Speech intelligibility improvement by sound recording techniques

by

Robert Ruiz and Laurent Faiget

Reprinted from

Journal of
Building Acoustics

VOLUME 7 NUMBER 1 2000

MULTI-SCIENCE PUBLISHING CO. LTD.
5 Wates Way, Brentwood, Essex CM15 9TB, United Kingdom

Speech Intelligibility Improvement by Sound Recording Techniques

Robert Ruiz[1] and Laurent Faiget[2]
[1]*LA.R.A, Université de Toulouse-le Mirail, 5 allées Antonio Machado, 31058 Toulouse cedex 1, France*, email: rruiz@univ-tlse2.fr
and
[2]*01dB, 6 av. Louis Blériot, 31570 Sainte Foy d'Aigrefeuille, France*,
email: lfg@toulouse.01db.com

Received 22 June 1999 and accepted 1 November 1999

ABSTRACT
Speech intelligibility studies have modelised the influences of the acoustical features of the room and/or the properties of the voice signal and/or the electroacoustical characteristics of the loudspeakers on scores. A complementary element is added here : the sound recording system. The purpose of this paper is to demonstrate that sound recording microphone techniques can modify intelligibility scores: monophonic techniques (with an omnidirectional and a cardioïd microphones) and a stereophonic O.R.T.F one. The sense and the extent of the variation are discussed: scores are increased when recordings are listened. The experimental conditions are recalled [1] but new results are obtained leading to a new detailed analysis. A discussion is started up concerning the reliability of intelligibility tests in an auralization approach.

INTRODUCTION

To guarantee a good intelligibility in a room is always a delicate trial for the acoustician especially for rooms of great volumes. The object is not here to compare criteria used for intelligibility prediction [2], nor to discuss of absorbing, reflecting materials placement, but to study the influence of sound recording on scores.

The speech is played in a first hall and heard in a second one. In the first one, intelligibility scores are known. The parameter of the study is the sound recording microphone technique. In most of the preceding studies [3,4,5,6], the listener is present in the hall where the speech sound source is. Here, the recordings are reproduced in another

room where listeners are tested. The topic is to know if the intelligibility scores measured in the first hall are modified by the recordings.

Severe acoustical situations must be considered in order to obtain low scores without additional masking noise. The required experimental conditions are found in a large reverberant hall where scores can vary, in a large extent, with the distance from the listener or the microphone(s) to the speech source. On the contrary, listening is performed in a dead acoustics, that is a room with a very short reverberation time. The poor reverberant sound field of the listening room is not superimposed on the sound field reproduced by the loudspeakers, i.e the direct sound. Such experimental conditions are favourable to study microphone technique influence on intelligibility scores. They are similar to those of a recording in a large hall and a listening in a housing room. Applications concern the field of transmission sound quality, auralization techniques and room acoustics prediction softwares.

After the detailed description of the experiment and an explanation of the procedure, results are presented and discussed.

I. EXPERIMENT

A. ROOMS

1. Reverberant Hall

It is an empty church of great volume (15900m^3). Reverberation time is measured in 14 locations (Table 1).

	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz
RT_{60} (s)	7,2	7,6	7,5	6,5	5,4	3,8
σ(s)	1,2	0,9	0,4	0,4	0,2	0,2

Table 1 : Mean reverberation time and standard deviation (reverberant hall).

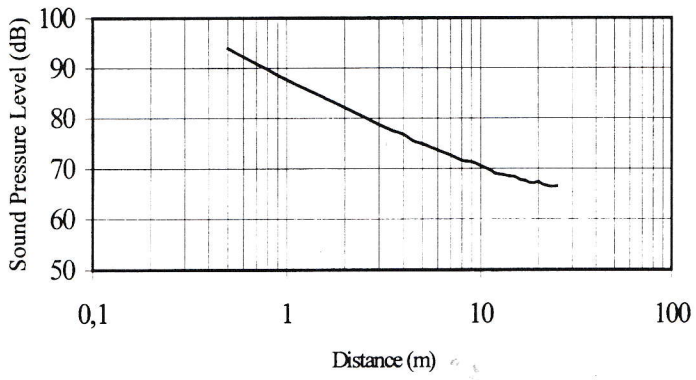

Figure 1 Spatial decrease of sound level pressure in the reverberant hall.

Variations of sound pressure level as a function of distance are done from a white noise source at the altar along the nave (Figure 1). From these measurements, critical distance is estimated about 12m from the edge of the altar where the speech source will be situated.

Acoustical criteria C_{50} is measured at 2m,4m, 8m and 16m from the altar along the nave. The values of C_{50}, expressed in dB, are the followings: 5,8 dB at 2m; 1,1 dB at 4m; -2,9 at 8m and –8,3 at 16m. They clearly indicate a great predominance of the reverberant sound field far from the source. Between 4m and 16m, intelligibility scores will vary in a large extent. Recognition of the vocabulary will be excellent near the speech source and will become difficult when C_{50} will be negative.

2. *Dead Room*

It is a smaller room (8,7×5,6×3m) intended to broadcasting. Reverberation time measurements are done in three positions near the stereophonic listening location where listeners will be.

	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz
RT_{60} (s)	0,8	0,4	0,3	0,3	0,4	0,3
σ(s)	0,1	0,04	0,04	0,02	0,05	0,01

Table 2 : Mean reverberation time and standard deviation (dead room).

In such conditions, the reverberant field in the dead room will not disturb the direct sound reproduced by the loudspeakers, carrier of the useful information. Reflections coming a few milliseconds after the direct sound will take part in enriching and colouring the original signal. Early reflections in the listening room are not strong.

B. SPEECH MATERIAL

Speech must be reproducible to test intelligibility in various locations and at different distances. It has been recorded on a D.A.T. The phonetic material is a set of three phonetically balanced lists (A, B and C) of 34 triphonemic french words each one (306 phonemes). They are named PB word lists. Duration of a list is about 10 minutes. Each word to recognize is preceded by an introductive sentence without any semantic link with it. Its role is to ensure sufficient reverberant conditions for the listening. The listener writes on a specific sheet of paper the word or the phonemes he has recognized. It is specified that spelling does not matter. A trial list is heard before the beginning of the test.

Two intelligibility scores are computed from the answers : the percentage of correctly recognized phonemes and the one of correctly recognized words. The first one is called phoneme-score and the second one word-score.

Lists are played in the reverberant hall over a type 101 Bose loudspeaker (Table 3 and

Figure 2). It has been used for the validation of the intelligibility prediction model proposed by the authors, which take into account electroacoustical features (frequency response and directivity) of the loudspeaker used to reproduce the speech in the room [6].

	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz
Q	1,2	1,3	2,0	3,2	3,8	6,4

Table 3 : Directivity factor Q of the type 101 Bose loudspeaker.

Signal-to-noise ratio, even far from the source, is always greater than 25 dB(A). Low intelligibility scores are not due to small signal-to-noise ratios.

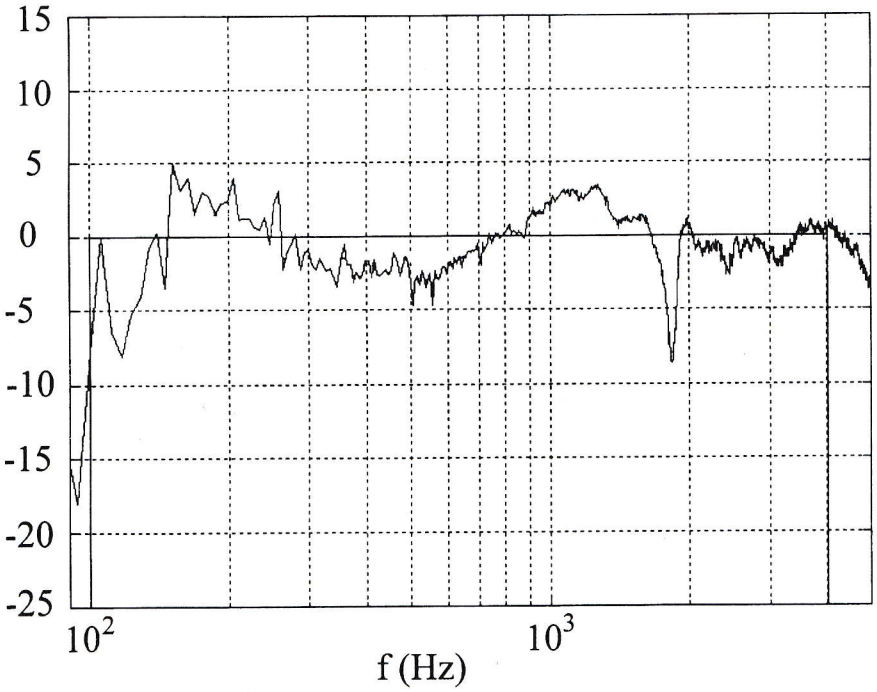

Figure 2 Frequency response of the type 101 Bose loudspeaker used to play the word lists in the reverberant hall and in the listening room.

C. EXPERIMENTAL PROCEDURE

Figure 3 shows the three phases of the experimental procedure.

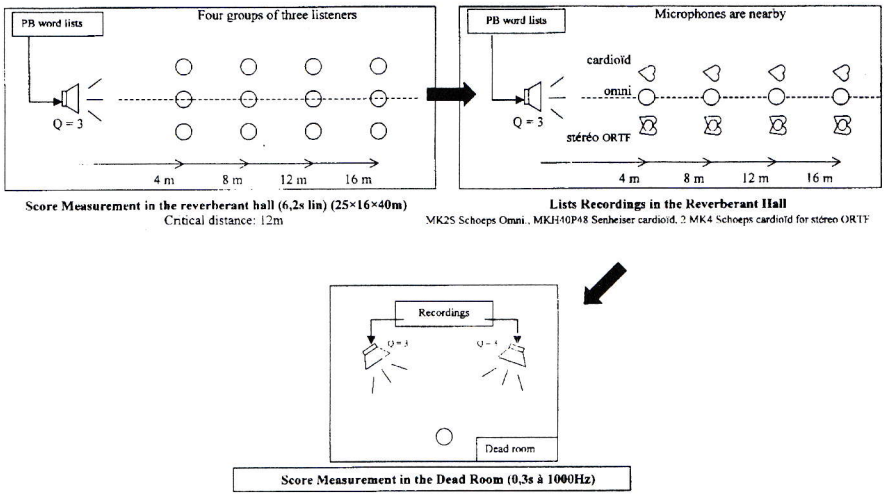

Figure 3. Experimental procedure

1/ Speech understanding is altered when distance from the speech source rises: the direct sound energy decreases leading to a reverberant sound field more and more present and, the time interval between the direct wave and the first reflections reduces. These effects are the main causes of the deterioration of intelligibility. To obtain a large extent of scores, four distances are chosen in the reverberant hall: 4, 8, 12 and 16m from the loudspeaker. The two first are in a predominant direct sound area and the last one belongs to a diffuse sound field area. At 12m, a balance exists. Twelve subjects, divided in four groups of three, are seated in front of the loudspeaker at these locations. They are chosen from the general public of students with normal hearing. This first phase of the experiment leads to obtain, for each distance, the mean intelligibility scores in the reverberant hall.

2/ At the same positions, later, with the same loudspeaker but without any listener, the same word lists are recorded on a D.A.T. with three different microphone techniques: two monophonic ones (with an omnidirectional microphone and with a cardioïd one) and a stereophonic one. Microphones were at the same height than the listeners.

Finally, acoustical measurements, intelligibility tests and recordings have all been done at the same locations.

3/ New intelligibility tests are performed in the dead room with a new set of listeners. None of them (12 again) was present during the first tests in the reverberant hall.

None of them knew this church. Listenings are performed individually. Three subjects listen to the recordings performed at 4m: the A word-list for the omnidirectional recording, the B word-list for the cardioïd one and the C word-list for the stereophonic one. Then three other subjects listen to the same lists but recorded at 8m (id for 12m and 16m).

Listeners are seated in front of two loudspeakers (type 101 Bose) in stereophonic position: equilateral triangle with sides of 3,2 m. The perceived positions of the monophonic and the stereophonic speech are at the centre between the loudspeakers. A third-octave band equalization is done at the listener position to avoid both spectral influence of the room and of the loudspeakers (80-8000 Hz).

Intelligibility scores are then averaged by distance value and by microphone technique.

D. MICROPHONE TECHNIQUES

Monophonic recordings are performed with an omnidirectional MK2S Schoeps microphone and a cardioïd MKHP48 Senheiser one. Stereophonic recordings are done with two cardioïd MK4 Schoeps microphones in an ORTF system. They are angled 110° apart and spaced 17 cm horizontally [11]. It is a near-coincident technique which uses phase and level differences between the transducers to give a good localization accuracy with spaciousness and depth. The O.R.T.F technique is a good compromise between coincident and spaced microphone techniques.

II. RESULTS

Mean intelligibility phoneme and word scores are computed for each distance and each microphone technique (Tables 4 and 5; Figures 4 and 5).

Distance from the source	Score in the reverberant hall	Omni mono recording score	Cardioïd mono recording score	Stereo recording score
4m	97,3 (1,2)	95,0 (2,2)	98,7 (0,5)	97,3 (1,9)
8m	93,7 (2,0)	95,0 (0,8)	94,0 (1,4)	96,3 (2,0)
12m	not performed	84,7 (1,9)	97,0 (1,4)	94,7 (2,9)
16m	77,7 (4,2)	78,0 (5,0)	80,3 (4,1)	90,3 (4,9)

Table 4: Mean intelligibility scores based on a phoneme recognition (standard deviations in brackets).

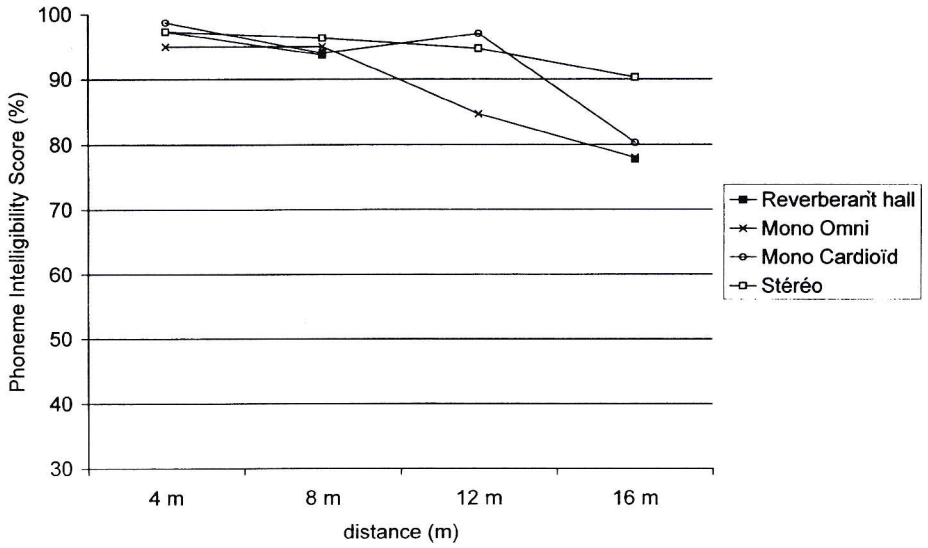

Figure 4 Phoneme intelligibility score as a function of distance from the speech source in the reverberant hall and for the listenings of the monophonic and stereophonic recordings.

Distance from the source	Score in the reverberant hall	Omni mono recording score	Cardioïd mono recording score	Stereo recording score
4m	93,0 (1,4)	89,0 (5,7)	97,0 (0,0)	92,0 (6,2)
8m	79,3 (6,9)	91,0 (2,4)	88,0 (2,4)	93,0 (2,8)
12m	not performed	74,7 (0,9)	92,0 (3,7)	91,0 (4,9)
16m	46,0 (8,6)	56,0 (8,8)	67,7 (8,5)	85,3 (8,4)

Table 5: Mean intelligibility scores based on a word recognition (standard deviations in brackets).

Significance of the difference between the means is estimated by a t-test [7] (Tables 6,7 and 8). The level of significance is expressed in percentages. For example, the classical unilateral 0,05 level becomes the 10% bilateral percentage in the tables. In the experiment, the sample size is three and the mean intelligibility scores are often associated with large standard deviations. A great significance, that is a low percentage level, can be difficult to obtain.

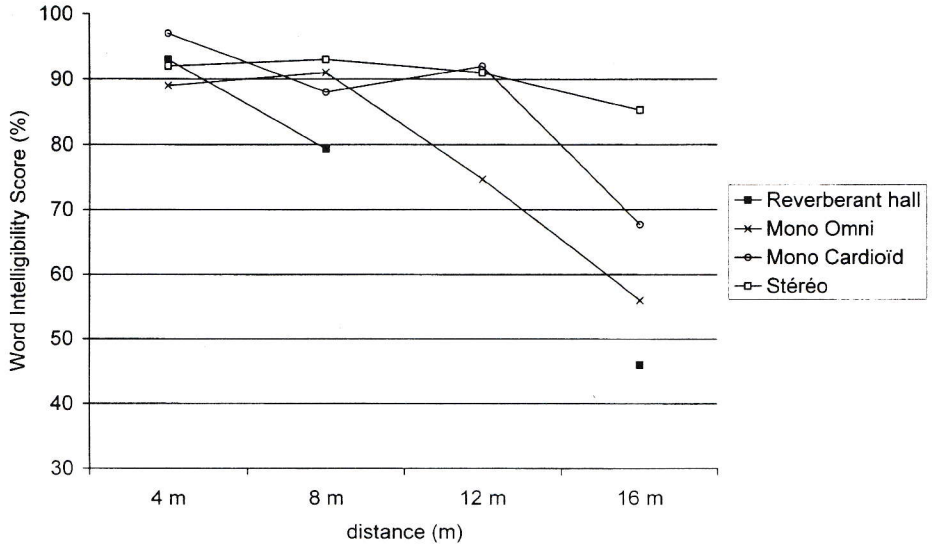

Figure 5 Word intelligibility score as a function of distance from the speech source in the reverberant hall and for the listenings of the monophobic and stereophonic recordings.

	Omni score	Cardioïd score	Stereo score
Score in the reverberant hall: 4m	Phonemes: $S_{50\%}$ Words: $S_{50\%}$	Phonemes: $S_{50\%}$ Words: $S_{2\%}$	Phonemes: NS Words: NS
Score in the reverberant hall: 8m	Phonemes: $S_{50\%}$ Words: $S_{10\%}$	Phonemes: NS Words: $S_{20\%}$	Phonemes: $S_{50\%}$ Words: $S_{10\%}$
Score in the reverberant hall: 16m	Phonemes: NS Words: $S_{50\%}$	Phonemes: NS Words: $S_{10\%}$	Phonemes: $S_{10\%}$ Words: $S_{1\%}$

Table 6 : Significance of the difference between mean scores in the reverberant hall and mean scores in the dead room at the same distances. $S_{x\%}$ means that the score difference is significant at the x% level; NS means that the score difference is not significant.

	Cardioïd score	Stereo score
Omni score : 8m	Phonemes: $S_{50\%}$ Words: $S_{50\%}$	Phonemes: $S_{50\%}$ Words: $S_{50\%}$
Omni score : 12m	Phonemes: $S_{0.2\%}$ Words: $S_{0,5\%}$	Phonemes: $S_{2\%}$ Words: $S_{1\%}$
Omni score : 16m	Phonemes: NS Words: $S_{50\%}$	Phonemes: $S_{10\%}$ Words: $S_{5\%}$

Table 7 : Significance of the difference between mean scores in the dead room for the omnidirectional recordings and mean scores in the dead room for the cardioïd and the stereo recordings at the same distances.

	Stereo score
Cardioïd score : 8m	Phonemes: $S_{50\%}$ Words: $S_{20\%}$
Cardioïd score : 12m	Phonemes: $S_{50\%}$ Words: NS
Cardioïd score : 16m	Phonemes: $S_{50\%}$ Words: $S_{20\%}$

Table 8 : Significance of the difference between mean scores in the dead room for the cardioïd recordings and mean scores in the dead room for the stereo recordings at the same distances.

III. ANALYSIS OF THE RESULTS

A speech intelligibility score improvement is observed after listening of the recordings, excepted for those performed with an omnidirectional microphone. Whatever is the recording technique and the distance from the speech source, intelligibility scores based on a phoneme or a word recognition are at least equal or greater than the reverberant hall ones (Tables 4,5 and Figures 4,5).

Results are analysed separating the distance effects and the microphone technique influences. The word scores are compared with the phoneme ones.

A. DISTANCE FROM THE SPEECH SOURCE

At the distance 4m, all the scores are similar and high. It shows that microphone techniques do not modify understanding when it is almost perfect in the reverberant hall. The recording and the reproduction electroacoustical devices do not change intelligibility.

From 8m to 16m, the differences between scores grow up. Scores dispersion also increases with the distance and similarly in the reverberant hall and for the recording listenings. The microphone techniques do not contribute to decrease the high natural variability of scores measured in situ.

B. MICROPHONE TECHNIQUE

1. Monophony:

The scores after the listenings of the omnidirectional recordings are very similar to those measured in the reverberant hall whatever is the distance but not whatever is the counting mode of the vocabulary (phonemes or words).

The analysis of the sound field for the modelization of intelligibility are based, for most of the models, on the computing of sound energy ratios issued from echograms [2,3,6], themselves issued from the impulse responses I.R. These ones are either measured for a future acoustics treatment of the room or simulated for a prediction. The I.R

measurement and the computing of the usual acoustics criteria are very often performed using Maximum Length Sequences (M.L.S) and they need some specific metrological precautions [8]. The other main intelligibility prediction technique, based on a Modulation Transfer Function, also uses the impulse response of the room [5,9]. The I.R estimation is always the first operation of the intelligibility prediction.

When I.R is measured, the technique uses an omnidirectional microphone and, when it is simulated by specific softwares, the method implicitly assumes an omnidirectional directivity at the reception point in the virtual room.

So, when the impulse responses are convolved with the speech to perform intelligibility tests based on a phoneme recognition, the obtained scores will probably be a very good approximation of the real ones. Auralization can be a good mean to test intelligibility without going into the room or before its construction. Even if the speech is played through headphones, essentially to cut off the listener from the acoustics of the room where he is, the results will be the same. Indeed, headphone listening of monophonic recordings moves the listening point from the centre of the loudspeakers to the centre of the head slightly back.

One can notice that for great distances from the speech source, when acoustical conditions are not favourable to intelligibility, omnidirectional recording leads to an increase of the word-scores (56% instead of 46% in situ, Table 5). Even if the comparison is not very significant (Table 6), it should be surely better with a large sample size. The improvement exists at 8m. Therefore, it seems that the intelligibility estimation from an auralization technique needs a phoneme-score computing to be in agreement with real scores in the room. If the method is performed with a word-score computing, the prediction can overestimate the reality.

The *cardioïd technique* induces a more important increase of the intelligibility scores. It is less adapted to recognition tests by impulse responses convolutions. The increase can be explained by the fact that the ratio of the direct sound energy to the reverberant one is greater for the more directive microphone and by the fact that direct sound holds the information to recognize and the reverberant sound field the detrimental one.

Here, it is much more a tendancy than a very significant result. Dispersion of the scores for great distances contributes to the difficulty in obtaining better results for the t-tests with a small sample size.

Furthermore, microphone is generally more directive for high frequencies which are very linked to speech intelligibility (consonants) [4]. Frequency stability of the omnidirectional microphone directivity is an important feature to guarantee reliable intelligibility scores in auralization techniques.

2. Stereophony:

Scores are very high even for the recordings far from the speech source. The gain supplied by the recording is 13% for the phoneme-score and 39% for the word-score. The acoustical conditions in the reverberant hall leads to the recognition of one phoneme out of two and the recording to the recognition of more than eight phonemes out of ten. This improvement is not due to an excessive decrease of reverberant sound or to an

excessive increase of direct sound. Stereophonic techniques are more efficient to reproduce the original acoustics than monophonic ones. Sound ambiance of the recording hall is present in the listening room.

Dispersion of the scores is the same than in situ. Two reasons can act jointly:

- dispersion is mainly due to inter-individual differences, independent of the acoustical situations (these last ones are not extremely favourable or unfavourable);
- the stereophonic reproduction only acts on mean scores and does not modify the natural variability.

With a few exceptions, the result applies to the dispersions of the mean scores of the two others microphone techniques.

All the listenings shares the playback of the recorded sound field in a half space in front of the listener. In monophony, it is "compact" and, in stereophony it takes up all the space of the azimutal plane between the two loudspeakers. This fact probably participates in the small increase of scores for the "compact" reproduction and the greater one for the "large" listening. In the first case, sound waves coming from the "back" of the microphone are mixed with direct sound and played in front of the listener. Moreover, for the omnidirectional recording their energy is not weakened by the microphone directivity. In the second case, these back-waves are also mixed with the direct sound but spread out in front of the listener. A spatial separation of front-back waves is realised by the stereophonic technique in the frontal azimutal plane of the listening. This separation is surely improved by the use of cardioïd microphones instead of two monophonic ones. The substantial improvement in the intelligibility scores of the stereophonic recordings can also be caused by the emphasis of direct sound in the stereophonic arrangement. Perhaps the ratio of direct sound energy to reflected sounds one is increased by the use of two cardioïd microphones leading to a better phoneme or word recognition.

Concerning the applications to auralization techniques, it seems that impulse response measurements with an artificial head be well adapted to intelligibility tests with an headphone listening (or via loudspeakers with an adapted signal processing). Similarity between in situ scores and those of recordings could be better. But an experiment is necessary to validate the hypothesis. When public address systems are tested, monophonic recordings are not suitable because the useful speech information comes from more than one direction. A microphone technique well correlated with the reproduced sound sources localization is needed. Again, a validation is necessary, comparing for example, an omnidirectional stereo and an artificial head technique with in situ intelligibility scores.

C. PHONEME-SCORE AND WORD-SCORE

General tendancies of the score changes with distance and microphone technique are similar whatever is the counting mode: phonemes or words. The curves of Figures 4 and 5 have the same sense of variation with the distance and the recording technique. The result is logical because words and phonemes are issued from the same written texts.

The gain difference introduced by the microphone technique between phoneme and word scores is important. At 16m for example, the word score is 85,3% for the stereo recording and 46% in situ; the phoneme score is 90,3% for the stereo and 77,7% in situ. Considering the word counting, the difference corresponds to 13 words (or 39 phonemes because words are triphonemic) and to 12 phonemes for the phoneme counting (or 4 words). It does not exist a linear relationship between the two counting modes. Figure 6 shows the 3-order polynomial regression (determination coefficient $R^2 = 0,99$) [10].

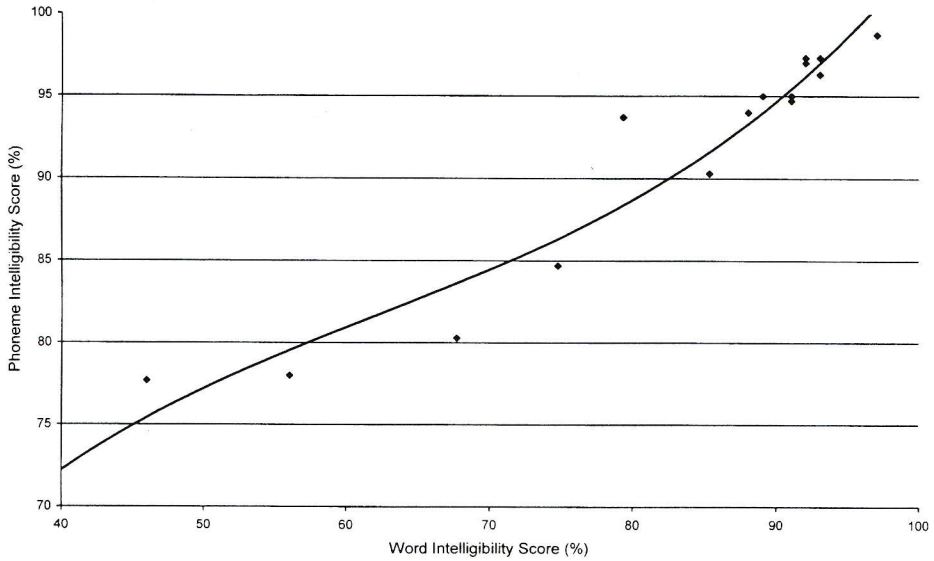

Figure 6 Third-order polynomial function between phoneme and word scores.

The counting of the recognized phonemes always leads to greater scores than the word counting. Sometimes, particularly when the acoustical conditions are unfavourable, the listener only writes a few phonemes of the word to recognize, which increases the phoneme score but not the word one. Even if the numerical differences seems to be important, they are normal. The question is to know what is the more appropriate or adapted counting mode to estimate the speech intelligibility. The correlation is not known. But is is likely that the two types of score underestimate the "perceived" score, because of the redundancy and the foreseeable nature of the speech which are absent of the word lists. In practice, empirical comparisons between predicted and estimated scores allows the acoustical designer to choose the numerical limits of the scores to guarantee a good intelligibility.

Classical intelligibility models [3,5] are well correlated with word-scores. So, if auralization is performed with a phoneme-score computing in order to avoid the increase of the word-score (cf III.B.1), a conversion of phoneme-scores to word-scores is needed to use the models.

CONCLUSION

The study reveals that it can be preferable to listen to speech in another room than the one where it is pronounced, in order to better recognize the phonemes or the words of the speech. The result is surprising because this operation needs additional electroacoustical devices and another room, which are able to decrease scores. But under the classical and ordinary conditions of the experiment, the technique improves the recognition.

The result is applied to the intelligibility prediction from the simulated or estimated impulse response of the room. When these techniques are used, the favourable effect of the increase can act in disfavour of the quality of the prediction, because scores can be overestimated. Finally, if the auralization is used only in the goal to predict intelligibility scores, a monophonic recording seems to be enough. An artificial head is recommended if a public address system is tested and if the future listener is close to the loudspeakers.

The experiment indicates tendancies and raises questions, from which others experiments can answer. It shows the role and the influence of the recording microphone technique to improve the intelligibility score. The change of scores with the microphone technique reminds that in good acoustical conditions, the influence of the factors acting on intelligibility is small but, when the acoustics becomes unfavourable, their modification can greatly improve the scores [6]. The sound recording technique confirms this property.

Further experiments are possible by using the M-S microphone stereophonic technique. By changing the S gain, the stereo width and the equivalent polar pattern can be varied from a monophonic recording to a very large stereophonic scene. Recordings of the same lists, with variable stereo perspectives at different distances from the source, should be useful to determine the best compromise between the intelligibility score improvement and the natural, the fidelity of the recording.

ACKNOWLEDGEMENTS

The authors thanks Isabelle Ballet, Engineer of the Ecole Supérieure d'AudioVisuel, who has led the intelligibility tests and all the listeners for their voluntary taking part.

REFERENCES

[1] R.Ruiz, I.Ballet, "Influence of Distance and Sound Recording System on Intelligibility in Highly Reverberant Conditions," *16th International Congress on Acoustics and 135th Meeting of the A.S.A*, Seattle, U.S.A (1998).

[2] K.D.Jacob, "Correlation of Speech Intelligibility Tests in Reverberant Rooms with Three Predictive Algorithms," J. Audio Eng. Soc., vol 37 (12), pp. 1020-1029 (1989).

[3] S.Bradley, "Predictors of Speech Intelligibility in Rooms," J. Acoust. Soc. Am., vol 80 (3), pp. 837-845 (1986).

[4] V.M.A.Peutz, "Articulation Loss of Consonants as a Criterion for Speech Transmission in a Room," J. Audio Eng. Soc., vol 19, pp. 915-919 (1971).

[5] H.J.M.Steeneken, T.Houtgast, "A Physical Method for Measuring Speech-Transmission Quality," J. Acoust. Soc. Am., vol 67 (1), pp. 318-326 (1980).

[6] L.Faiget, R.Ruiz, "Speech Intelligibility Model Including Room and Loudspeaker Influences", J. Acoust. Soc. Am., vol 105 (5), pp. 3345-3354 (1999).

[7] T.H.Wonnacott, R J.Wonnacott, *Statistique*, (John Wiley and Sons Inc, 1995).

[8] L.Faiget, C.Legros, R.Ruiz, "Optimization of the Impulse Response Length: Application to Noisy and Highly Reverberant Rooms", J. Audio Eng. Soc. 46(9), 741-750 (1998).

[9] M.R.Schroîder, "Modulation Transfer Functions: Definition and Measurement", Acustica 49, 179-182 (1981).

[10] H.Fletcher, R.H.Galt, "The perception of speech and its relation to telephony", J. Acoust. Soc. Am. 22(2), 89-151 (1950).

[11] B.Bartlett, "Stereo Microphone Techniques", Focal Press (1991).