Audio Engineering Society
Convention Paper
Presented at the 123rd Convention
2007 October 5–8 New York, NY, USA

Quality impact of diotic versus monaural hearing on processed speech

Arnault Nagle[1], Catherine Quinquis[1], Aurélien Sollaud[1], Anne Battistello[1] and Dirk Slock[2]

[1] France Telecom Research and Development, 2 avenue Pierre Marzin, 22300 Lannion, France
{first_name.name}@orange-ftgroup.com

[2] Institut Eurecom, BP 193 F-06904 Sophia Antipolis cedex, France
slock@eurecom.fr

ABSTRACT

In VoIP audio conferencing, hearing is done over handsets or headphones, so through one or two ears. In order to keep the same loudness perception between the two modes, a listener can only tune the listening level. The goal of this paper is to show that monaural or diotic hearing has a quality impact on speech processed by VoIP coders. It can increase or decrease the differences in perceived quality between tested coders and even change their ranking according to the sound level. This impact on the ranking of the coders will be explained thanks to the normal equal-loudness-level contours over headphones and the specifics of some coders. It is important to be aware of the impact of the hearing system and its associated sound level.

1. INTRODUCTION

In the context of Voice over IP (VoIP) conferencing, whatever the configuration is (centralized, distributed or semi-centralized with the use of a forwarding bridge), monophonic streams are the most frequently used.

Audio frames are encoded and packetized on the sender side, and then the processing on the network depends on the conferencing configuration. On the receiver side, after decoding and possibly mixing, for convenience, a participant frequently uses headphones instead of a handset to play the monophonic sound coming from the VoIP client. To hear the same content on the two ears is called *diotic hearing*.

However in ITU-T, on which we can rely to propose a coder for a conferencing application, audio quality tests are sometimes done in monaural hearing using a handset, so on one ear.

In this paper, our main goal is to show that the quality ranking of coders can be modified according to the hearing system and its associated listening level.

This paper presents the detailed procedure and results of a subjective test that was conducted in France Telecom.

In section 2, we present the common experiment between narrowband and wideband tests. In section 3, we describe the narrowband processing and its associated results. The same description for wideband processing is presented in section 4. We finally discuss our tests before concluding.

2. COMMON EXPERIMENT PROTOCOL

2.1. Absolute Category Rating method

Narrowband and wideband tests were run following the Absolute Category Rating method to simulate a transmission between a sender and a receiver with one encoding operation and one decoding operation. The experiment used two sentence-pairs for two male and two female talkers. The speech material used was extracted from the France Telecom speech database. This database contains quiet background speech of 8 seconds sampled at 16 kHz, for a bandwidth of 8 kHz.

Both tests were performed using 32 different listeners, divided into four groups of eight listeners each. The processed speech material was presented to each group, seated in an acoustically conditioned sound room following the P.800 requirements [1].

2.2. Two sessions, one for each hearing system

In each test, two sessions with training were done. The training consists of a small test example before the actual test, to familiarize listeners with the Mean Opinion Score (MOS) scale. In one of the sessions, all test stimuli have been presented monaurally and in randomized order, one order for each of the four listener groups. The hearing was done over a Sennheiser HD 25 headset with flat response in the audio-bandwidth of interest: 50Hz-7kHz. The other ear was left open.

In the other session, the **same** stimuli have been presented diotically to the subjects, in the same randomized order as the first session and over the same headphone. In each of the four listener grous, half of the listeners did the monaural session first and next the diotic session, and the other half did the opposite.

Obviously, the best choice was to compare the two hearing systems simultaneously in a randomized interleaved fashion. Indeed, before each sample listening, we could have informed subjects to leave free or not one of their ears. However with this procedure, listeners could not remain concentrated. The second solution was, to constantly leave the headset in place, to play a comfort noise on the assumed free ear. However, it was difficult to choose the comfort noise properly.

2.3. The choice of the listening level of each session

In order to compare the two sessions fairly, listening levels were normalized to have the same auditory perception. We had to choose between two methods to implement this:

- The first method was to use the weighted decibel scale (dBA dBB or dBC), which take ear sensitivity into account. This method attenuates more or less by some dB SPL (Sound Pressure Level) according to the processed octave band.

- The second method was to attenuate by some dB SPL over the whole frequency spectrum. It is this option that was chosen due to our VoIP application. Indeed, in audio conferencing, a listener can only tune the sound level to adjust his auditory perception when he changes his system of hearing.

The ITU-T performed monaural tests at 79 dB SPL. So, due to our chosen method, the next step is to evaluate, in dB SPL, the attenuation to be applied to the diotic content. However, the current state of the art discusses such attenuation for either pure tones or broadband noise.

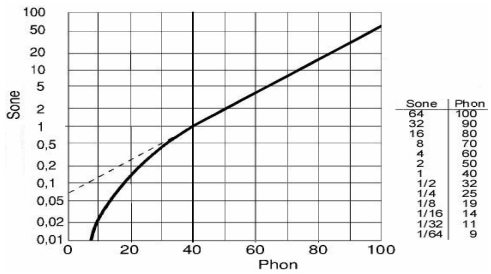

Figure 1 : Sone-Phon matches for diotic hearing in free field

Concerning pure tones, it was shown in [2] that a sound heard with one ear has a loudness (expressed in sones) twice smaller than the same sound heard with both ears. The scale of sones is the relative loudness (subjective impression of sound level) of a pure tone at 1kHz with respect to its loudness at 40 dB SPL. Thanks to this

information and Figure 1, a link can be made between loudness and phons. For a pure tone at 1kHz, the scale of phons is the sound level expressed in dB SPL. For a pure tone at a frequency f, it is the sound level expressed in dB SPL of a pure tone at 1kHz that sounds equally loudly (see Figure 2). It can be estimated that we have to attenuate the diotic content by 10 phons. So, according to Figure 2, which establishes the relation between phons and dB SPL, hence we decided to attenuate the diotic content by 10 dB SPL.

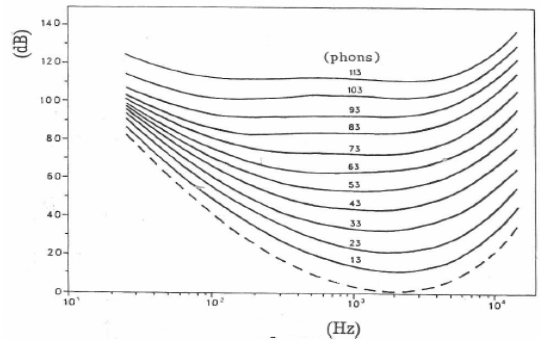

Figure 2 : Normal equal-loudness-level contours over headphones

This 10 dB attenuation for pure tones is found also for the case of broadband noise at 80 dB SPL in monaural hearing over headphones, see Figure 3 [3]. The lowpass noise considered in [3] has constant spectrum and is bandlimited to [100-500kHz].

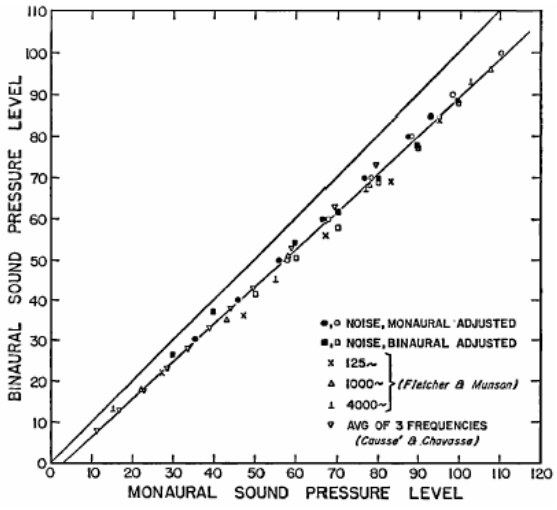

Figure 3 : Monaural-Binaural loudness matches from several experiments. The line with a slope of 45° would be the locus of matches if there was no binaural summation of loudness [3]

Since we found the same attenuation level for both pure tones and lowpass noise, we expect it to be also applicable to speech signals.

The listening level was chosen at 79 dB SPL for the monaural session and a decrease of 10 dB SPL was applied per channel over the whole frequency spectrum for the diotic session (69 dB SPL), according to the normal equal-loudness-level contours over headphones ([2], [4]) and our application constraint (frequency independent attenuation). In a pre-test performed in the same conditions as the test, we validated those values.

2.4. The choice of coders

We considered the coders that are the most likely to be used in the context of VoIP and for interoperability with mobile networks. For each test, the direct condition (without coder) follows the same processing as the other conditions (with coders), but obviously without the encoding and decoding steps.

Whatever the hearing session was, packet loss was introduced with frame erasure rate of 3% or 6%. For each coder, we used its associated "Packet Loss Concealment" or "Frame Erasure Correction" algorithm to correct the packet loss.

For each test, there are 16 conditions = 5 coders x 3 packet loss levels (0%, 3%, 6%) + the direct condition.

3. NARROWBAND PROCESSING AND RESULTS

3.1. Narrowband processing

The tested coders are ITU-T G.729.1 [5] at 8 and 12 kbits/s, 3GPP AMR [6] at 4.75 and 12.2 kbits/s, and ITU-T G.711 [7] at 64 kbits/s with Appendix I [8].

The processing begins with 16 kHz French speech source files which are next handled by a MIRS 16 filtering [7]. Then we adjust them to -26dB$_{ov}$ using the P.56 speech voltmeter [7] and we down-sample them from 16kHz to 8kHz using a high-quality filter [7]. Next, we apply the different coders with or without packet loss. Finally, we up-sample the files from 8kHz to 16kHz using a high-quality filter [7] and we process them by a RXIRS16 filtering [7]. The files are available

to be played back in a monaural or diotic way according to the session.

3.2. Narrowband results

Factor	Degrees of freedom	F-ratio	Significance
Hearing	1	157.94	*0.00*
Coder	5	509.24	*0.00*
Packet Loss	2	1564.14	*0.00*
Speaker	3	40.39	*0.00*
Sample	1	12.99	*0.00*
Order of hearing	1	0.06	0.81

Table 1 Effects of the different factors (N=8192)

First, we perform an ANalysis Of VAriance (ANOVA) in order to evaluate the effects of the different factors (see Table 1) with N the number of observations (32 listeners x 2 sessions x 16 conditions x 4 speakers x 2 samples per speaker). The meaning of *significance* is the probability that the corresponding factors did not influence the test results. Except the Order of hearing (Monaural => Diotic or Diotic => Monaural), the factors Hearing, Coder, Packet Loss, Speaker, and Sample have a significant impact on the test.

In the following, results are shown for each hearing session and for each percentage of packet loss.

3.2.1. Without packet loss

In Figure 4, the Mean Opinion Score (MOS) [1] without packet loss is presented. For each coder, the left and right histograms are respectively the results for the monaural and diotic hearings.

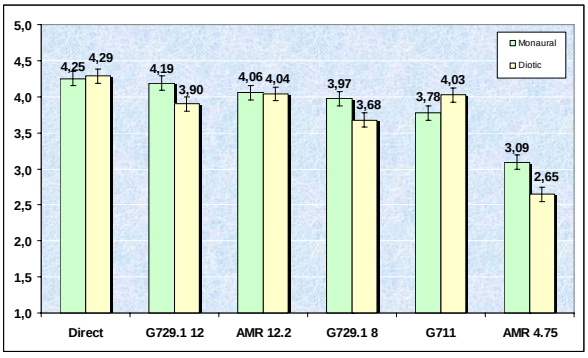

Figure 4 : MOS for each coder in monaural hearing (left histogram) and diotic hearing (right histogram) without packet loss

In order to establish classes of equivalence, we perform a t-test (see Table 2) to compare the coders between each other. The confidence interval (CI) at 95 % is defined as:

$$CI = 1.96 \times \sqrt{\frac{MSE}{N}} \qquad (1)$$

where N=256 and MSE are respectively the number of samples and the mean square error computed in the ANOVA procedure and whose value is 0.65580. So the confidence interval value is 0.0991. Please note that it is the same whatever the coder, the percentage of packet loss and the hearing session. We defined standard error as $\sqrt{MSE\ /N}$ where $MSE\ /N$ is the best estimator of the real variance.

Ranking	Monaural hearing	Diotic hearing
1	Direct <=> G.729.1 12	Direct
2	G.729.1 12 <=> AMR 12.2	AMR 12.2 <=> G.711
3	AMR 12.2 <=> G.729.1 8	G.711 <=> G.729.1 12
4	G.711	G.729.1 8
5	AMR 4.75	AMR 4.75

Table 2 : Ranking of the coders without packet loss

To compare a coder between the two hearing sessions, we perform a second t-test (CI=0.0991). Please note that the two sessions have not been performed at the same time:

- Direct Monaural is equivalent to Direct Diotic. Idem for AMR 12.2.
- AMR 4.75 Monaural is not equivalent to AMR 4.75 Diotic. Idem for G.711, G.729.1 8 and 12.

3.2.2. With 3% packet loss

In Figure 5, the MOS for 3 % packet loss is presented.

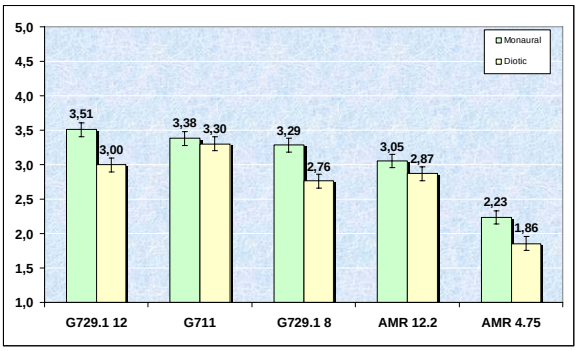

Figure 5 : MOS for each coder in monaural hearing (left h.) and diotic hearing (right h.) for 3 % packet loss

In order to establish classes of equivalence, we perform a second t-test to compare the coders between each other (see Table 3).

Ranking	Monaural hearing	Diotic hearing
1	G.729.1 12 <=> G.711	G.711
2	G.711 <=> G.729.1 8	G.729.1 12 <=> AMR 12.2
3	AMR 12.2	AMR 12.2 <=>G.729.1 8
4	AMR 4.75	AMR 4.75

Table 3 : Ranking of the coders for 3 % packet loss

To compare a coder between the two hearing sessions, we perform a t-test. Please note that the two sessions have not been performed at the same time:

- G.711 Monaural is equivalent to G.711 Diotic.
- AMR 4.75 Monaural is not equivalent to AMR 4.75 Diotic. Idem for AMR 12.2, G.729.1 12, G.729.1 8.

3.2.3. With 6% packet loss

In Figure 6, the MOS for 6 % packet loss is presented.

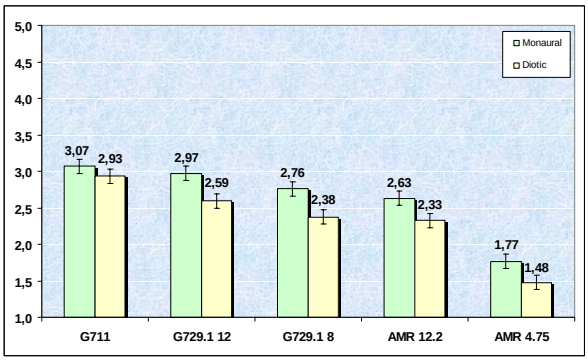

Figure 6: MOS for each coder in monaural hearing (left h.) and diotic hearing (right h.) for 6 % packet loss

In order to establish classes of equivalence, we perform a t-test to compare the coders between each other (see Table 4).

Ranking	Monaural hearing	Diotic hearing
1	G.711 <=> G.729.1 12	G.711
2	G.729.1 8 <=> AMR 12.2	G.729.1 12
3	AMR 4.75	G.729.1 8 <=> AMR 12.2
4		AMR 4.75

Table 4 : Ranking of the coders for 6 % packet loss

To compare a coder between the two hearing sessions, we also perform a t-test. Please note that the two sessions have not been performed at the same time. The results show that the monaural hearing is not equivalent to the diotic hearing for each coder.

4. WIDEBAND PROCESSING AND RESULTS

4.1. Wideband processing

The tested coders are ITU-T G.729.1 [5] at 16 and 32 kbits/s, 3GPP AMR-WB [9] at 12.65 and 23.85 kbits/s, and ITU-T G.722 [7] at 64 kbits/s with appendix IV [10].

The processing begins with French speech source files at 16kHz which are next handled by P341 filtering [7]. Then we adjust them to -26dB$_{ov}$ using the P.56 speech voltmeter [7]. Next, we apply the different coders with or without packet loss. The files are available to be played back in a monaural or diotic way according to the session.

4.2. Wideband results

Factor	Degrees of freedom	F-ratio	Significance
Hearing	1	94.82	*0.00*
Coder	5	69.75	*0.00*
Packet Loss	2	3244.08	*0.00*
Speaker	3	206.36	*0.00*
Sample	1	41.08	*0.00*
Order of hearing	1	45.57	*0.00*

Table 5 : Effects of the different factors (N=8192)

First, we perform an ANOVA in order to evaluate the effects of the different factors (Table 5) with N the number of observations. All factors Hearing, Coder, Packet Loss, Speaker, Sample and Order of hearing (Monaural -> Diotic or Diotic-> Monaural) have a significant impact on the test.

In the following, results are shown for each hearing session and for each percentage of packet loss.

4.2.1. Without packet loss

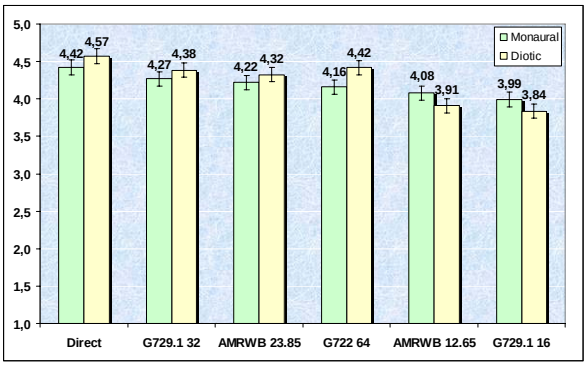

Figure 7: MOS for each coder in monaural hearing (left h.) and diotic hearing (right h.) without packet loss

In Figure 7, the MOS without packet loss is presented. In order to establish classes of equivalence, we perform a t-test (see Table 6) to compare the coders between each other. The definition of the confidence interval is the same as in section 3.2.1. The MSE value is 0.61281, so the CI value is 0.0959. This value is the same whatever the coder, the percentage of packet loss or the hearing session.

Ranking	Monaural hearing	Diotic hearing
1	Direct	Direct
2	G.729.1 32 <=>	G.722 <=> G.729.1
	AMR-WB 23.85<=> G.722	32 <=> AMR-WB 23,85
3	G.722 <=> AMR-WB 12.65	AMR-WB 12.65 <=> G.729.1 16
4	AMR-WB 12.65 <=>G.729.1 16	

Table 6. Ranking of the coders without packet loss

To compare a coder between the two hearings, we perform a second t-test (CI= 0.0959). Please note that the two sessions have not been performed at the same time:

- AMR-WB 23.85 Monaural is equivalent to AMR-WB 23.85 Diotic. Idem for G.729.1 32.
- AMR-WB 12.65 Monaural is not equivalent to AMR12.65 Diotic. Idem for G.722, G.729.1 16 and Direct.

4.2.2. With 3% packet loss

In Figure 8, the MOS for 3 % packet loss is presented.

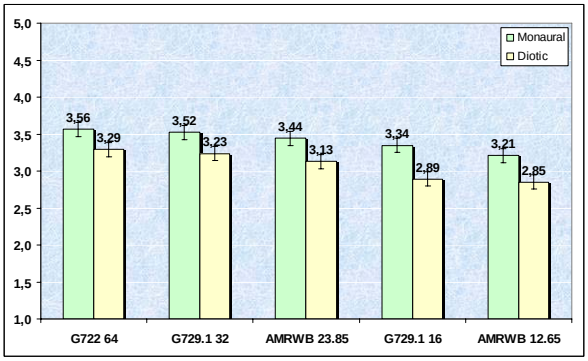

Figure 8: MOS for each coder in monaural hearing (left h.) and diotic hearing (right h.) for 3 % packet loss

In order to establish classes of equivalence, we perform a t-test to compare the coders between each other (see Table 7).

Ranking	Monaural hearing	Diotic hearing
1	G.722 <=> G.729.1 32 <=> AMR-WB 23.85	G.722 <=> G.729.1 32
2	AMR-WB 23.85 <=> G.729.1 16	G.729.1 32 <=> AMR-WB 23.85
3	G.729.1 16 <=>	G.729.1 16 <=>

	AMR-WB 12.65	AMR-WB 12.65

Table 7 : Ranking of the coders for 3 % packet loss

To compare a coder between the two hearings, we perform a t-test. Please note that the two sessions have not been performed at the same time. The results show that the monaural hearing is not equivalent to the diotic hearing for each coder.

4.2.3. With 6% packet loss

In Figure 9, the MOS for 6 % packet loss is presented.

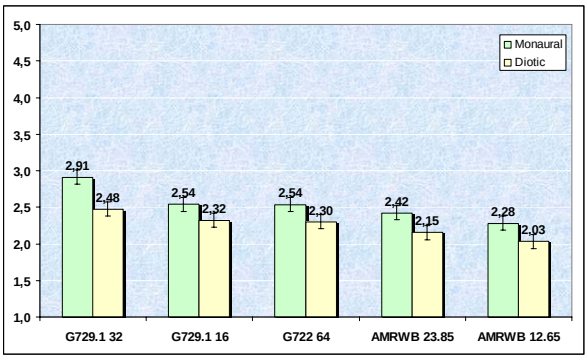

Figure 9: MOS for each coder in monaural hearing (left h.) and diotic hearing (right h.) for 6 % packet loss

In order to establish classes of equivalence, we perform a second t-test to compare the coders between each other (see Table 8).

Ranking	Monaural hearing	Diotic hearing
1	G.729.1 32	G.729.1 32
2	G.729.1 16 <=> G.722 <=> AMR-WB 23.85	G.729.1 16 <=> G.722
3	AMR-WB 12.65	AMR-WB 23.85 <=> AMR-WB 12.65

Table 8 : Ranking of the coders for 6 % packet loss

To compare a coder between the two hearings, we perform a t-test. Please note that the two sessions have not been performed at the same time. The results show that the monaural hearing is not equivalent to the diotic hearing for each coder.

5. DISCUSSION

This study highlights two different results. First, in narrowband, coders ranking can change according to the hearing system. Last, in wideband, diotic hearing, compared with monaural hearing, seems to help subjects to better distinguish differences between coders. Let's examine more precisely the results.

In narrowband, the most surprising results involve G.711 and G.729.1 at 12 kbits/sec. Indeed, without packet loss, G.711 passes from the fourth position in monaural hearing to the second one in diotic hearing, and G.729.1 at 12 kbits/sec from the first one to the third one. It is the same effet at 3 % packet loss but it is less obvious at 6 %.

In wideband, whatever the packet loss are, results are better distinguished in diotic hearing than in monaural hearing. Overall, better the bitrate is, better ranked the coder is. The better behaviour of G.729.1 in presence of packet loss find itself back as before in [11]. Let's try to explain those results.

First of all, the two sessions of each test do not happen at the same instant. So, it is difficult to assert that a coder becomes better or worst. However, the ranking of each session can not be challenged and can help us to put forward few assumptions, coming under psychoacoustics.

Although our method of normalization can not be questioned due to our application, it is one of the lead to explain the G.711 ranking. Especially in low frequencies and in the absence of speech, the factor to attenuate by 10 dB SPL is too much for the diotic hearing (see Figure 2). So G.711 white quantization noise is less disturbing in diotic hearing than in monaural hearing.

After the choice of the method of normalization, the value in dB SPL of the attenuation can be challenged. Our value was justified in 2.3 but those tests highlight the importance of the listening level. It could be interesting to perform a test with different sessions whose listening levels are different. One hypothesis is that ranking (e.g. G.711 ranking) could change and at one sound level, the diotic ranking could be uncovered.

Masking phenomenon can be put forward too. Indeed, e.g. for g.711, speech and noise are mixed and masking can play a part of listener decision.

The hearing system could have a psychoacoustics impact on both tests. Apart from the normalization method and its associated sound level, it is clear that in diotic hearing, each ear is less "upset" than the selected ear for monaural hearing. Probably, subjects feel less the aggressiveness of noise in diotic hearing. Moreover, in the diotic session, the noise is located at the center of the head, making hearing perhaps more real and less disturbing. Those two last hypotheses were highlighted by subjects and could justify a better ranking of G.711 in diotic hearing.

The hearing system could be more selective for CELP (Code Excited Linear Predictive) embedded coders such as G.729.1, AMR or AMR-WB. Due to their speech distortion trend, diotic hearing could be more selective in this kind of drawback. Indeed, it is well-known [12] that diotic hearing enables a better discrimination on pure tones than monaural hearing for the same content played on one and two ears. Pure tones are detected 3 dB lower in diotic hearing than in monaural hearing. This hypothesis could explain a decline in perceived quality of CELP embedded coders in diotic hearing. In our tests and without forgetting that they were not compared in the same session, it can be emphasized that Direct condition is always better scored in diotic hearing than in monaural hearing.

However for AMR at 12.2 kbits/sec, listeners judge it with a good opinion score in both sessions. It could be justified by a minor speech distortion compared with G.729.1 distortion.

Overall, this test highlights the difficulty to know what impact the most on subjects. Numerous factors have an influence on listener rating decision, e.g. in narrowband: G.711 white quantization noise, G.729.1 silence cleaning, low-bitrate coders, sound level etc. This kind of information could be interesting to evaluate and could help developers to create a coder. A study about the different category of noise detected by listeners can be useful to propose a new kind of test.

6. CONCLUSION

In this paper, tests showed that the hearing has an impact on quality test results. It can increase or decrease the differences between tested coders and even change their ranking. Some leads are put forward but no conclusion can be drawn. Obviously, we can not say that the ranking of one of the monaural or diotic hearings is the good one. Each ranking is the best one for the relevant hearing. We can only assert that it is important to be aware of the impact of the hearing system and its associated sound level. We have to choose a coder according to the application we aim at, its kind of hearing, the ranking of the coder for this hearing system and perhaps, when it is possible, with the supposed most suitable sound level.

7. ACKNOWLEDGEMENTS

The authors want to thank Martine Apperry[1] for her help with test adjustments.

8. REFERENCES

1. ITU-T, *Rec. P.800 Methods for Subjective Determination of Transmission Quality*. 1996.
2. Fletcher, H., Munson, WA, *Loudness, its definition, measurement and calculation.* JASA 5, 1933: p. 82-108.
3. Reynolds, G., Stevens, S., *Binaural Summation of Loudness.* JASA, 1960. **32**(10).
4. Scharf, B., Houtsma, AJM, *Audition II. Loudness, pitch, localization, aural distortion, pathology*, in *Handbook of perception and human performance*, K. Boff, Kaufman, L, Thomas, JP, Editor. 1986: New York.
5. ITU-T, *REC. G.729.1 G.729 based Embedded Variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729.* 2006.
6. 3GPP, *TS 26.073 AMR speech Codec; C-source code. V.5.1.0.* 2003.
7. ITU-T, *ITU-T Software Tool Library 2005 User's Manual*. 2005, Geneva: ITU.
8. ITU-T, *Rec. G.711 Appendix I (09/99) : A high quality low-complexity algorithm for packet loss concealment with G.711*. 1999.
9. 3GPP, *TS 26.173 ANSI-C code for the Adaptive Multi-Rate - Wideband (AMR-WB) speech codec. Version 5.8.0.* 2003.
10. ITU-T, *Rec. G.722 (1988) Appendix IV (11/06) : A low-complexity algorithm for packet loss concealment with G.722.* 2006.
11. DECT-NG07_033. *ITU-T G.722 PLC selection phase: additional information.* in *NG_DECT #7*. 2006.
12. Jesteadt, W., Wier, C., *Comparison of monaural and binaural discrimination of intensity and frequency*. JASA, 1977. **61**(6).