. For example, if I(σ) is the intensity of the picture at scale σ, then the six intensity maps I(c,s) are computed through the Equation 1:

$$I(c,s) = |I(c) - I(s)| \qquad \textbf{Equation 1}$$

The six maps of a same feature are then average to form a feature map, for example for intensity:

$$\bar{I} = \frac{1}{6}\sum_{c,s} I(c,s)$$

Then the three feature maps ($\bar{I}$ for intensity, $\overline{C}$ for color, and $\overline{O}$ for orientation) are **normalized** to promote maps with a small number of strong peaks while globally suppressing maps with numerous comparable peaks. The local maxima of the map are compared to the global maximum of the same map. When the difference is large the map is strongly promoted, when the difference is small, the map contains nothing unique and is suppressed. For example, in figure Figure 15, the only interesting feature is orientation as all objects of the stimulus image have the same intensity. The normalization process therefore reduces the impact of the intensity map.

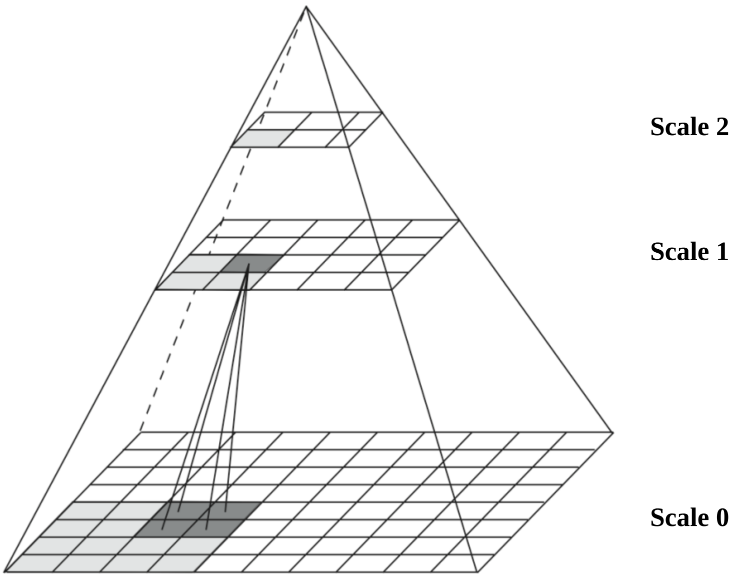


Figure 14. Pyramidal representation. Each pixel from a layer of the pyramid is generated from pixels of the previous layer through lowpass filtering. The filtered picture is then subsampled by a factor of 2. The base of the pyramid (scale 0) is the original image.

Mathematically, each original feature map $\overline{M_i}$ is scaled by a factor D_i to obtain the normalized feature map M_i^* (Equation 2). The factor D_i corresponds to the local *vs* global maxima comparison (Equation 3).

$$M_i^* = \overline{M_i} \times D_i$$

Equation 2

$$D_i = (G_i - \overline{L_i})^2$$

Equation 3

with G_i= Global peak of map *i*

and L_i= Local peaks of map *i*

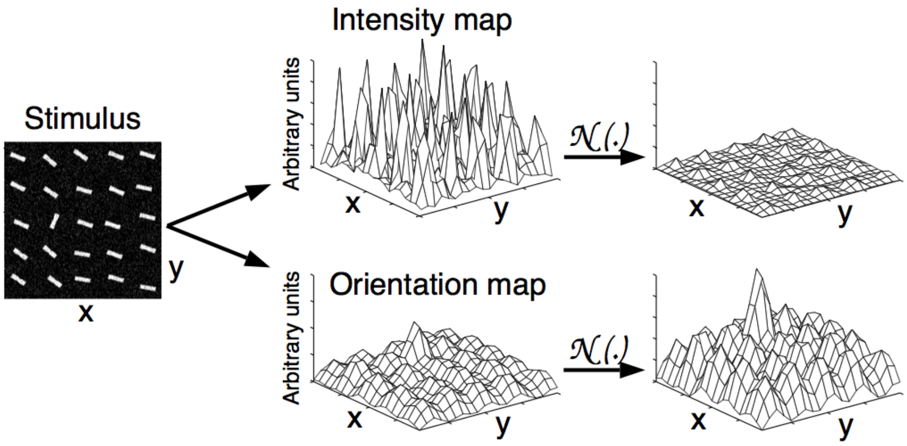


Figure 15. The normalization process applied to intensity and orientation maps in the case where the salient object is distinguished by its orientation only. From [14].

At the end, the normalized feature maps are combined in a two-dimensional saliency map corresponding to the salient points of the image. An example is presented in Figure 16. This model has been extended in auditory saliency models (see section 0).

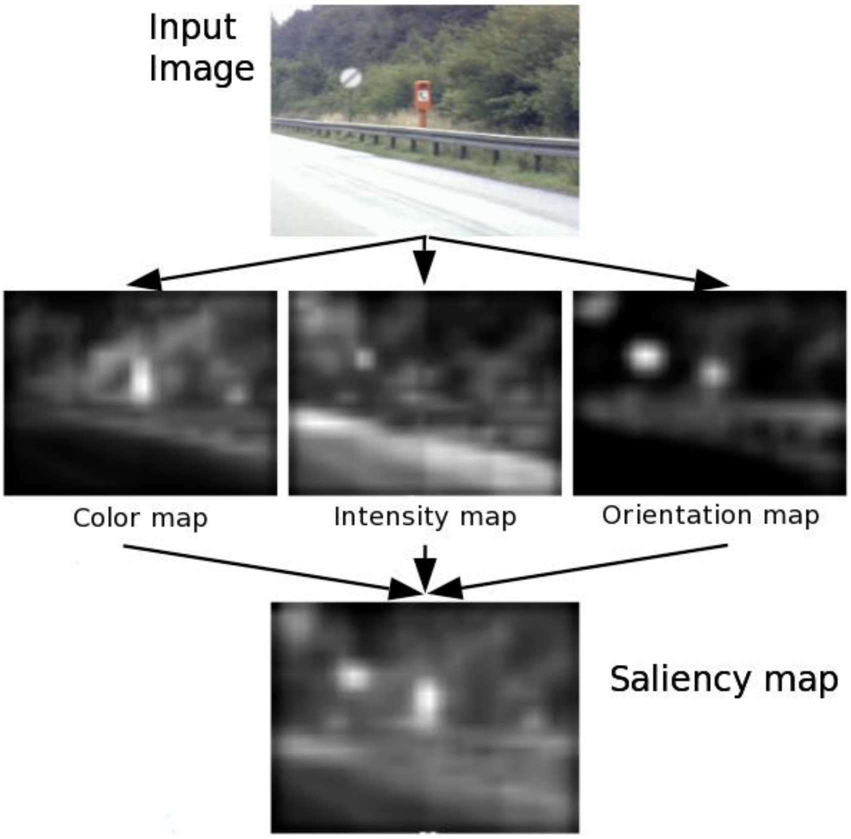


Figure 16. Example of saliency map and feature maps extracted from a photograph. From [14].

APPENDIX 2: MODEL OF AUDITORY SALIENCY BASED ON THE ITTI & KOCH MODEL

Since several saliency model have been developped for vision, an easy way to define auditory saliency model is to extend visual models to audio.

Three auditory models have been proposed that extend the visual saliency model of Itti & Koch [14] to the auditory model [16], [15], [7]. Basically the key idea is to obtain a visual saliency map from an "auditory image", i.e. a visual, spectro-temporal, representation of the sound (a spectrogram in [16], an auditory spectrogram or cochleogram in [15] and [7])

The first auditory saliency model was proposed in 2005 by [16]. It is very closed to the visual saliency model of Itti & Koch and relies on a very similar structure with : first, a parallel extraction of different features at different scales, then a center-surround differentiation substracting the coarser to the finer scale to obtain different feature maps, and finally a normalization and a linear combination of these feature maps to obtain the final saliency map.

The only differences between those two models are the feature detectors and the normalization:

- features extracted in the visual models are those extracted by the visual cortex like luminance contrast, orientation or color. The auditory features used in auditory scene perception to distinguish the different sources are related to spectral or temporal modulations. However, although visual and auditory features differ in their interpretation, mathematically they are very similar: the filters used to extract frequency or temporal contrast in audio can be interpreted as detectors for horizontal and vertical orientations in vision and the filters used to extract sound intensity are identical to those extracting luminance.
- while the visual saliency model is applied on still images, the auditory saliency model incorporates a temporal component. The normalization procedure (used to promote feature with few but highly conspicuous peaks) is therefore adapted so that causality restrictions imposed by the temporal domain are incorporated in a sliding window normalization. Based on known properties of forward and backward masking, the sliding window used to compute local and global maxima for the normalization is asymmetric, extending 225 msec into the past and 75 msec into the future.

Kalinli et al. [15] based their model on the same structure (see Figure 17) but suggested to add two more auditory features (pitch and orientation). Furthermore the spectrogramm used in this model mimics the early auditory processing, so is slightly different from the original spectrogram used in Kayser's model.

While Duangudom argue that different features are employed in her model, these features are still related to intensity or frequency and temporal modulations. The main improvement of this model compared to Kalinli and Kayser models is therefore the use of a pre-processing stage that consists in applying a weighting frequency filters before computing the auditory spectrogram to reinforce the contribution of key dimensions: present frequencies within the sound and loudness [7].

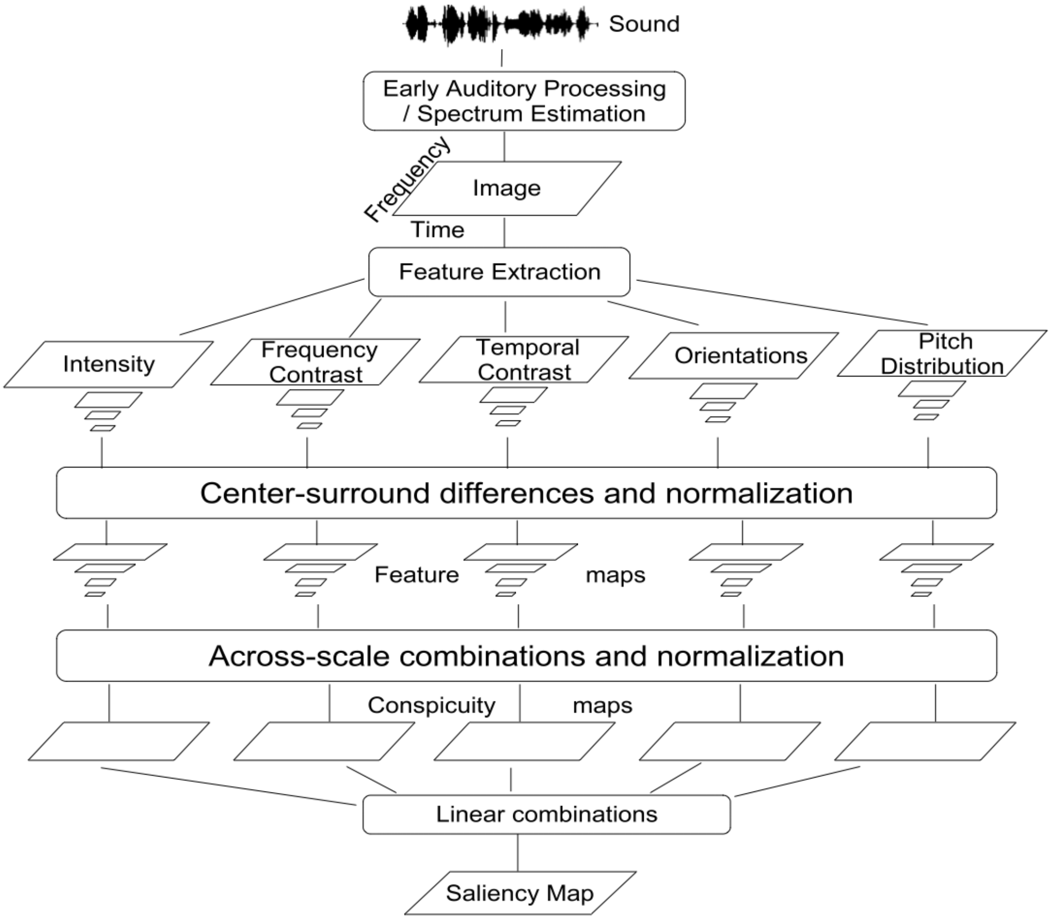


Figure 17. Auditory saliency map structure of [15] adapted from [16].

An implementation of the Kayser model is temporarily available online at the address: http://transfert.u-psud.fr/download.php?file=35ImplementationKayser_forICityForAllReport.zip

APPENDIX 3: NATURAL STATISTICS FOR VISUAL AND AUDITORY SALIENCY MEASURES

Contrarily to the center-surround mode, the Saliency Using Natural statistics mode [27] requires a comparison with other images. The salience at any point of the picture is still based on the rarity of the feature responses at that point but rarity is computed on statistics, not only depending on statistics of the particular image being viewed (difference with neighbors), but also derived from natural image statistics obtained in advance from a large collection of natural images. The SUN model uses only one single feature map, learned using Independent Components Analysis (ICA) of each natural images of the database.

This model has been extended in an **Auditory Salience Using Natural statistics** model (ASUN). This extension is based on a learning approach that compares a temporal frame with the recent past frames ("local statistics") and with long past frames ("lifetime statistics").

The features used to analyze the signal are not defined in advance as in other auditory saliency models but are directly computed by the system through a Principal Component Analysis (PCA). Three stages, summarized in Figure 18, are necessary to obtain the features. As for previous auditory models, the first stage is a conversion from the audio signal to a visual, spectro-temporal representation of the sound. The ASUN model relies on a cochleogram, close to the auditory spectrogram used in [15] obtained by applying a 64-channel gammatone filterbank (i.e. decomposition to 64 frequency dimensions). The second stage is to split the cochleogram in several patches of width representing 8 ms and height representing one frequency band regrouping 7 frequency dimensions from the 64 frequency dimensions of the cochleogram (overlap of 8 samples and 4 frequency dimensions). The third stage consists of calculating the two or three first components through a PCA for each patch.

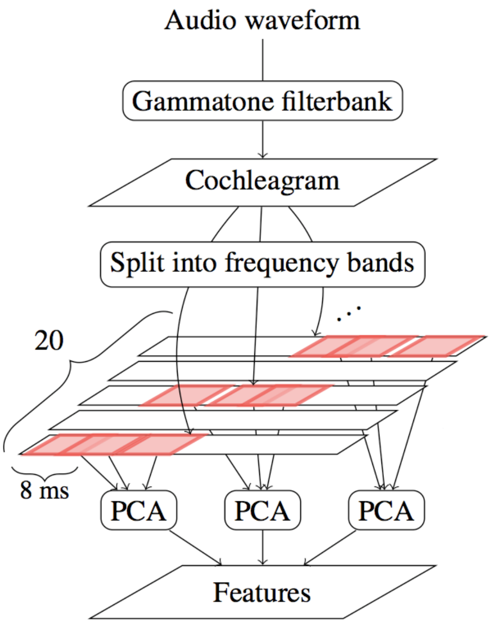


Figure 18. Schematics for the feature extraction procedure in the ASUN model. Input signals are first converted to smoothed cochleagram which is then separated into 20 bands of 8 ms patches. The number of dimensions of each band is reduced through PCA. From [25].

If F_t is the vector representing the features responses of the signal at time *t*, the saliency *s(t)* at t can be define as the rarity in relation to the recent past (from the input signal) as well as to the long-term past beyound a delay *k* as in equation Equation 4 also equivalent to Equation 5 under the assuption of independance between local and lifetime statistics.

$$s(t) \propto -\log P(F_t = f_t \mid \underbrace{F_{(t-1)},\ldots,F_{(t-k)}}_{recent\ past}, \underbrace{F_{(t-k-1)},\ldots}_{long\ past})$$

Equation 4

$$\begin{aligned} S(t) \quad \propto \quad & -\log P(F_t = f_t \mid F_{(t-1)},\mathrm{K}\ ,F_{(t-k)}) \\ & -\log P(F_t = f_t \mid F_{(t-k-1)},\mathrm{K}\ ) \\ = \quad & s_{local}(t) + s_{lifetime}(t) \end{aligned}$$

Equation 5

The probability of feature occurence $P(F = f_t)$ is computed based on prior experience. The lifetime statistics are computed through Independant Component Analysis on 1200 seconds of sound samples randomly chosen from a large database including environmental sounds (animal, urban...) and speech sounds. The local statistics was estimated using the same method considering at each time step *t* the probability distribution of the input signal from *t-k* to *t-1*. Unfortunately, due to computational limits, the local statistics are only computed with discontinuity every *k*=250 msec.

APPENDIX 4: DISCRETE ENERGY SEPARATION ALGORITHM

Unlike previous methods inspired from the center-surround and SUN visual models, the **Discrete Energy Separation Algorithm** (DESA) is not based on the computation of a visual saliency map from a visual representation of the sound. It considers only the temporal modulations of amplitude and frequency in multiple frequency bands. Developped by Evangelopoulos and colleagues, it is employed for different applications like video summarization (detection of salient event in the movie) [10] and speech detection in noise [9]. According to [5], DESA is a very popular measure as it is very easy to compute.

For each temporal frame *m* (40~ms), the input signal *s* is separated in several frequency bands through Gabor filters (6 bands centered on 281, 562, 1125, 2250, 4500, and 9000 Hz). Then, for each frequency band and every sample *k* of the frame, the Teager-Kaiser energy is obtained with the Equation 6.

$$\Psi[s[k]] = s^2[k] - s[k+1]s[k-1]$$

Equation 6

The frequency band that maximizes the Teager-Kaiser energy for this sample *k* is selected before computing two other measures on this frequency band : the instant amplitude (Equation 7) and the instant frequency (Equation 8).

$$a(s[k]) = 2\frac{\Psi(s[k])}{\sqrt{\Psi(s\acute{Y}[k])}}$$

Equation 7

$$f(s[k]) = \frac{1}{2\pi}\arcsin\left(\sqrt{\frac{\Psi(s\acute{Y}[k])}{4\Psi(s[k])}}\right)$$

Equation 8

with $\dot{s}$ the derivative of the signal *s*.

Each feature (Teager-Kaiser energy, instant amplitude and instant frequency) is then averaged over all the audio samples of the frame, so three global features are obtained for each frame : the Mean Teager Energy (MTE), the Mean Instant Amplitude (MIA) and the Mean Instant Frequency (MIF). They are normalized independantly to be in the range [0;1] and finally combined to compute the auditory saliency value *S* of the current frame m with the Equation 9.

$$S(m) = \omega_1 MTE(m) + \omega_2 MIA(m) + \omega_3 MIF(m)$$ **Equation 9**

The weightings ω_i can be identical or adapted to promote one of the features. If the analyzed sound contains great energy variations, one is likely to give a preferential weighting to the MTE. On the contrary, if the sound is more based on frequency variations, as in a moving police siren presenting Doppler effect, the MIF will be preferred.

The complete procedure is summarized in Figure 19. Even if the DESA measure is very correlated to human ratings of saliency [5], it is not usable for real time measurement since it relies on a comparison between past and future energy.

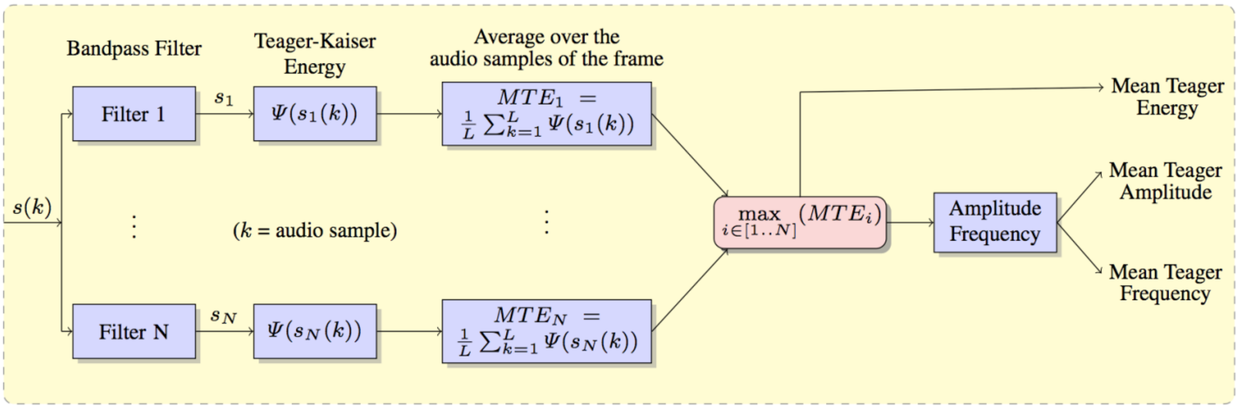


Figure 19. Global procedure of the DESA model. From [5]

1.1. Other models

Saliency is also used in robotics where both visual and auditory saliency maps are associated. However, in such cases, the auditory saliency map only corresponds to the position of the sound source as in the ICub project [20]. The auditory saliency computation is in that case equivalent to sound source separation algorithm. It does not consider complex scenes where different sources can be listened in the same time at different saliency levels.

APPENDIX 5: EXPERIMENTAL METHODOLOGIES TO EVALUATE AUDITORY SALIENCY MODELS

Evaluating the efficiency of an auditory saliency model is much more difficult than the evaluation of visual saliency models as no audio equivalent of eye-tracking is available to directly track a physical correlate of auditory saliency. To evaluate how well models can predict listeners behaviour, saliency values obtained with the models are compared to subjective performances or indirect ratings of different more or less adapted tasks.

1.2. Direct ratings

A very simple solution to measure how participants will perceive the saliency of different sounds is to directly ask them to rate the saliency of these sounds. For example, in [8] participants had to listen to two different auditory stimuli (sequential presentation), each composed of one auditory scene/background + one specific sound, and then decide which of the two stimuli they find more salient in a **two alternative forced choice task** (**2AFC**). They had also to rate the difference of saliency on a scale from 1 (equal salience) to 7. The same procedure was previously employed in [16]. This methodology is problematic as experimentators are forced to give a definition of saliency and then to consider that every partipant understand the saliency definition in the same way. Alternatively, the question posed to participant can be slightly modified to avoid the explanation of the word "saliency". In [25], the authors still used a 2AFC paradigm but asked participants to choose "the most interesting sound" instead of the most salient one. The problem of a common definition across participants persists.

1.3. Detection tasks

The auditory saliency map predicts which sounds or features of a complex auditory scene will naturally capture our attention and, hence, are more easily detected, even for low signal-to-noise ratio. By varyating the sound level relatively to the background noise level (naturalistic ambiant sound or gaussian noise), it is possible to determine the detection threshold that would be highly correlated to the attractiveness of that sound. According to a group meeting about auditory saliency [23], this **detection task paradigm** is the most employed technique to evaluate the perceived auditory saliency. This procedure is for example employed in [16].

1.4. Dual-task experiments

A common procedure employed in experimental psychology to measure the cognitive load induced by a task is the **dual-task paradigm**. It consists in having users engaged in two tasks simultaneously. This method relies on the assumption that we have limite**Error! Reference source not found.**d-capacity resources so, *"when a great deal of cognitive capacity is consumed by the primary task, less capacity is available to devote to the secondary task"* [4]. The mental effort measure is therefore obtained by comparing single-task performances to dual-task performances.

The dual-task paradigm is very promising for the I'City For All project for two different reasons. The first one is because dual-task experiments are frequently used to measure the effect of age or hearing problems on auditory abilities, or more generally to assess cognitive load in speech listening. While normal intelligibility test sometimes reveal no difference in word recognition performance between normal and hearing-impaired listeners, dual-task experiments hightlight that, to achieve the same performances, hearing-impaired listeners just allocate more cognitive ressources. This of course lead to a non desired auditory fatigue. A review of dual-task experiments for **assessing listening effort** is available in [12]. Although various tasks can be used, the procedure is very similar from a study to another. The primary task concerns the listening activity (e.g. speech recognition test in quiet or in different SNR) while the secondary task may be in the same modality (auditory memory task/ recall) or in another one (reaction to a light probe or to colour change, reaction to a tactile pattern). Participants are told to focus on the word recognition task while any additional tasks have to be considered secondary.

Moreover, dual-task experiment are of main interest to evaluate auditory saliency. As perceptual processes driven by saliency do not require attention, they occur very rapidly and effortlessly. Therefore the mental effort required to achieve a task should be reduced if salient sounds are involved. An attempt to use dual-task paradigm for measuring saliency perception was recently presented in [7]. Participants achieved two auditory tasks in parallel. The first one, requiring high cognitive processing, consist in counting how many low frequency tones (100~Hz) appear in a sequence of 25 tones (100~Hz or 200~Hz). The second task consists in detecting the presence of a modulated tone among four tones (*present/absent* experiment). As modulation is one of the main feature of auditory saliency, this second task is supposed to be of reduced cognitive load. This main limitation of this study is in the use of non-ecological laboratory stimuli (pure or modulated tones) for both tasks. To evaluate the efficiency of the auditory saliency models on our applied project of vocal announces, we suggest to define more complex tasks using speech stimuli.

Finally, some variant of dual-task experiment are employed for studying selective attention in multi-talker listening. In [19], the authors measure the cortical activity with electrodes to determine which of two simultaneous speakers is actually attented. After a calibration step using the TIMIT corpus [11]. the actual experiment is based on a **reaction to a target call-sign** relying on the Coordinate Response Measure (CRM) [1] corpus. The CRM corpus contains sentence in the form "ready (*call sign*) go to (*color*)(*number*)". This form of corpus is very similar to corpus formed on matrices, allowing a huge amount of different sentences with the same grammar and duration. Participants had to listen to two voices in parallel, one male and one female, and report the number and the color associated with the target call-sign pronounced by only one voice.

1.5. Proposition

We suggest to use a dual-task paradigm mixing the experiment of [19] and the dual-task experiment of [7]. The idea is to consider that a salient vocal announce will be distractive for a continuous task like reading or speaking. Therefore we propose that the peripheral task would correspond to a reaction to target call-sign, very similar to the one used in CRM corpus. The call-sign will be a destination and messages will mimics the transport vocal announces (e.g. "Train for (*Destination*), departure at (*time*) platform (*letter*)"). In parallel of this peripheral task, participants will have to perform a very engaging primary task with an attentional cost as high as reading or speaking with someone. This kind of experiment would allow us to 1/ use more realistic stimuli than the pure tone signals used by Duangudom 2/ use speech as a simulus to evaluate auditory saliency models.

APPENDIX 6: SOURCE CODES OF DIFFERENT VISUAL SALIENCY MODELS

Generally the implementations of visual saliency models can be downloaded from the authors' website. A list of saliency methods and associated codes referenced in the state-of-the-art is published at http://cg.cs.tsinghua.edu.cn/people/~cmm/saliency/.

The code of the SUN model from Lingyun Zhang [27] is available under Matlab format at http://cseweb.ucsd.edu/~l6zhang/code/imagesaliency.zip.

A fast computation of the SUN model [3] has been implemented and is available at http://mplab.ucsd.edu/~nick/NMPT/main.html, as part of the Nick's Machine Perception Toolbox (NMPT). Coded in c++ language, it nevertheless requires the OPEN CV library.

The model of Attention based on Information Maximization (AIM model) from [2] is also available: http://www.cs.umanitoba.ca/~bruce/datacode.html.

The simple Matlab implementation of the spectral residual approach [13] allows a fast computation of a visual saliency map in the spectral domain:

```
clear

clc

%% Read image from file

inImg = im2double(rgb2gray(imread('yourImage.jpg')));

inImg = imresize(inImg, 64/size(inImg, 2));

%% Spectral Residual

myFFT = fft2(inImg);

myLogAmplitude = log(abs(myFFT));

myPhase = angle(myFFT);

mySpectralResidual = myLogAmplitude - imfilter(myLogAmplitude,
... fspecial('average', 3), 'replicate');  % imfilter function is part of
the image processing toolbox

saliencyMap = abs(ifft2(exp(mySpectralResidual + i*myPhase))).^2;

%% After Effect

saliencyMap = mat2gray(imfilter(saliencyMap, fspecial('gaussian', [10, ...
10], 2.5)));

imshow(saliencyMap);
```

[1] Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (2000). A speech corpus for multitalker communications research. J. Acoust. Soc. Am., 107:10651066.

[2] Bruce, N. and Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. Journal of Vision, 9(3):1–24

[3] Butko, N. J., Zhang, L., Cottrell, G. W., and Movellan, J. R. (2008). Visual salience model for robot cameras. In Proceedings of the 2008 IEEE International Conference on Robotics and Automation (ICRA), page 23982403, Pasadena, CA, USA

[4] Cennamo, K. S. (1993). Learning from video: Factors influencing learners preconceptions and invested mental effort. Educational Technology Research and Development, 41(3):33–45.

[5] Coutrot, A., Guyader, N., Ionescu, G., and Caplier, A. (2013). Video viewing: do auditory salient events capture visual attention? Annals of Telecommunications, in press:1–9.

[6] De Coensel, B. and Botteldooren, D. (2010) A model of saliency-based auditory attention to environmental sound. In Proceedings of the 20th International Congress on Acoustics (ICA), Sydney, Australia.

[7] Duangudom, V. (2012). Computational auditory saliency. PhD thesis, Georgia Institute of Technology.

[8] Duangudom, V. and Anderson, D. V. (2007). Using auditory saliency to understand complex auditory sceneseuropean signal processing conference (eusipco 2007. In European Signal Processing Conference (EUSIPCO 2007, Poznan, Poland.

[9] Evangelopoulos, G. and Maragos, P. (2006). Multiband modulation energy tracking for noisy speech detection. IEEE Transactions on Audio, Speech, and Language Processing, 14(6):2024–2038.

[10] Evangelopoulos, G., Zlatintsi, A., Skoumas, G., Rapantzikos, K., Potamianos, A., Maragos, P., and Avrithis, Y. (2009). Movie summarization based on audiovisual saliency detection. In Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-09), Taipei, Taiwan.

[11] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). The timit acoustic-phonetic continuous speech corpus. Linguistic Data Consortium.

[12] Gosselin, P. A. and Gagn, J.-P. (2010). Use of a dual-task paradigm to measure listening effort. Revue canadienne d'orthophonie et daudiologie -, 34(1):43–51.

[13] Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR '07).

[14] Itti, L. and Koch, C. (1998). A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Learning, 22:12541259.

[15] Kalinli, O. and Narayanan, S. (2007). A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In Interspeech.

[16] Kayser, C., Petkov, C. I., Lippert, M., and Logothetis, N. K. (2005). Mechanisms for allocating auditory attention : An auditory saliency map. Current Biology, 15:194–1947.

[17] Lu, S. and Lim, J.H. (2012). Saliency Modeling From Image Histograms. European Conference on Computational Vision. Florence, Italy

[18] Marr, D. and Hildreth, E. (1980). Theory of edge detection. Proceedings of the Royal Society of London. Series B, Biological Sciences, 207:215217.

[19] Mesgarani, N. and Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485:233236.

[20] Ruesch, J., Lopes, M., Bernardino, R., Hrnstein, J., Santos-victor, J., and Pfeifer, R. (2008). Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In IEEE International Conference on Robotics and Automation (ICRA 2008).

[21] Schauerte, B. and Stiefelhagen, R. (2013). "Wow! Bayesian Surprise for Salient Acoustic Event Detection". In Proc. 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, May 26-31.

[22] Schauerte, B., Kühn, B., Kroschel, K., Stiefelhagen, . R. (2011) "Multimodal Saliency-based Attention for Object-based Scene Analysis". In Proc. 24th International Conference on Intelligent Robots and Systems (IROS), IEEE/RSJ, San Francisco, CA, USA.

[23] Telluride Attention Team (2011). Discussion comparing auditory saliency models. In Tel luride Neuromorphic Cognition Engineering Workshop.

[24] Treisman, A. and Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12:97–136.

[25] Tsuchida, T. and Cottrell, G. W. (2012). Auditory saliency using natural statistics. In the Annual Meeting of the cognitive science society COGSCI 2012.

[26] Wolfe, J. M. and Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience, 5:1–7.

[27] Zhang, L., Tong, M. H., Marks, T. K., Shan, H., and Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. Journal of Vision, 8:1–20.

Audio Sharpness Index

1. Principle and definition of the Sharpness Index

A Sharpness Index (SI) was recently proposed in 0 for image processing, providing a simple and efficient objective measure of the perceived sharpness of an image, though this was not assessed through formal subjective tests. The SI has the advantage of being a non-intrusive measure, so that it can be used as an optimization criterion in blind denoising/debluring algorithms.

The principle of the SI is to measure the sensitivity of the total variation of an image (actually any regularity measure) to the convolution of the image by a white gaussian noise. The sharper is an image, the more sensitive is its total variation.

This principle can be applied to sound processing, adapting the formulas of 0 to a one-dimensionnal signal. We define the SI of a sound *u* of length *N* samples as:

$$SI(u) = -\log_{10}\Phi(\frac{\mu - TV(u)}{\sigma})$$

where Φ denotes the complementary error function, *TV(u)* denotes the total variation of *u*, defined as:

$$TV(u) = \sum_{n=1}^{N-1}(|u(n) - u(n-1)|)$$

and μ and σ^2 are defined respectively by :

$$\mu = \sqrt{\frac{2N}{\pi}}\|\partial u\|_2$$

$$\sigma^2 = 2\frac{\|\partial u\|_2^2}{\pi}\sum_{m=0}^{N-1}\omega\left(\frac{R_{\partial u}(m)}{\|\partial u\|_2^2}\right)$$

where:

- $\partial u(n) = u(n+1) - u(n)$;
- $\|\partial u\|_2^2 = \sum_{n=0}^{N-1}\partial u(n)^2$;

- $R_{\partial u}(m) = DFT^{-1}\left[\left|DFT(\partial u)\right|^2\right]$ (DFT denoting the Discrete Fourier Transform);
- $\forall t \in [-1,1],\ \omega(t) = t \ \ \arcsin\ t + \sqrt{1-t^2} - 1$

2. The SI as clarity measure ?

We tested the SI on various music and speech signals, impaired by different levels of noise, reverberation and low-pass filtering, known to reduce the sound clarity. We will present here the results for speech corrupted by noise and reverberation, since these are the main distortions to considered in our application. In each case, we will compare the behavior of the SI to that of the Speech Transmission Index (STI), which was proved to be well correlated with speech intelligibility [23].

When speech is corrupted by white noise, as illustrated by Figure 20, the SI has the same variations as the STI, but in a different range of SNRs (10 to 40 dB instead of -15 to 15). This could be interesting to measure the intelligibility for hearing-impaired people, that may be reduced at SNRs greater than 15 dB, where the STI saturates.

In the case of reverberation, as illustrated by Figure 21, the SI is a decreasing function of the reverberation time *T60* in a similar manner as the STI, though decreasing faster. Again, this sharper decrease could be an advantage in the case of hearing-impaired people, who are more sensitive to the increase of reverberation.

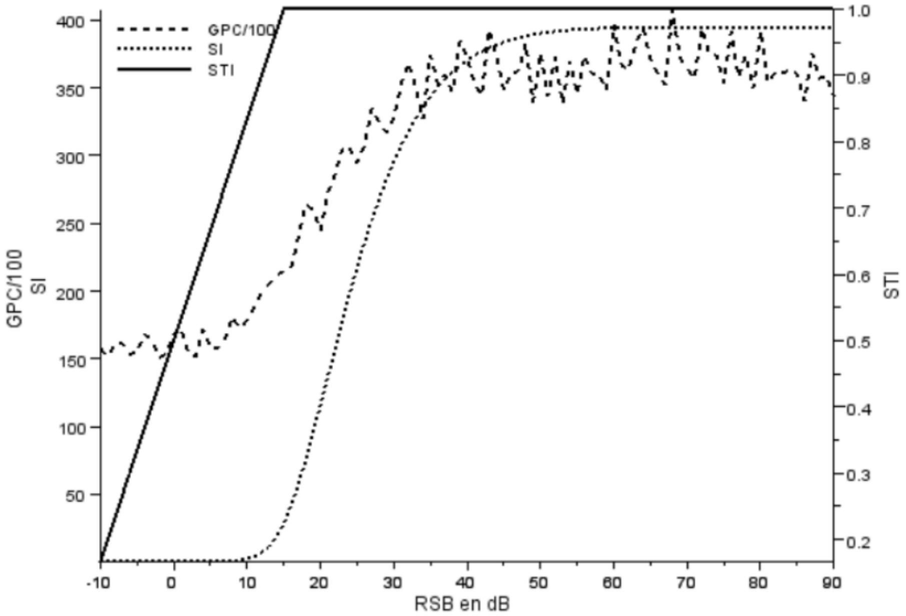


Figure 20. SI and STI according to the SNR.

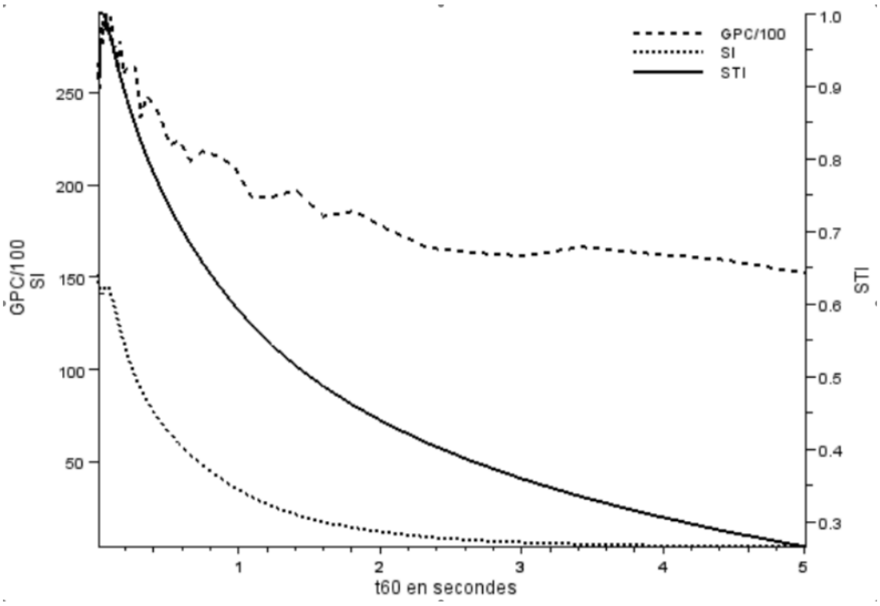


Figure 21. SI and STI according to the reverberation time T60.

The combined effect of noise and reverberation on the STI and on the SI is shown on Figure 22 and Figure 23. Note that the T60 range is the same, whereas the SNR range is adapted to the sensitivity of each index.

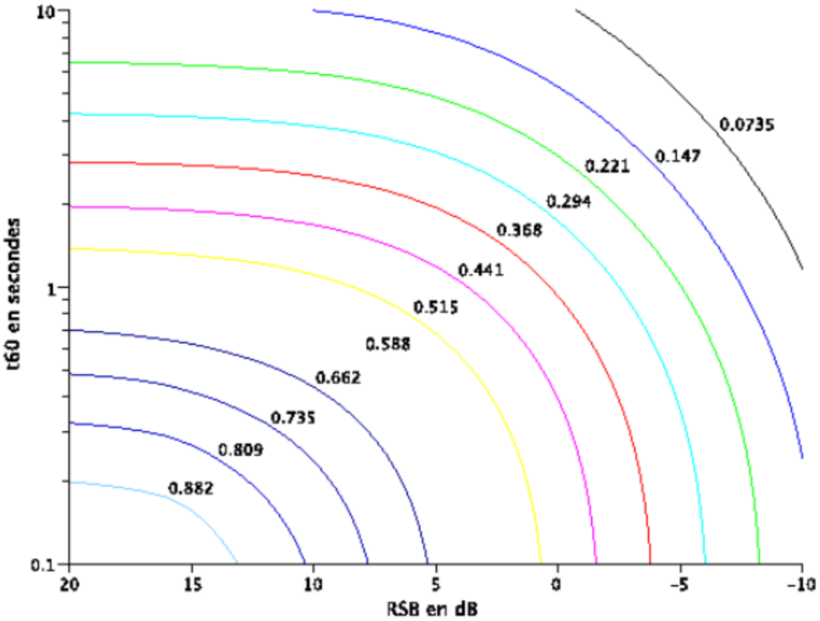


Figure 22. Contour of STI according to T60 and SNR.

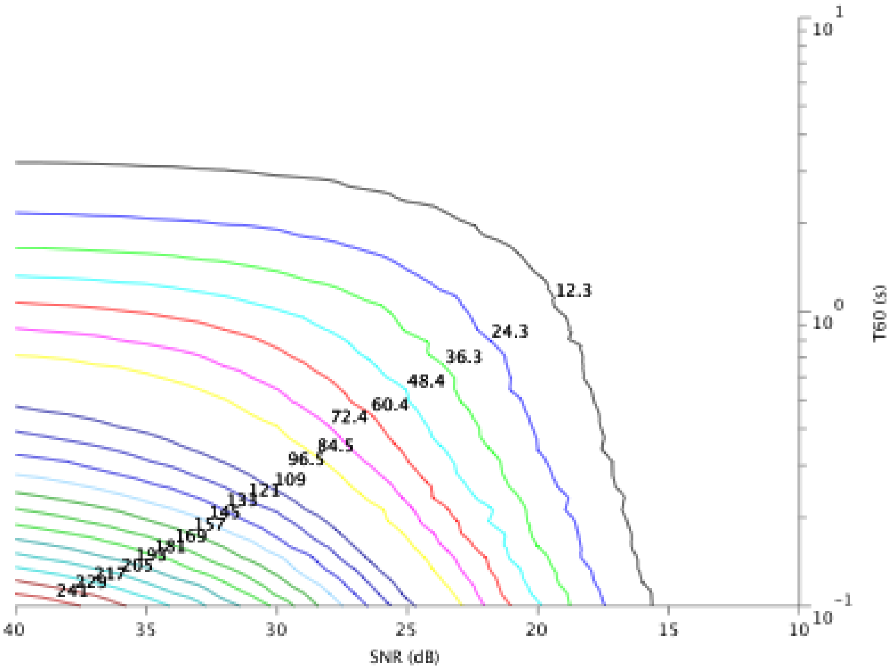


Figure 23. Contour of SI according to T60 and SNR

Conclusions

Since the STI is known to be well correlated to the intelligibility of speech, the contour of SI compared to that of STI shows that the SI is probably not relevant as an intelligibility measure, at least for normal-hearing people. However, Figures Figure 20Figure 21 show that the SI decreases as factors of intelligibility loss increase. Consequently, the SI could be a good criterion to maximize in an intelligibility-enhancement algorithm. It has two main advantages:

- its computational complexity is low;
- it does not require the original signal, which avoids the problem of synchronization between test and reference signals in classical algorithms.

Scilab source code of the SI function

```
function SI = SharpnessIndex(s)

    exec("omega.sci");

    exec("TotalVariation.sci");

    N = length(s);

    N_tot = 2^ceil(log2(N))*2;

    TVs = TotalVariation(s);

    ds = [s(2:$) s(1)] - s;

    ds = [ds zeros(1,N_tot-N)];

    ds22 = sum(ds.^2);

    Rds = real(ifft(abs(fft(ds)).^2));

    Rds = max(-ds22,min(ds22,Rds)); // to avoid small
                                  exceedings

    E_TV_S = sqrt(2*N*ds22/%pi);

    Var_TV_S = 2*ds22*sum(omega(Rds/ds22))/%pi;

    if SI>15 then

        v = (E_TV_S-TVs)/sqrt(2*Var_TV_S);

        SI = v^2/log(10) + log10(v) + 0.5*log10(%pi);

    end

endfunction
```

with :

```
function TV = TotalVariation(x)

    TV = sum( abs(x(2:$)-x(1:$-1)) );
```

```
endfunction



function w = omega(x)

    w = x.*asin(x) + sqrt(1-x.^2) - 1;

endfunction
```

References

[1] G. Blanchet and L.Moisan (2012) *An explicit sharpness index related to global phase coherence,* in Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference, pp.1065-1068.

[2] T. Houtgast and H.J.M Steeneken (1985) 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*, pp. 1096-1077.