Intelligibility assessment in forensic applications

Giovanni Costantini[1,2], Andrea Paoloni[3], Massimiliano Todisco[1,3]

[1]Department of Electronic Engineering, University of Rome “Tor Vergata”, Rome, Italy

[2]Institute of Acoustics “O. M. Corbino”, Rome, Italy

[3]Fondazione “Ugo Bordoni”, Rome, Italy

E-mail: costantini@uniroma2.it, pao@fub.it, massimiliano.todisco@uniroma2.it

Abstract

In the context of forensic phonetics the transcription of intercepted signals is particularly important. However, these signals are often degraded and the transcript may not reflect what was actually pronounced.

In the absence of the original signal, the only way to see the level of accuracy that can be obtained in the transcription of poor recordings is to develop an objective methodology for intelligibility measurements.

This study has been carried out on a corpus specially built to simulate the real conditions of forensic signals. With reference to this corpus a measurement system of intelligibility based on STI (Speech Transmission Index) has been evaluated so as to assess its performance. The result of the experiment shows a high correlation between objective measurements and subjective evaluations. Therefore it is recommended to use the proposed methodology in order to establish whether a given intercepted signal can be transcribed with sufficient reliability.

Keywords: objective intelligibility, forensic phonetics, speech transmission index, transcript reliability

1. Introduction

Intelligibility of speech refers to the amount of speech items that a normal listener can understand. More specifically the standard ISO 99 21 defines intelligibility as "the measurement of effectiveness in understanding speech." Intelligibility can be assessed at sentence level, at word level, and for each phoneme.

Intelligibility plays a key role in communications; indeed, ensuring full intelligibility is the main purpose of any communication channel or any recording system.

In forensic applications it is crucial that the meaning of sentences and mentioned names reflect those actually uttered by the speakers rather than the views of the transcribers.

In many cases, however, there are harsh contrasts between the prosecutor and the defender about the transcription of poor recordings. (Fraser, 2003)

To assess the reliability of a transcript it would be useful to have a measure of the intelligibility of the signal to be transcribed. Unfortunately, no subjective measurement can be used in forensic applications, because the content of the message is not known in advance, and therefore it is impossible to determine the percentage of words that have been accurately transcribed.

The only way to assess the intelligibility in forensic applications is to set up a system based on acoustic parameters which is able to predict the intelligibility of the measured signal.

Such a system in the forensic field would be also very useful for evaluating the performance of speech enhancement systems, and more generally, would be very useful in many other fields to avoid the high cost of the subjective evaluation of signal intelligibility.

In a previous paper (Costantini, 2010) we proposed an implementation of the STI (Speech Transmission Index) and verified a good agreement between the data provided by the STI and subjective measures of intelligibility in a series of experiments based on phoneme recognition.

Since in real applications on intercepted signals, especially in eavesdropping, the signal exhibits both adaptive noise (background noise) and multiplicative noise (reverb), we thought it useful to carry out an additional experiment which would verify the correlation between the measures obtained with STI and subjective listening results on the same audio material.

The present paper is organized as follows: section 2 describes the corpus used in the experiment; section 3 describes the organization of the subjective tests; section 4 presents the objective measurements methodology; section 5 summarizes and discusses the results; conclusions and comments are provided in section 6.

2. Speech Corpus

Both subjective and objective tests are conducted using the corpus collected during the European project SAM EUROM 1 (Chen, 1995). In particular 24 Italian, meaningful or meaningless, sentences, have been used.

Degradations considered include additive Babble noise and convolutive reverberation disturb.

The noisy speech appeared in three different grades of signal to noise ratio (S/N = +4dB, 0dB, -4dB) each with two types of reverb (T60 = 0.95*s* and 2.03*s*), used to simulate Office and Lobby environment, so we obtain six differently degraded signals. Each sentence is read by 4 different voices: two men and two women.

At the end of operations, therefore, can be found to have 24 different signals, each formed by different sentences. Table I shows the complete speech corpus: each S/N ratio with its reverberation rate is evaluated by 4 different phrases, spoken by different voices.

3. Subjective Intelligibility Assessment

A first experiment was conducted to obtain intelligibility scores using subjective listening tests. The speech corpus was submitted to a group of 24 normal-hearing listeners using software developed to this purpose under the Max/MSP (Cycling74) environment, that randomly delivers each item many times to the listener agree. One test set consists of 24 different test signals. The listener fills in the proper space the sentence he/she has heard. Fig. 1 shows the application interface used for test. The averaged results of the subjective tests, regarding sentences, words and phonemes are shown in Fig. 2.

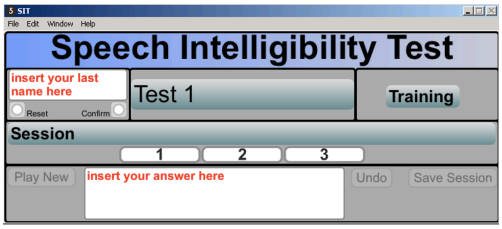

Figure 1: Interface used for the subjective listening tests

S/N	+ 4 dB	0 dB	- 4 dB
Office	HO CANTATO TANTO CHE SONO RAUCO E SENZA FIATO	HA AVUTO L'INTUITO DI RIMUOVERE TUTTI I POSSIBILI OSTACOLI	MI HA ZITTITO CON UN SUONO GUTTURALE, QUASI MAGNETICO
	SONO STANCO DI IMMETTERE DATI NEL COMPUTER	CHE TI SALTA IN MENTE DI ORDINARE SOLO PER TE?	IN FONDO, E' PIU' SIMPATICO IL GUFO CHE IL LEONE
	MI SONO ARRABBIATO CON LUI E HO URLATO A LUNGO	SUONA ANCHE IL LIUTO, MA UN PO' MALE	CHISSA' SE E' MEGLIO L'OLIO DI SOIA O QUELLO DI MAIS
	DALL'ODORE SI DIREBBE COGNAC DENATURATO	LO AGITI UN PO' E HAI GIA' OTTENUTO UN COCKTAIL SCECHERATO	PER LE GOCCIOLE DI CREMA SERVONO MOLTI TUORLI
Lobby	E' IL PERIODO PIU' IELLATO DEI MIEI ULTIMI ANNI	E' UN VERO AMATORE DI PESCA SUBACQUEA	COGLIETE L'OCCASIONE PER IMPIANTARE UNA MAGLIERIA
	GLI HO DETTO LA VERITA' E LUI SE NE E' ANDATO MOGIO MOGIO	NON LO VEDO ARRIVARE: SARA' ULTIMO	IL GALLO SI E' AVVENTATO PER GHERMIRE LA PREDA
	FINISCI COLL'AVERE UN ALGORITMO SDOPPIATO	CI SONO MOMENTI IN CUI SEI ANNOIATO DI TUTTO	TUTTA LA ZONA DELL'OLGIATA E' MOLTO RICCA
	ALT, FERMATEVI O MI SENTO MALE	LA REGIA MI E' SEMBRATA ACCURATA, MA NON BRILLANTE	C'E' UNO SCREZIO SERIO CON TUTTA LA MIA FAMIGLIA

Table I: Speech corpus

We note that, for the same S/N, Office room reverb noise leads to higher values of the intelligibility than the Lobby room noise. Moreover, sentence assessment leads to significantly lower values of intelligibility than those obtained with single word and single phoneme assessment. A measure of sentence intelligibility which is less than that obtained for words or phonemes may look surprising: in reality it is self-evident that results for words and phonemes taken from the same audio material (as the sentences) can only provide better percentages. Indeed, if we correctly transcribe, say, 12 sentences out of 24 (50%), we will have that 50% of words (those in the correct sentences) are correct, but we may have also other words correctly transcribed from partially transcribed sentences. The same reasoning could be applied to the phonemes.

4. Objective Measures

Objective measurements do not measure intelligibility but determine physical parameters to predict intelligibility according to a certain model.

Many objective speech intelligibility measurements have been proposed in the past in order to predict the intelligibility of speech (Herman), (Ma, 2009).

Most of the literature in this field comes from the IT world, where the problem is to study the impact of the transmission channel and the encoders on intelligibility of speech (Kitawaki, 2007), (Liu, 2008).

Three frequently used objective measurement methods were evaluated for use, based on: the signal-to-noise ratio, with the noise filtered by an A-weighting curve (S/NA) (Hu, 2007), the Articulation Index (AI) (Kryter, 1962), (Kryter, 1969) and the Speech Transmission Index (STI) (Payton, 1999).

Unfortunately, all those objective measurements need the clean signal to be available for comparison with the noisy signal. All of them can be referred to as *double-sided methods* and are not suitable for predicting the intelligibility in forensic applications. To this end, we propose a *single-sided* intelligibility measurement based on STI.

In the Speech Transmission Index theory the intelligibility of speech is related to the preservation of the spectral differences between successive speech elements, the phonemes. This can be described by the envelope function.

The envelope function is determined by the specific sequence of phones of a specific utterance.

The STI-based measure is computed as follows. The noisy signal is first bandpass-filtered into seven octave bands from 125 Hz to 8000 Hz.

The envelope of each band is computed using the power of the signal. In particular, if we consider a discrete time-domain signal *x(n)* filtered in the k^{th} octave band we define the envelop function as

$$Env_k(m) = \frac{1}{N_e - 1} \sum_{n=mh}^{mh+N_e-1} H(n-mh)\left[x(n)\right]^2 \qquad (1)$$

where N_e is the window size, h is the hop size, $m \in \{0, 1, 2,\ldots,$ M$\}$ is the hop number, *H(n)* is a finite-length sliding Hanning window and n is the summation variable.

After that, we compute the normalized spectrum envelope as follows

$$s_{k,f_i} = \frac{\left| \sum_{p=0}^{N_s-1} w(p) Env_k(p) \cdot e^{-\frac{i2\pi p f_i}{F_s}} \right|}{\sum_{p=0}^{N_s-1} Env_k(p)} \qquad (2)$$

where N_s is the window size, F_s is the sampling rate, f_i denotes any of the 14 frequencies in the range 0.63 Hz to 12.5 Hz at 1/3-octave step, *w(p)* is a finite-length rectangular window and p is the summation variable. The SNR in each band is computed as

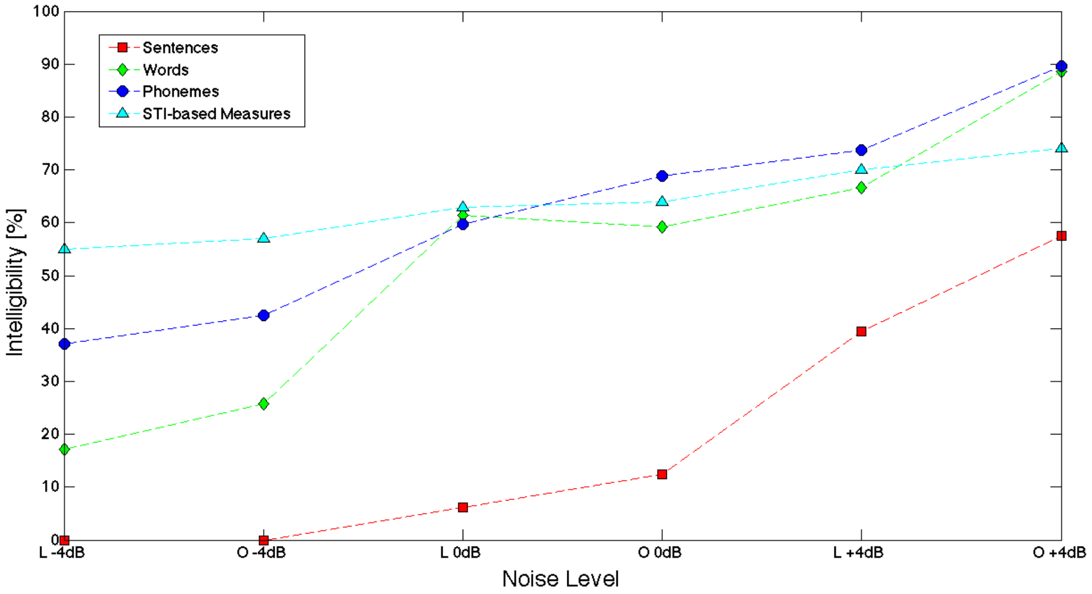

Figure 2: Subjective measures on Sentences, Words and Phonemes and Objective STI-based measures

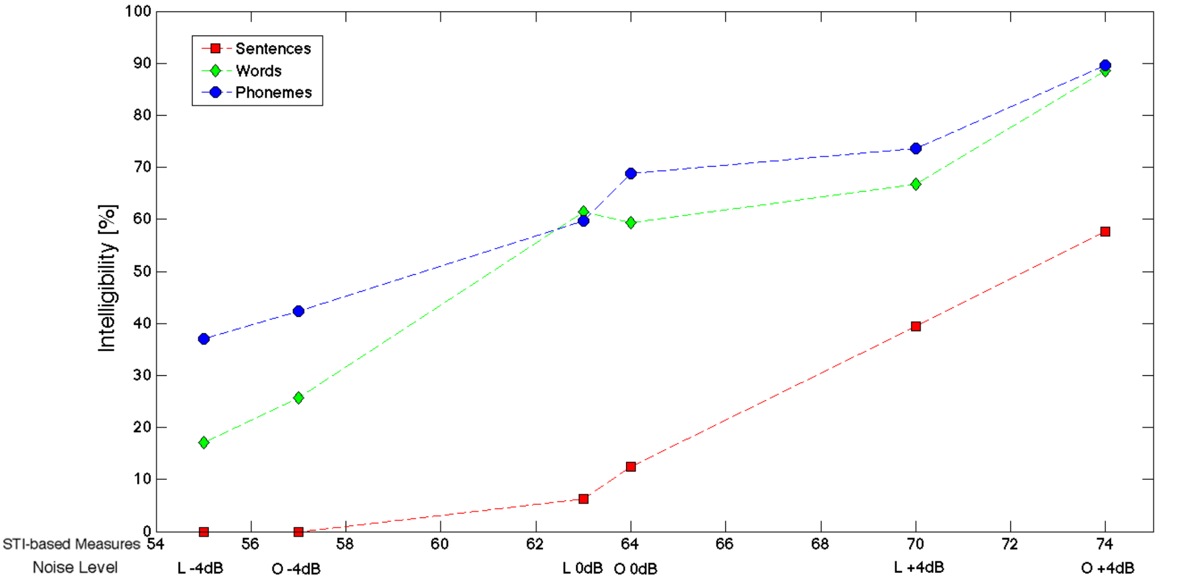

Figure 3: STI-based Measures Vs Intelligibility under different noisy environments

$$SNR_{k,f_i} = 10\log_{10}\left(\frac{s^2_{k,f_i}}{1-s^2_{k,f_i}}\right) \quad (3)$$

and subsequently limited to the range of [-15dB, 15dB]. The Transmission Index (TI) in each band is computed by linearly mapping the SNR values between 0 and 1 using the following equation

$$TI_{k,f_i} = \frac{SNR_{k,f_i} + 15}{30} \quad (4)$$

For each octave band, the average TI over a specified frequency range gives the Modulation Transfer Index (MTI), as

$$MTI_k = \frac{1}{n}\sum_{i=1}^{n} TI_{k,f_i} \qquad (5)$$

Finally, the STI-based measure is obtained as a weighted mean of the MTI over seven octave bands, and is written as

$$STI = \sum_{k=1}^{7} W_k \cdot MTI_k \qquad (6)$$

The sum of these weighting factors W_k is 1, as of [7].

5. Results

Performances of the objective measures are presented in terms of the Pearson product-moment correlation coefficient r between subjective intelligibility ratings and the objective measure, and is given by

$$r = \frac{\sum_{i=1}^{n}\left(S_i - \bar{S}\right)\cdot\left(O_i - \bar{O}\right)}{\sigma_S \cdot \sigma_O} \qquad (7)$$

where S and O are the subjective and objective scores, with mean $\bar{S}$ and $\bar{O}$ and standard deviation σ_S and σ_O, while n is the number the different levels of degraded signal considered. The coefficient ranges from -1 to 1, with 1 indicating the highest-correlation between the two measurements. The experiment was conducted using for the intelligibility assessment the STI-based measure on the speech corpus described in section 2.

The experiment has shown the correlation between subjective and objective data in particular conditions that are typical of forensic applications.

Table II shows the correlation between subjective and objective measures, for all degradations taken into account. The results of these experiments are summarized in Fig. 2-3. Fig. 2 shows intelligibility assessment obtained with the objective method compared with subjective measures. We note that objective values follow words and phonemes intelligibility measures.

Fig. 3 shows STI-based measures versus intelligibility related to sentences, words and phonemes.

	Correlation
Sentences	**0,95**
Words	**0,96**
Phonemes	**0,98**

Table II: Correlation between subjective and objective measures

6. Conclusions

The overall results of this study show that the STI function provides a good estimate of speech intelligibility. In particular, the experiments carried out have proven that our proposed STI measurement procedure is able to predict with sufficient accuracy speech intelligibility in conditions very close to those most frequently found in forensic applications, where both additive and multiplicative noise are involved.

Moreover, we developed a Windows application that operates a short-time STI-based measure; this application allows us to compute the objective intelligibility locally on a noisy signal, using window length of 500*ms*.

Interested readers are invited to download our system from the site indicated below and test it on their signals.

http://voice.fub.it/SSIM/

7. References

Chen D., Fourcin A., et alii (1995). EUROM A spoken language resource for the EU. *ESCA EUROSPEECH*. Madrid.

Costantini, G, Paoloni, A, Todisco, M, (2010). Objective Speech Intelligibility Measures Based on Speech Transmission Index for Forensic Applications. *39th International AES Conference on Audio Forensics: Practices and Challenges*. Hillerød, Denmark, pp. 182-188.

Cycling74 Max/MSP, documentation available on the web at: http://cycling74.com/products/maxmspjitter/

Fraser, H., (2003). Issue in transcription: factors affecting the ryability of transcripts as evidence in legal cases. *Speech Language and the Law*, 10(2), pp. 203--226.

Herman, J.M., Steeneken. The Measurement of Speech Intelligibility, *TNO Human Factors*, Soesterberg, the Netherlands.

Hu Y., Loizou, P.C., (2007). A Comparative Intelligibility Study of Speech Enhancement Algorithms. *Acoustics, Speech and Signal Processing. ICASSP 2007*. Volume 4, Issue 1, Page(s): IV-561 - IV-564.

Kitawaki N., Yamada, T, (2007.) Subjective and Objective Quality Assessment for Noise Reduced Speech. *ETSI Workshop on Speech and Noise in Wideband Communication*. Sophia Antipolis, France.

Kryter K., (1962). Methods for the calculation and use of the Articulation Index. *JASA 34*, 1689–1697.

Kryter, K., ANSI S3.5-1969, (1969). American National Standards Methods for Calculation of the Articulation Index. *American National Standards Institute*, New York.

Liu W. M., Jellyman K. A., Evans N. W. D., and Mason J. S. D., (2008). Assessment of Objective Quality Measures for Speech Intelligibility, *INTERSPEECH 2008, 9th Annual Conference of the International Speech Communication Association Brisbane*, Australia.

Ma J., Hu y., Loizou C. (2009). Objective measures for predicting speech intelligibility in moist conditions based on new band importance functions. *JASA 125*.

Payton K. L., (1999). A method to determine the speech transmission index from speech waveforms. *JASA 106*, 3637-3648.