A standardised test to evaluate audio-visual speech intelligibility in French

Loïc Le Rhun[a,*], Gerard Llorach[b,*], Tanguy Delmas[a,*], Clara Suied[c], Luc H Arnal[a], Diane S Lazard[a,d]

[*] *These authors contributed equally*

Affiliations:

[a] *Institut de l'Audition, Inserm unit 1120, Institut Pasteur, Paris, France;* [b] *Auditory Signal Processing, Dept. of Medical Physics and Acoustics, University of Oldenburg Oldenburg, Germany;* [c] *Institut de Recherche Biomédicale des Armées, Département Neurosciences et Sciences Cognitives, Brétigny-sur-Orge, France;* [d] *Institut Arthur Vernes, ENT surgery department, Paris, France*

Mails: loiclerhunca@gmail.com; Gerard.llorach.to@uni-oldenburg.de; clarasuied@gmail.com; tanguy.delmas@pasteur.fr; luc.arnal@pasteur.fr; dianelazard@yahoo.fr

Corresponding author: Tanguy Delmas, tanguy.delmas@pasteur.fr

A standardised test to evaluate audio-visual speech intelligibility in French

Objective: Lipreading, which plays a major role in the communication of the hearing impaired, lacked a French standardised tool. Our aim was to create and validate an audio-visual (AV) version of the French Matrix Sentence Test (FrMST).

Design: Video recordings were created by dubbing the existing audio files.

Sample: Thirty-five young, normal-hearing participants were tested in auditory and visual modalities alone (Ao, Vo) and in AV conditions, in quiet, noise, open and closed-set response formats.

Results: Lipreading ability (Vo) varied from 1% to 77%-word comprehension. The absolute AV benefit was 9.25 dB SPL in quiet and 4.6 dB SNR in noise. The response format did not influence the results in the AV noise condition, except during the training phase. Lipreading ability and AV benefit were significantly correlated.

Conclusions: The French video material achieved similar AV benefits as those described in the literature for AV MST in other languages. For clinical purposes, we suggest targeting SRT80 to avoid ceiling effects, and performing two training lists in the AV condition in noise, followed by one AV list in noise, one Ao list in noise and one Vo list, in a randomised order, in open or close set-format.

5 Key words: Speechreading, Lipreading, audio-visual interaction, French Matrix Sentence Test, dubbing

Introduction

Lipreading plays a major role in the communication of the hearing impaired but has been little explored, both from the point of view of its usefulness in everyday life (Pimperton et al., 2017) and of its predictive value in the case of hearing rehabilitation by cochlear implants (Lazard & Giraud, 2017) or conventional hearing aids (Dell'Aringa et al., 2007). During the COVID pandemic, the general population became increasingly aware of the importance of lipreading even among normal-hearing listeners, when wearing facial masks (Bottalico et al., 2020). Several studies have demonstrated that the combination of masks and background noise has a negative impact on speech intelligibility, even in normal listeners (Sönnichsen et al., 2022; Tofanelli et al., 2022; Yi et al., 2021). Yi et al. (2021) showed a decrease of 20% of comprehension of sentences in normal listeners when adding a mask to the locutor. These results underpin the limits of understanding speech in adverse listening conditions when relying on the auditory modality alone (unimodal condition, Ao). Thus, visual cues help disambiguate the auditory input thanks to lips' movements (Arnal et al., 2009; Bourguignon et al., 2020; Erber, 1975) and facial expressions (Kessous et al., 2010).

Auditory tests help hearing professionals in their diagnosis (Jerger & Hayes, 1977; Joly et al., 2022) and in their choice of rehabilitation (Niemeyer, 1976). While many tests assessing speech understanding in Ao are available and validated in the international literature (e.g., logatomes, phonemes, words, sentences, in quiet or in background noise), calibrated automatized audio-visual testing material is scarce (Llorach et al., 2022; van de Rijt et al., 2019) and does not exist in French language. Adding an audio-visual assessment tool thus seems important to explore deafness and its compensation, and more generally audio-visual synergy in deprived and non-deprived subjects.

A wide variety of speech audiometry tests are available in French, from words in quiet (e.g., Fournier & Aubin, 1951; Lafon, 1972) to sentences in noise (e.g., MRT: Modified Rhyme

Test (Zimpfer et al., 2020), FIST: French Intelligibility Sentence Test (Luts et al., 2008), FrMST: French Matrix Sentence Test (Jansen et al., 2012). The FrMST, which is the French adaptation of the international MST, caught our attention because it has been widely used in clinical practice in France since its validation in 2012 (Jansen et al., 2012), and is available in different languages (for comparison purposes). Its stereotypical pattern –28 lists of 10 sentences of 5 words, each generated from 50 words (10 nouns + 10 verbs + 10 numbers + 10 objects + 10 colours)– is of interest when performing neurofunctional imaging, such as EEG (Luck, 2014). In practice, the MST uses a staircase procedure presenting the sentences in a fixed background noise (Brand et al., 2011; Brand & Kollmeier, 2002). The signal presentation level varies to provide a specified speech reception threshold (SRT): 20%, 50% or 80% SRT.

Four teams have already validated an audio-visual version of the MST into their respective languages (New Zealander English, Malay, Dutch, and German) (O'Beirne et al., 2015; Trounson, 2012; Jamaluddin, 2016; van de Rijt et al., 2019; Llorach et al., 2022). Instead of recreating the audio-visual material, Llorach et al. (2022) dubbed the original German MST audio sentences. They then validated their material in a group of young normal-hearing subjects in open-set and closed-set formats, in quiet and in background noise, in uni- and bi-modal conditions.

Recording new MST speech material from scratch is a demanding process, which includes recording, correct cutting of the speech sentences, setting the gain of each word for equal intelligibility. This step requires an evaluation with participants and further processing of the audio files before its final validation. Moreover, newly recorded MSTs have shown up to 6 dB speech perception differences compared to previous versions, depending on the talker (Hochmuth et al., 2015). Therefore, dubbing the original validated audio sentences with videos ensures results more comparable to those of the literature. The main disadvantage of

dubbing is AV asynchronies. In their study, Llorach et al. (2022) described the recording set up and the selection process of the shortest asynchronies (ideally, less than 240 ms, which represents the upper acceptable limit before deleterious effects on speech intelligibility (Grant et al., 2003). The scripts and guidelines to record, process, and cut the video material automatically can be found in the github repository (https://github.com/gerardllorach/audiovisualdubbedMST/).

Here, we reproduced the methodology (creation and validation) of Llorach et al. (2022), and compared the results of the two studies in terms of AV training effect, test-retest variability, lipreading ability in the visual only (Vo) condition, and audio-visual gain.

Material and Method

Dubbing the auditory material and recording the videos

The French version has been validated in 2012 (Jansen et al., 2012) and is largely used in clinical practice. The initial corpus included 280 sentences. We manually discarded the sentences for which the original audio recording was distorted. From the remaining sentences, we created balanced lists of 20 sentences, i.e., each word appeared no more than twice in a list, and no sentence was repeated. Forty-five lists were created based on a final set of 150 sentences. The MATLAB scripts to create these lists can be found in the github repository (see Statements).

The same female speaker who originally recorded the auditory French version participated in the video recordings.

The videos were recorded in an anechoic room, transformed into a film studio (Figure S1). The set up followed the one described in Llorach et al. (2022). The background behind the speaker was a green screen -Chroma key. We set up a homogeneous illumination of the face

using lights (ESDDI PS055 photo studio kit, Yuehai Street, Nanshan District, Shenzhen, China). The videos were recorded with a camera at 50 fps/full HD (Sony a7S III, Sony Group Corporation, Tokyo, Japan), and a condenser microphone (MKE 600 Shotgun Microphone, Senheiser, Germany) positioned at the height of the knees in front of the speaker.

The speaker had to simultaneously say the sentence she was hearing through an earphone while she was being recorded. The two audio tracks (microphone signal with the spoken speech and the original speech) were recorded in the camera as a stereo signal together with the video. The camera recorded the speech with a 48 kHz sampling rate and a 16-bit linear pulse-code modulation (LPCM) sample format.

We used a mono-to-stereo cable (Hosa YMM261, Hosa Technology, INC, California, USA) to combine the recorded speech and the original speech into a stereo signal for the camera audio input. The recorded speech came from the audio recorder (Zoom H6, Zoom North America, NY, USA) which was receiving the signal from the condenser microphone. The audio recorder provided phantom power to the condenser microphone. The original speech was reproduced from a computer. The audio output of the computer was connected to a jack splitter: one signal for the earphone and the other for the camera. Each sentence was presented 4 times in a row to the speaker. An identification signal and three pure tones were added before this presentation to warn the speaker and help her get prepared to dub. The identification signal was used to identify the sentence number during the post-processing of the files and lasted one second. The speaker heard the three pure tones through an earphone placed in her left ear and listened to the first presentation to know the five words to repeat. She had to dub the three following sentences while simultaneously hearing them, with a neutral face, without too much exaggeration of the articulation, while keeping her eyes fixed and centred on the camera, avoiding blinking. She had to start and finish the sentences with a

still face. A training session was performed to get used to the dubbing. The recorded session lasted one day, including breaks.

To facilitate the dubbing process, we displayed a temporal representation of the acoustic stimulus on a screen placed just above the camera. This visual guide sequentially displayed three vertical bars stemming for the pure tones followed by the acoustic waveform of the sentence being played. The waveform changed from blue to red as the signal was being reproduced (Figure S1, left side). During the recording procedure, the sentences to play were selected through a user interface in MATLAB. Several takes could be recorded for the same sentence. These scripts to record the material can be found in the github repository.

The resulting video files, which contained the visual speech, the original sentences of the French MST and recorded speech, were processed automatically via scripts. During the recording session, the camera recorded continuously i.e., a video file could contain several sentence takes. Each sentence take contained an identification signal, three pure tones, and four sentence repetitions, which were present in one of the audio channels of the video files. These identification signals were used to automatically identify the sentence takes from the video files. The duration of the sentence was retrieved from the original MST audio files and each take was cut into repetitions (three sentence repetitions for each take).

Selection of the videos

As mentioned earlier, an automated video editing script sequentially cut and named the videos with the sentence code, the take number, and the repetition number. We manually discarded videos in which the speaker smiled or showed other non-neutral facial expressions.

The analysis script also allowed quantifying the asynchrony between the original auditory sentence and the sentence dubbed by the speaker during the recording, based on an asynchrony score, as described in Llorach et al. (2022). This approach uses dynamic time

warping (DTW) (Sakoe & Chiba, 1978) to obtain an asynchrony score that represents the absolute temporal difference between the mel spectrograms of the original and the dubbed sentences. It detects time offsets asynchronies (words spoken too early or too late) and/or words spoken slower or faster than the original words. It designates the best match between the recordings and the original sentences, allowing to keep the most synchronized videos. The absolute asynchrony scores per sentence based on the final selected sentences used for the AV test ranged from 26.3 ms to 276.2 ms with a median of 50.6 ms, and a mean of 61.4 ms ± 35.5 ms (see Figure S2).

Validation of the AV material

Participants

This project was approved by the Ethics committee CPP Tours-Region Centre-Ouest 1 (project identification number 2020T317 RIPH3 HPS). All the subjects gave their written consent.

The test was validated on a gender-balanced cohort of 35 normal-hearing participants (18 females and 17 males). Two participants did not perform the retest session. Participants were 20 to 29 years old (mean age: 24.4 years with a standard deviation (SD) of 2.8), with no known hearing problems or difficulty understanding in noisy environments. They had normal or corrected vision and no specific lipreading skills. All participants were native French speakers and their level of education ranged from 0 to 8 years after high school (mean: 4 years). They all undertook a pure tone audiogram (ELIOS-ECHODIA portable audiometer, Electronique du Mazet, France) at the beginning of the first session. The mean pure tone average (PTA) was 1.22 dB HL (range: -4.25; 5).

Set up

The experiment was performed in an anechoic booth. The participant was sitting on a chair wearing binaural audiometric headphones (Sennheiser HDA 300, Sennheiser electronic GmbH & Co. KG, Germany) and placed at about 80 cm (arm-distance) in front of a 27" 4k screen (LG 27UN83A-W, LG Electronics, Korea). The experiment was controlled by an experimenter sitting in the booth next to the participant. The recorded videos and the original auditory sentences were played using VLC® (version 3.0.16).

The acoustic signal was sent with a SLL 2 sound card (Solid State Logic, UK). The acoustic levels were calibrated using an artificial ear connected to a sound level meter.

To synchronize the audio and video signals, we used a video where number wrote on the screen grow at the frame rate of the screen (50 Hz), and every 25 frames a sound was played. An external camera recorded the display screen of the experiment at 50 fps. The acoustic signal was connected directly to the external camera instead of the headphones. Using the recording of the camera, we compared the audio signal, and the number written on the screen. We found a consistent delay of 3 frames, *i.e.* 60 ms between the audio and the video. This 60 ms lag was then corrected in the experiment set up.

Stimuli

Each video had a green background and started with a 500 ms still face. The participants were tested in quiet and in noise. The noise used was the long-term stationary noise of the average speech spectrum (LTASS) from the original French MST (Jansen et al., 2012; Kollmeier et al., 2015; K. Wagener et al., 2003). The adaptive procedure (Brand et al., 2011; Brand & Kollmeier, 2002) varied the presentation level of the sound material according to the participant responses. The adaptive procedure is an extended staircase method that changes its step size depending on the responses.

In quiet, the presentation level of the sentences started at 25 dB SPL, except for the first tested subject. In noise, the presentation level of the sentences started at 60 dB SPL, with a fixed noise at 65 dB SPL.

Experimental conditions

Three different modalities were tested, Audio only (Ao), Visual Only (Vo) and Audio-visual (AV) in 9 conditions. Ao and AV modalities were tested in quiet and in noise, in open-set and closed-set response formats (8 conditions). The Vo modality was presented in the LTASS noise to be more ecological (i.e., in everyday life, normal-hearing people rely on lipreading in noisy settings) and in closed-set response format (open-set format being too difficult for naive participants). In the open-set format, participants repeated aloud what they understood while the experimenter selected the correct answers on his numerical interface. In the closed-set format, participants selected their responses on a matrix presented on the screen. The interface presented all the possible 50 words, as well as a "no response" option for each category of words.

For each condition, the participant was tested with a list of 20 sentences. The adaptive procedure was set to reach an individual Speech Reception Threshold of intelligibility of 80% (SRT80), which is equivalent to 4 out of 5 correctly identified words per sentence. The SRT80 was chosen by Llorach et al. (2022) to avoid ceiling effects in the AV conditions due to the Vo contribution. Indeed, some participants are able to understand 50% or more of the content thanks to lipreading alone (Vo) (van de Rijt et al., 2019).

SRT80 is expressed in dB SPL (speech presentation level to reach 80% of intelligibility) for the quiet conditions and in dB SNR (signal-to-noise ratio presentation level to reach 80% of intelligibility) for the noise conditions. For the Vo in noise condition and closed-set format

lists, we report the percentage of correctly understood words among the 20 sentences. No feedback was given.

Procedure

A test-retest procedure (two sessions) was performed, similarly to Llorach et al. (2022). Sessions were spaced between 3 to 59 days apart. The first session lasted about 1 hour and 15 minutes and the second lasted 1 hour.

All participants undertook a training phase before each session to control for the learning effect: four AV lists of 20 sentences in noise were used during the first session (Test), and one AV list in noise for the retest (Figure 1). The response format of the training lists was either in open-set or closed-set format, randomly chosen. This response format was used for the whole training session. For the retest (session 2), only one training list was presented in the same response format as during session 1 (open-set or closed-set).

During session 1 (Test) and after the training, subjects started with the same response format (open-set or closed-set) as during the training. The order of the conditions (Ao and AV in quiet and in noise, and Vo in noise) was randomised. Once these conditions tested in the initial format, the same conditions with the other response format were tested in a randomised order.

For the retest (session 2), the same succession of response formats as in session 1 was used. Then, all the conditions were again randomly tested in quiet and in noise.

We added another condition to the second session to compare our results with those obtained by Jansen et al. (2012) who performed the validation of the French version of the MST. The Ao condition in noise and closed-set format was tested with an adaptive procedure of SRT50. This extra-condition was performed after the training list of session 2 (retest).

Statistical analysis

Results are expressed in means ± standard deviation (SD). ANOVAs for repeated measures were performed using JASP (v0.16.3, University of Amsterdam, Netherlands), followed by *post-hoc* multiple comparisons with Bonferroni corrections. If the sphericity assumption was not met, a Greenhouse Geisser correction was used. We searched for correlations between the Vo scores (lipreading skills) and the AV benefit of four AV conditions (Pearson correlations).

Results

SRT50

In 2012, Jansen et al. (2012) validated the French version of the MST with a group of 30 normal-hearing participants (mean age: 22 years, range: 20 to 29 years). The adaptive procedure described by Brand et al. (2011; 2002) was used to reach the SRT50 of the participants, in a closed-set format. Their publication validated a mean SRT50 of -6 dB SNR (±0.6). We performed the same evaluation and obtained a mean SRT50 of -7.75 dB SNR (±0.79).

Evaluation of the training effect

The use of the MST starts with a systematic training session to get used to the type of material and background noise. The audio French version advises two training lists (Jansen et al., 2012). In the validation of the AV version of the German MST, Llorach et al (2022) tested four lists during the training phase of the test session and one list during the retest session (AV Noise condition, closed or open-set format). We applied the same protocol. In the present study, the test and retest sessions were spaced 17 days apart on average (range: 3-59).

Figure 1 shows the training effect (mean SRT80% in dB SNR for the AV condition in noise in the response format performed, open-set or closed-set) during the test and retest sessions, in

the present study. On average, the training effect between the first training list and the test was -4.5 dB SNR during the first session, regardless of the response format, and -1.93 dB SNR between the training retest and retest lists. During the second session, participants retained similar SRT scores as the ones obtained for the fourth training list of the first session. The repeated-measures ANOVA with a Greenhouse-Geisser correction (Maulchy's test X^2(5) = 22.4, $p < 10^{-3}$) showed a significant effect of training lists ($F(2, 65.90) = 51.7$, $p < 10^{-3}$) independently of the response format. *Post hoc* multiple comparisons with Bonferroni corrections showed that the average SRT of the first list of the first session was significantly different from those of the three other training lists ($p<10^{-3}$, Figure 1). The subsequent SRTs of the training lists 2 to 4 did not significantly differ. The repeated-measures ANOVA showed a significant improvement during the second session between the training retest and the retest lists ($F(1, 31) = 25$, $p < 10^{-3}$). There was a significant score difference of 1.5 dB SNR (±0.4) between the open-set and closed-set formats during the first training session ($F(1, 33) = 11.6$, $p = 0.002$), but the interaction with the number of training lists was not significant ($F(2, 65.90) = 2.58$, $p = 0.084$).

Audio-only (Ao) and Audio-visual evaluation

Table 1 shows the mean SRTs (±SD) obtained in the different Ao and AV conditions (noise/quiet, closed/open response format) averaging test and retest results (except for the two participants who only performed the first session). There was no effect of the response formats.

The absolute SRT benefit obtained by adding the visual modality to the auditory alone condition (AV - Ao) was 4.6 dB SNR in noise and 9.25 dB SPL in quiet (mean between the two response formats).

The audio-visual conditions led some participants to reach a SRT80 at speech presentation levels below 0 dB SPL in quiet and SNRs below -20 dB SNR in noise (Figure S3). At these levels, it is impossible to obtain any acoustic cue from the auditory modality. These scores do not represent audio-visual speech perception, but only lipreading. For data analysis, we did not exclude these outliers (very proficient lipreaders) but limited their SRT scores to -20 dB SNR and 0 dB SPL, as Llorach et al. (2022) did. Indeed, for the female German MST, the speech detection threshold has been evaluated at -16.9 dB SNR in the audio-only condition in noise (Schubotz et al., 2016). This threshold can be theoretically lowered by -3 dB when adding visual speech, as demonstrated in Bernstein et al. (2004). No speech detection threshold in noise has been published for the French MST. We consequently decided to extrapolate the German values to our results in French. The outlier scores represented 3 scores from 3 different subjects out of 445 SRTs (0.7%) (*versus* 5% in Llorach et al. (2022).

Lipreading (Vo) and audio-visual benefit

Participants had a wide range of lipreading abilities (Figure 2). The individual Vo scores in noise and closed-set format ranged from 1% to 77%-word intelligibility, with an average of 51.3% (±16.2%), in accordance with that of Llorach et al. (2022) in the same testing conditions (50% ±21.4%). There was an average intelligibility improvement between the test and the retest sessions of 4.8 % (i.e., 5 words over 100 presented during the 20 sentences). This improvement was significant (paired t-test: $p<10^{-3}$). Lipreading scores (Vo) correlated with the audio-visual benefit (SRT difference between audio-visual and audio-only condition, $p <10^{-3}$) in all AV conditions (Figure S4). The Pearson r ranged from -0.55 (in quiet and open-set format) to r =-0.74 (in noise and open-set format).

Test-retest differences

To assess test-retest differences, we compared the within-subject standard deviations (test minus retest) to the between-subject standard deviations. According to the 2σ criterion, if the between-subject standard deviation is larger than the double of the within-subject standard deviation, it is possible to distinguish if a participant performs differently from another (K. C. Wagener & Brand, 2005). Figure 3 shows the within- and between-subject standard deviations of SRTs in the conditions tested, except for the Vo condition, which is expressed as a percentage of correct words. The standard deviations of the within-subject differences (test minus retest) are shown as grey bars. The between-subjects standard deviations are shown as white bars. The black line represents the 2σ criterion, which is the double of the within-subject standard deviation. Except for the Vo condition, none of the conditions exceeded the 2σ criterion in both studies. These results mean that it was possible to distinguish proficient and unskilled lipreaders.

Discussion

The present study designed a new audio-visual material to evaluate lipreading skills and audio-visual benefit in a standardized manner and tested it in a young normal-hearing sample. We used the French audio MST corpus validated by Jansen et al. (2012), and the methodology described in Llorach et al. (2022). We compared our results with those two studies.

Population

The samples of the studies were similar in terms of age (mean 24.4 +/-2.8 years) and gender balance. We tested 8 more subjects than Llorach et al. (2022) and 5 more subjects than Jansen et al. (2012). They all had normal hearing.

SRT50

A difference of -1.75 dB SNR was observed between the current study (-7.75 ± 0.79 dB SNR) and Jansen et al. (2012) (-6 ± 0.6 dB SNR) when measuring SRT50 in audio-only and closed-set format. Our subjects outperformed. This could be due to differences in calibration and technical configuration, and/or timing of the SRT50 assessment. The SRT50 assessment was performed during the second session (retest). A greater training effect may have impacted these results.

Validation of the visual material

The absolute SRT benefit obtained by adding the visual modality to the auditory alone condition (AV - Ao) was 4.6 dB SNR in noise and 9.25 dB SPL in quiet in the present study, *versus* 5.0 dB SNR in noise and 7.0 dB SPL in quiet in Llorach et al. (2022), and 3 dB SNR in noise in Van de Rijt et al. (2019). Our 4.6 dB difference in the noise condition is in accordance with the literature and validates our video material for AV assessment. It also shows that the audio-visual asynchronies of the material, which were slightly smaller in the German AV version, did not impact on the results.

Compared to the SRT scores reported in Llorach et al. (2022), the SRT scores of the present study were higher in the AV and Ao conditions, meaning that the test was more difficult in French than in German. Similar differences between these languages have been mentioned before when assessing SRT50 (Kollmeier et al., 2015).

During the Visual-only condition, some subjects scored up to 77% of comprehension. The average was 51.3% (±16.2%) in the present study, *versus* 50% (± 21.4%, range: 0%-84%) in Llorach et al. (2022). This is why Llorach et al. (2022) choose to use 80% rather than 50% of intelligibility. Indeed, if the AV condition used a SRT of 50%, an important proportion of the subjects would have been able to lipread half of the material without using acoustic information. Still, in both studies, some subjects outperformed, achieving understanding at a

level where the sentences were not audible anymore. Llorach et al. (2022) decided to limit these values to the level where acoustic information disappears (-20 dB SNR and 0 dB SPL) rather than excluding these data points. It would have been equivalent to removing the best audio-visual scores. This method avoided unrealistic audio-visual SNR benefits for the analysis but maintained variability.

In the current study, only 0.7% of the AV SNR were limited compared to 5% in Llorach et al (Llorach et al., 2022). This difference could be explained by the fact that we used the German values as a cut-off in the present study, as no speech detection threshold in noise was available for the French material.

Another hypothesis is that our speaker was harder to lipread than the German actress, and/or that French is less accessible through visual cues. However, we carefully checked that the actress articulated the words naturally and remained neutral during the dubbing. It is known that there is a significant effect of the talker on lipreading performance (Yakel et al., 2000). A way to explore this variability would be to dub with different actors.

Training effect

An improvement of 1.8 dB SNR between the 1st and the sixth list is expected in the audio-only French MST, i.e., for SRT50 (Jansen et al., 2012). Here, we found a 4.5 dB SNR improvement for the SRT80 in the AV condition in noise in closed-set and open-set format, between the first training list and the test list (fifth list). This greater effect probably arose from the participants learning to lipread the material. Indeed, Lander & Davies (2008) demonstrated that lipreading performance increased overall with practice, but that performance increased significantly more as participants became increasingly familiar with the same speaker. Accordingly, there was an average intelligibility improvement in the Vo

only condition between the test and the retest session of 4.8% in the present study and 6.1% in Llorach et al. (2022).

The training effect in the AV condition in noise statistically disappeared after the two first training lists. During the second training session, participants retained similar SRT scores as the one obtained for the fourth list of the first session, similarly to Llorach et al. (2022).

Test-retest differences

We found a small within- and between-subjects variability in the audio-only in noise conditions, which was expected from a homogenous group of young normal-hearing participants. This variability increased in quiet conditions. These results are in accordance with the literature (Llorach et al., 2022; Smoorenburg, 1992; Souza et al., 2007).

The Visual-only scores were highly variable across subjects, as expected (Jamaluddin, 2016; Llorach et al., 2022; van de Rijt et al., 2019). A larger between and within subjects' variability was found in the AV conditions in the present study as compared to Llorach et al. (2022). This variability can be explained by the different individual lipreading abilities between samples.

Response format

There was no difference between the response formats once subjects had been trained. This difference was not significant in the audio-visual German MST (Llorach et al., 2022). For most languages, a closed-set response format means lower SRTs (i.e., easier task), but not for German and Polish, as reported in Kollmeier et al. (2015).

Conclusion

We have created an audio-visual version of the French version of the MST by dubbing the original validated audio material.

The SRT80 values for young normal hearing participants in the audio-visual condition were −9.2 dB SNR in noise and 16.8 dB SPL in quiet.

The absolute audio-visual benefit was 4.6 dB SNR in noise and 9.25 dB SPL in quiet. Lipreading scores (visual-only sentences) ranged from 1% to 77%. Our results are in accordance with those already published for AV MSTs in other languages.

This AV material can be used for clinical purposes, targeting SRT80 instead of SRT50 to avoid ceiling effects related to visual information.

In practice, it is not possible to carry out all the conditions described here, especially in hearing impaired subjects. For clinical purposes, we suggest performing two AV training lists in noise (closed-set or open-set format, not significant difference in French), followed in a randomised order using the same response format as the one of the training lists by: one list in audio-visual condition in noise, one list in audio only condition in noise, and one list in visual only condition in noise.

Further research will evaluate the Audio-visual FrMatrix on hearing-impaired cohorts, and especially their audio-visual gain in quiet and in noise.

Acknowledgements: We thank Marie-Amélie Bizouard and Gregory Gerenton for their technical help, Jan Wouters and Volker Hohmann for relevant discussions. This work was funded by Fondation Pour l'Audition FPA RD-2020-10 (for LLR, TD, LHA, DL) and by the Deutsche Forschungsgemeinschaft (for GL, DFG, Cluster of Excellence EXC 1077/1 "Hearing4all", and SFB1330 Projects B1 and C4).

Statements: The video recordings for the female French Matrix Sentence Test will be given free of charge upon request to the corresponding author (Diane Lazard) if the new users certify they will cite the present article and the following:

Jansen, S., Luts, H., Wagener, K. C., Kollmeier, B., Del Rio, M., Dauman, R., James, C., Fraysse, B., Vormès, E., Frachet, B., Wouters, J., & van Wieringen, A. (2012). Comparison of three types of French speech-in-noise tests: A multi-center study. *International Journal of Audiology*, *51*(3), Art. 3. https://doi.org/10.3109/14992027.2011.633568

Llorach, G., Kirschner, F., Grimm, G., Zokoll, M. A., Wagener, K. C., & Hohmann, V. (2022). Development and evaluation of video recordings for the OLSA matrix sentence test. *International Journal of Audiology*, *61*(4), 311□321. https://doi.org/10.1080/14992027.2021.1930205

The audio recordings are not provided in the dataset, but research licenses are available from Hörzentrum Oldenburg gGmbH. Please refer to Hörzentrum Oldenburg gGmbH for the audio material (https://www.hz-ol.de/en/matrix.html).

The software that links audio and video in the design process is available open-source at GitHub - gerardllorach/audiovisualdubbedMST. This is a repository with guidelines and code

to create visual material for Matrix Sentence Test for speech audiometry.

https://github.com/gerardllorach/audiovisualdubbedMST

The research software that runs the test is available for collaborative work with Hörzentrum Oldenburg gGmbH. Please contact sales@hz-ol.de.

References

Arnal, L. H., Morillon, B., Kell, C. A., & Giraud, A.-L. (2009). Dual neural routing of visual facilitation in speech processing. *Journal of Neuroscience*, *29*(43), Art. 43.

Bernstein, L., Auer, E., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*, 5–18. https://doi.org/10.1016/j.specom.2004.10.011

Bottalico, P., Murgia, S., Puglisi, G. E., Astolfi, A., & Kirk, K. I. (2020). Effect of masks on speech intelligibility in auralized classrooms. *The Journal of the Acoustical Society of America*, *148*(5), Art. 5. https://doi.org/10.1121/10.0002450

Bourguignon, M., Baart, M., Kapnoula, E. C., & Molinaro, N. (2020). Lip-reading enables the brain to synthesize auditory features of unknown silent speech. *The Journal of Neuroscience*, *40*(5), 1053–1065. https://doi.org/10.1523/JNEUROSCI.1101-19.2019

Brand, T., Kissner, S., Jürgens, T., Berg, D., & Kollmeier, B. (2011, mars 5). *Adaptive Algorithmen zur Bestimmung der 80%-Sprachverständlichkeitsschwelle.*

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *The Journal of the Acoustical Society of America*, *111*(6), Art. 6. https://doi.org/10.1121/1.1479152

Dell'Aringa, A., Adachi, E., & Dell'Aringa, A. (2007). Lip reading role in the hearing aid fitting process. *Brazilian journal of otorhinolaryngology*, *73*, 95–99. https://doi.org/10.1016/S1808-8694(15)31129-0

Erber, N. P. (1975). Auditory-visual perception of speech. *The Journal of Speech and Hearing Disorders*, *40*(4), 481–492. https://doi.org/10.1044/jshd.4004.481

Fournier, J.-E., & Aubin, A. P. (1951). *Audiométrie vocale : Les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités*. Maloine.

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2003). *Discrimination of Auditory-Visual Synchrony : 2003 International Conference on Audio-Visual Speech Processing, AVSP 2003*. 31 35. http://www.scopus.com/inward/record.url?scp=85133580799&partnerID=8YFLogxK

Hochmuth, S., Kollmeier, B., Brand, T., & Jürgens, T. (2015). Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests. *International Journal of Audiology*, *54*(sup2), Art. sup2. https://doi.org/10.3109/14992027.2015.1046502

Jamaluddin, S. A. (2016). *Development and evaluation of the digit triplet and auditory-visual matrix sentence tests in Malay*.

Jansen, S., Luts, H., Wagener, K. C., Kollmeier, B., Del Rio, M., Dauman, R., James, C., Fraysse, B., Vormès, E., Frachet, B., Wouters, J., & van Wieringen, A. (2012). Comparison of three types of French speech-in-noise tests : A multi-center study. *International Journal of Audiology*, *51*(3), Art. 3. https://doi.org/10.3109/14992027.2011.633568

Jerger, J., & Hayes, D. (1977). Diagnostic Speech Audiometry. *Archives of Otolaryngology*, *103*(4), 216 222. https://doi.org/10.1001/archotol.1977.00780210072008

Joly, C.-A., Reynard, P., Mezzi, K., Bakhos, D., Bergeron, F., Bonnard, D., Borel, S., Bouccara, D., Coez, A., Dejean, F., Del Rio, M., Leclercq, F., Henrion, P., Marx, M., Mom, T., Mosnier, I., Potier, M., Renard, C., Roy, T., … Thai-Van, H. (2022). Guidelines of the French Society of Otorhinolaryngology-Head and Neck Surgery (SFORL) and the French Society of Audiology (SFA) for Speech-in-Noise Testing in Adults. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, *139*(1), 21 27. https://doi.org/10.1016/j.anorl.2021.05.005

Kessous, L., Castellano, G., & Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, *3*(1□2), 33□48. https://doi.org/10.1007/s12193-009-0025-5

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test□: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, *54 Suppl 2*, 3□16. https://doi.org/10.3109/14992027.2015.1020971

Lafon, J. C. (1972). Phonetic test, phonation, audition. *JFORL. Journal francais d"oto-rhino-laryngologie; audiophonologie et chirurgie maxillo-faciale*, *21*(3), 223□229. Scopus.

Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability? *Quarterly Journal of Experimental Psychology*, *61*(7), Art. 7. https://doi.org/10.1080/17470210801908476

Lazard, D. S., & Giraud, A.-L. (2017). Faster phonological processing and right occipito-temporal coupling in deaf adults signal poor cochlear implant outcome. *Nature Communications*, *8*(1), Art. 1. https://doi.org/10.1038/ncomms14872

Llorach, G., Kirschner, F., Grimm, G., Zokoll, M. A., Wagener, K. C., & Hohmann, V. (2022). Development and evaluation of video recordings for the OLSA matrix sentence test. *International Journal of Audiology*, *61*(4), 311□321. https://doi.org/10.1080/14992027.2021.1930205

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique, second edition*. MIT Press.

Luts, H., Boon, E., Wable, J., & Wouters, J. (2008). FIST□: A French sentence test for speech intelligibility in noise. *International Journal of Audiology*, *47*(6), Art. 6. https://doi.org/10.1080/14992020801887786

Niemeyer, W. (1976). Speech Audiometry and Fitting of Hearing Aids in Noises. *Audiology*, *15*(5), Art. 5. https://doi.org/10.3109/00206097609071802

O'Beirne, G. A., Trounson, R. H., McClelland, A. D., Jamaluddin, S. A., & Maclagan, M. A. (2015). Development of an auditory-visual matrix sentence test in New Zealand English. *Journal of International Advanced Otology*, *11*(Supplement 1), Art. Supplement 1.

Pimperton, H., Ralph-Lewis, A., & MacSweeney, M. (2017). Speechreading in Deaf Adults with Cochlear Implants□: Evidence for Perceptual Compensation. *Frontiers in Psychology*, *8*. https://www.frontiersin.org/article/10.3389/fpsyg.2017.00106

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), Art. 1. https://doi.org/10.1109/TASSP.1978.1163055

Schubotz, W., Brand, T., Kollmeier, B., & Ewert, S. D. (2016). Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features. *The Journal of the Acoustical Society of America*, *140*(1), Art. 1. https://doi.org/10.1121/1.4955079

Smoorenburg, G. F. (1992). Speech reception in quiet and in noisy conditions by individuals with noise□induced hearing loss in relation to their tone audiogram. *The Journal of the Acoustical Society of America*, *91*(1), 421□437. https://doi.org/10.1121/1.402729

Sönnichsen, R., Llorach Tó, G., Hochmuth, S., Hohmann, V., & Radeloff, A. (2022). How Face Masks Interfere With Speech Understanding of Normal-Hearing Individuals□: Vision Makes the Difference. *Otology & Neurotology: Official Publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology*, *43*(3), Art. 3. https://doi.org/10.1097/MAO.0000000000003458

Souza, P. E., Boike, K. T., Witherell, K., & Tremblay, K. (2007). Prediction of Speech Recognition from Audibility in Older Listeners with Hearing Loss□: Effects of Age, Amplification, and Background Noise. *Journal of the American Academy of Audiology*, *18*(01), 054□065. https://doi.org/10.3766/jaaa.18.1.5

Tofanelli, M., Capriotti, V., Gatto, A., Boscolo-Rizzo, P., Rizzo, S., & Tirelli, G. (2022). COVID-19 and Deafness□: Impact of Face Masks on Speech Perception. *Journal of the American Academy of Audiology*, *33*(2), 98□104. https://doi.org/10.1055/s-0041-1736577

Trounson, R. H. (2012). *Development of the UC Auditory-Visual Matrix Sentence Test*.

van de Rijt, L. P. H., Roye, A., Mylanus, E. A. M., van Opstal, A. J., & van Wanrooij, M. M. (2019). The Principle of Inverse Effectiveness in Audiovisual Speech Perception. *Frontiers in Human Neuroscience*, *13*, 335. https://doi.org/10.3389/fnhum.2019.00335

Wagener, K. C., & Brand, T. (2005). Sentence intelligibility in noise for listeners with normal hearing and hearing impairment□: Influence of measurement procedure and masking parameters. *International Journal of Audiology*, *44*(3), Art. 3. https://doi.org/10.1080/14992020500057517

Wagener, K., Josvassen, J. L., & Ardenkjaer, R. (2003). Design, optimization and evaluation of a Danish sentence test in noise. *International Journal of Audiology*, *42*(1), 10□17. https://doi.org/10.3109/14992020309056080

Yakel, D. A., Rosenblum, L. D., & Fortier, M. A. (2000). Effects of talker variability on speechreading. *Perception & Psychophysics*, *62*(7), Art. 7. https://doi.org/10.3758/BF03212142

Yi, H., Pingsterhaus, A., & Song, W. (2021). Effects of Wearing Face Masks While Using Different Speaking Styles in Noise on Speech Intelligibility During the COVID-19

Pandemic. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/article/10.3389/fpsyg.2021.682677

Zimpfer, V., Andéol, G., Blanck, G., Suied, C., & Fux, T. (2020). Development of a French version of the Modified Rhyme Test. *The Journal of the Acoustical Society of America*, *147*(1), EL55□EL61. https://doi.org/10.1121/10.0000559

Figure legends:

Figure 1: Audio-visual training effect in noise during the test and retest sessions.

Four training lists were used during the test session. One list was used during the retest session.

Results are expressed in mean SRT ± SD, with black dashed lines and circles for the participants who performed the training in closed-set response format, and continuous grey line for those who performed the training in open-set response format. *: $p < 10^{-3}$. According to Bonferroni *post hoc* tests; SRTs of Training 1 were significantly different from the SRTs of the three other training lists regardless of response formats. SRTs of Training 1 were significantly different between response formats

Figure 2: Boxplot and distribution of the lipreading scores: Visual only in noise and closed-set format.

The mean and the median are represented by a red cross and a red line, respectively. The outliers are represented by smaller red crosses. The boxplot is smoothed with the Matlab Kdensity function that returns a probability density estimate.

Figure 3: Within-subject (grey bars) and between-subject (white bars) standard deviations (SD) for all conditions. The 2σ criterion is indicated as a thick black line. On the left, SDs of speech in noise conditions expressed in dB SNR; middle, SDs of the visual-only, expressed in percentage; right, SDs of speech in quiet, expressed in dB SPL.

Supplement files legends:

Figure S1: Schematic drawing of the recording setup: the camera and the screen showing the visual aid are on the left side. The audio input of the camera is connected to the mono-to-stereo jack adapter. The input of this adapter is the recorded signal (blue path) and the original speech (red path). The speaker with a green background and an earphone playing the original speech signal is in the center. The condenser microphone, below the speaker, is connected to the audio recorder. The output of the audio recorder goes to the mono-to-stereo adapter. On the right: the computer that generates the three tones and the sentence repetitions (waveforms). The audio output of the computer is connected to the earphone and to the mono input of the mono-to-stereo adapter. An HDMI cable connects the computer screen to the visual aid on top of the camera to duplicate the screen.

Figure S2: Asynchrony scores of the final selected recordings. The mean and the median are represented by a red cross and a red line, respectively. The outliers are represented by smaller red crosses. The boxplot is smoothed with the Matlab Kdensity function that returns a probability density estimate.

Figure S3: Adaptive speech presentation levels for the Audio-visual conditions in quiet (a) and adaptive signal to noise ratios for the Audio-visual conditions in noise (b).

The adaptive procedure changed the speech levels to reach 80% of intelligibility (SRT80 for a fixed background noise). Each line shows a single list of 20 sentences per participant. Below the horizontal line at 0 dB SPL (a) and at -20 dB SNR (b), participants understood speech using only visual cues.

The presentation level in quiet started at 25 dB SPL, except for the first subject of the study who started with a presentation level of 60 dB.

Figure S4: Visual Only scores as a function of the audio-visual benefit A. in noise (in dB SNR) and B. quiet (in dB SPL), in closed-set (left column) and open-set (right column) formats (individual data points).

The scores of the test and retest sessions are dissociated, represented as two open circles per subject. r and p stand for the results of the Pearson correlations.

Figure 1

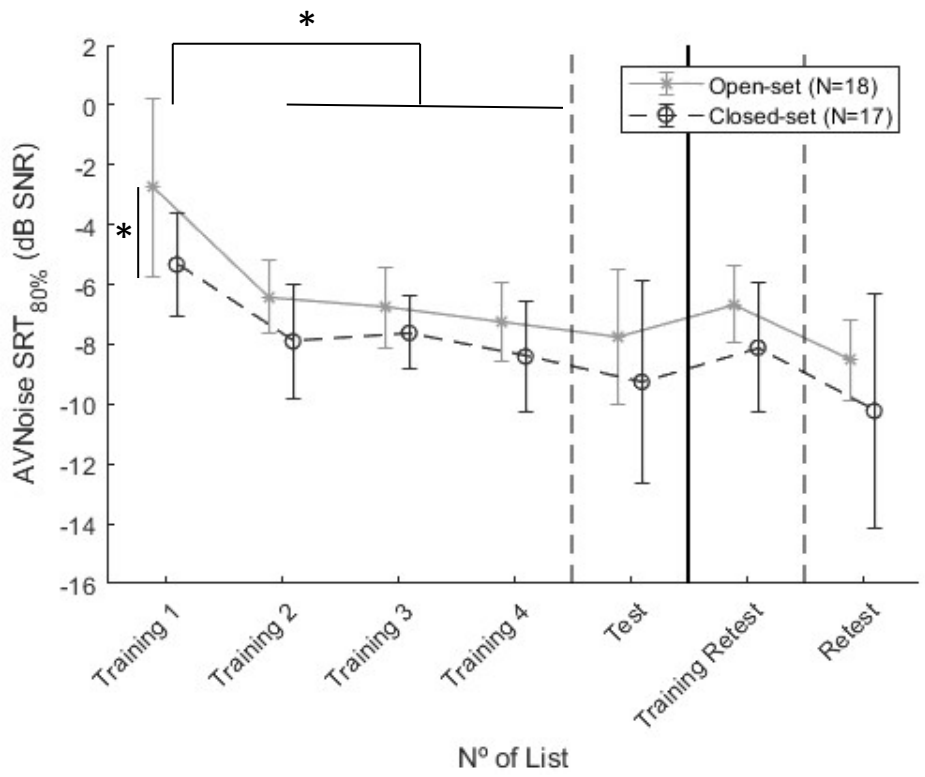
*
*
2
0
-2
-4
-6
-8
-10
-12
-14
-16
AVNoise SRT80% (dB SNR)
Open-set (N=18)
Closed-set (N=17)
Training 1
Training 2
Training 3
Training 4
Test
Training Retest
Retest
Nº of List

Figure 2

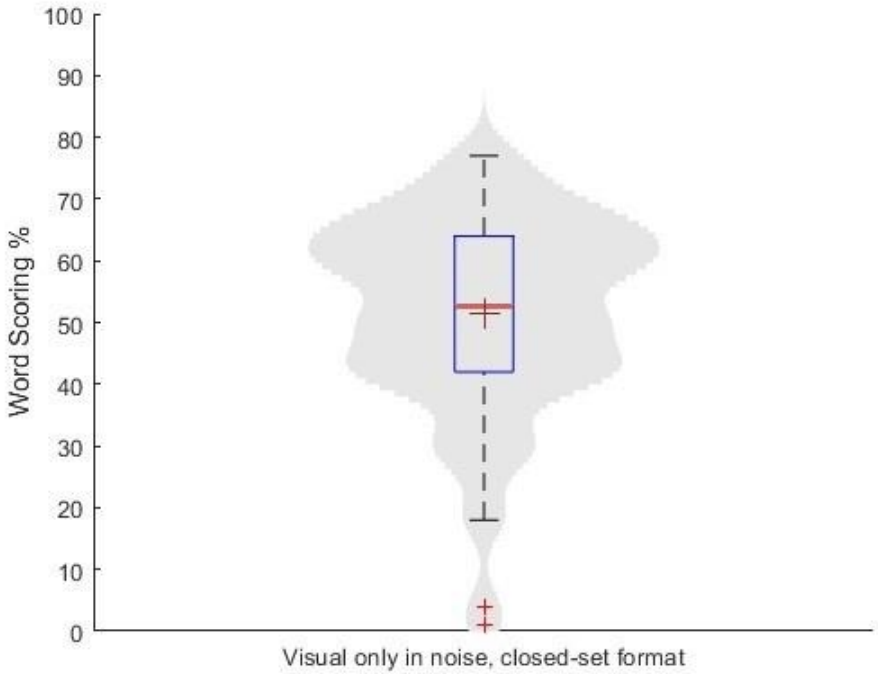
100
90
80
70
60
50
40
30
20
10
0
Word Scoring %
Visual only in noise, closed-set format

Figure 3

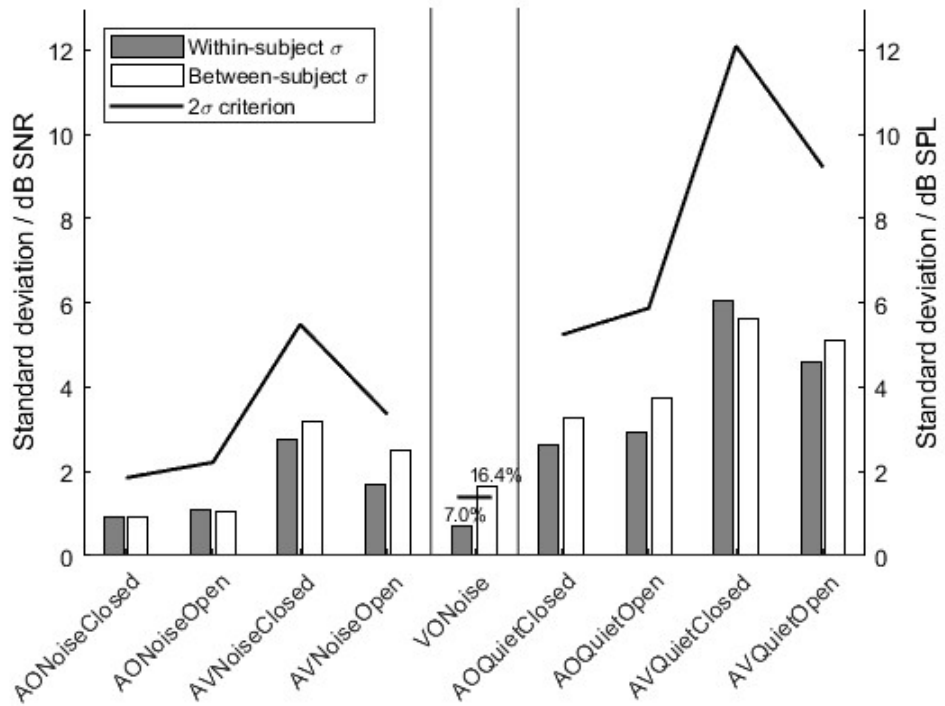
Within-subject σ
Between-subject σ
2σ criterion
Standard deviation / dB SNR
Standard deviation / dB SPL
0
2
4
6
8
10
12
7.0%
16.4%
AONoiseClosed
AONoiseOpen
AVNoiseClosed
AVNoiseOpen
VONoise
AOQuietClosed
AOQuietOpen
AVQuietClosed
AVQuietOpen

Table 1: Mean Speech Reception Thresholds SRT80% (±Standard Deviation) obtained in the different Audio-only (Ao) and Audio-visual (AV) conditions, averaging test and retest results.

Mean SRT (dB SNR)		Mean SRT (dB SPL)	
Ao Noise Closed	-4,8 (0,9)	Ao Quiet Closed	25,4 (3,2)
Ao Noise Open	-4,6 (1,1)	Ao Quiet Open	26,6 (3,8)
AV Noise Closed	-9,5 (3,2)	AV Quiet Closed	16,4 (5,6)
AV Noise Open	-8,9 (2,5)	AV Quiet Open	17,1 (5,1)

SRT80% expressed in dB SNR for the conditions in noise and in dB SPL for the conditions in quiet. For the conditions in quiet, we use the term SRT80% by extension, but the percentage expresses the sound presentation level for 80% word recognition.