



Research article

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

Loïc Le Rhun^{a,1}, Gerard Llorach^{b,1}, Tanguy Delmas^{a,c,1,*}, Clara Suied^d,
Luc H. Arnal^a, Diane S. Lazard^{a,e,f}

^a Institut Pasteur, Université Paris Cité, Inserm UA06, Institut de l'Audition, Paris, France

^b Auditory Signal Processing, Dept. of Medical Physics and Acoustics, University of Oldenburg Oldenburg, Germany

^c ECLEAR, Audition Lefevre – Audition Marc Boulet, Athis-Mons, France

^d Institut de Recherche Biomédicale des Armées, Département Neurosciences et Sciences Cognitives, Brétigny-sur-Orge, France

^e Princess Grace Hospital, ENT & Maxillo-facial Surgery Department, Monaco

^f Institut Arthur Vernes, ENT Surgery Department, Paris, France

ARTICLE INFO

Keywords:

Speechreading
Lipreading
Audio-visual interaction
French matrix sentence test
Dubbing

ABSTRACT

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, and open and closed-set response formats.

Results: Lipreading ability (Vo) ranged 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.

1. Introduction

During the COVID-19 pandemic, there was a growing recognition among the general population about the importance of lip-reading, even for those with typical hearing, particularly in situations where facial masks were worn [1]. Several studies have demonstrated that the combination of masks and background noise has a negative impact on speech intelligibility, even in normal listeners [2–4]. Yi et al. [4] showed a decrease of 20 % of sentence comprehension in normal listeners when adding a mask to the

* Corresponding author. Institut Pasteur, Université Paris Cité, Inserm UA06, Institut de l'Audition, Paris, France.

E-mail address: tanguy.delmas@pasteur.fr (T. Delmas).

¹ These authors contributed equally.

<https://doi.org/10.1016/j.heliyon.2024.e24750>

Received 1 March 2023; Received in revised form 7 December 2023; Accepted 12 January 2024

Available online 14 January 2024

2405-8440/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

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 from lips' movements, i.e. visemes, and facial expressions help disambiguate the auditory input [5–8]. Visemes are known to play a major role in the communication of the hearing impaired [9] and to have a predictive value in hearing rehabilitation by cochlear implants [10] or conventional hearing aids [11].

While many tests assessing speech understanding in Ao are available and validated in the international literature, calibrated automatized audio-visual (AV) testing material is scarce [12,13] and does not exist in French. Adding an AV assessment tool thus seems important to explore deafness and its compensation, and more generally AV synergy in deprived and non-deprived participants.

A wide variety of speech audiometry tests are available in French, from words in quiet (e.g., Fournier [14], Lafon [15]) to sentences in noise (e.g., MRT: Modified Rhyme Test [16], FrMST: French Matrix Sentence Test [17], VRB: Vocale Rapide dans le Bruit [18]). 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 [17], and is available in different languages (for comparison purposes). Its stereotypical pattern –sentences generated from combinations of 50 words (10 nouns + 10 verbs + 10 numbers + 10 objects + 10 colours)– is of interest when performing neurofunctional imaging, such as EEG [19]. In practice, the MST uses a staircase procedure presenting the sentences in a fixed background noise [20,21]. The signal presentation level varies to provide a specified speech reception threshold (SRT) (e.g., 50 %).

Four teams have already validated an audio-visual version of the MST into their respective languages (New Zealander English, Malay, Dutch, and German) [12,13,13,22–24]. Instead of recreating the audio-visual material, Llorach et al. [12] dubbed the original German MST audio sentences. Recording new MST speech material from scratch is a demanding process, which can induce differences in speech perception of up to 6 dB SNR compared to previous versions, depending on the talker [25]. Therefore, dubbing the original validated audio sentences with videos ensures results more comparable to those of the literature. To avoid AV lags, Llorach et al. [12] described the protocol for recording and selecting the best aligned videos from their audio recordings [12] (see the paragraph "Selection of the videos" for further details).

Here, we reproduced and improved the methodology (creation and validation) of Llorach et al. [12], and compared our results in terms of AV training effect, test-retest variability, lipreading ability in the visual only (Vo) condition, and AV gain in young normal hearing participants.

2. Material and method

All the device references used in this study are given in Table 1.

Scripts and guidelines for automatically recording, processing and cutting video material are available in the GitHub repository, at the following address: <https://github.com/gerardllorach/audiovisualdubbedMST/>.

2.1. Dubbing the auditory material and recording the videos

The French version was validated in 2012 [17] and is largely used in clinical practice. The initial corpus included 280 sentences. We manually discarded the sentences for which the original audio recording presented artifacts due to the merging of different sentences. From the remaining 150 sentences, forty-five lists of 20 sentences were created. In each list of 20 sentences, no word appeared more than twice and no sentence was repeated. 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 (Fig. 1). The set up followed the same principle as in Llorach et al. (2022). The speaker was filmed while repeating the sentence she was hearing through an earphone placed in her left ear. The camera recorded the video together with the two audio tracks (original sentence and microphone signal) separately. The camera recorded images at 50 fps/full HD and speech with a 48 kHz sampling rate and a 16-bit LPCM (linear pulse-code modulation) sampling format. Each sentence was presented 4 times in a row to the speaker. The speaker first heard three pure tones through the earphone to prepare herself and listened to the first presentation to know the five words to repeat. She then had to dub the three following sentences while simultaneously hearing them. The requirements were a neutral face, without exaggeration of the articulation, while keeping her eyes fixed and focused on the camera and avoiding blinking. Several takes could be recorded for the same sentence.

The resulting video files, which contained the visual speech, the original audio sentences of the French MST and the recorded

Table 1
References of the devices used.

Device	Reference	Brand
Recording studio set	ESDDI P8055 photo studio kit	ESDDI, Shenzhen, China
Camera	Sony a7S III	Sony Group Corporation, Tokyo, Japan
Condenser microphone	MKE 600 Shotgun Microphone	Senheiser, Germany
Mono-to-stereo cable	Hosa YMM261	Hosa Technology, INC, California, USA
Audio recorder	Zoom H6	Zoom North America, NY, USA
Audiometer	ELIOS-ECHODIA portable audiometer	Electronique du Mazet, France
Headphones	Sennheiser HDA 300	Sennheiser electronic GmbH & Co. KG, Germany
Screen	LG 27UN83A-W	LG Electronics, Korea
Sound card	SLL 2	Solid State Logic, UK

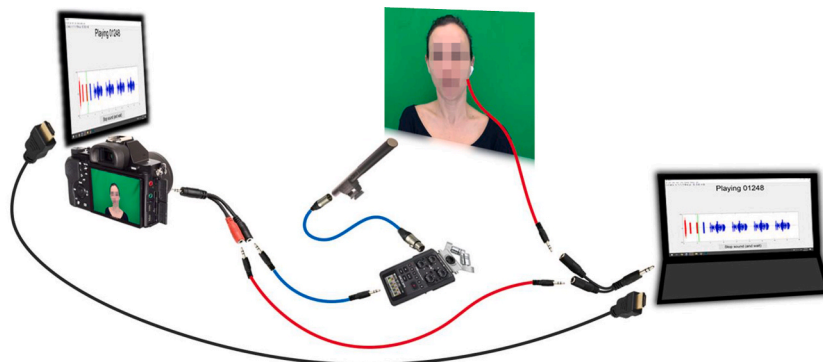


Fig. 1. Schematic drawing of the recording setup: the camera and the screen showing the visual aid are on the left side. The audio input of the camera is connected to the mono-to-stereo jack adapter. The input of this adapter is the recorded signal (blue path) and the original speech (red path). The 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.

speech were processed automatically via open-source scripts. During the recording session, the camera recorded continuously i.e., a video file may contain several takes of sentences. 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).

The MATLAB scripts used and guidelines to record the video material can be found in the GitHub repository (see Statements).

2.2. Selection of the videos

Videos in which the expressions were not neutral were manually removed. To select the best videos, i.e. those most synchronized with the original audio, we quantified the time misalignments between the original auditory sentences and the sentences dubbed by the speaker during the recording, using the same approach as in Llorach et al. [12], which relies on the dynamic time warping (DTW)

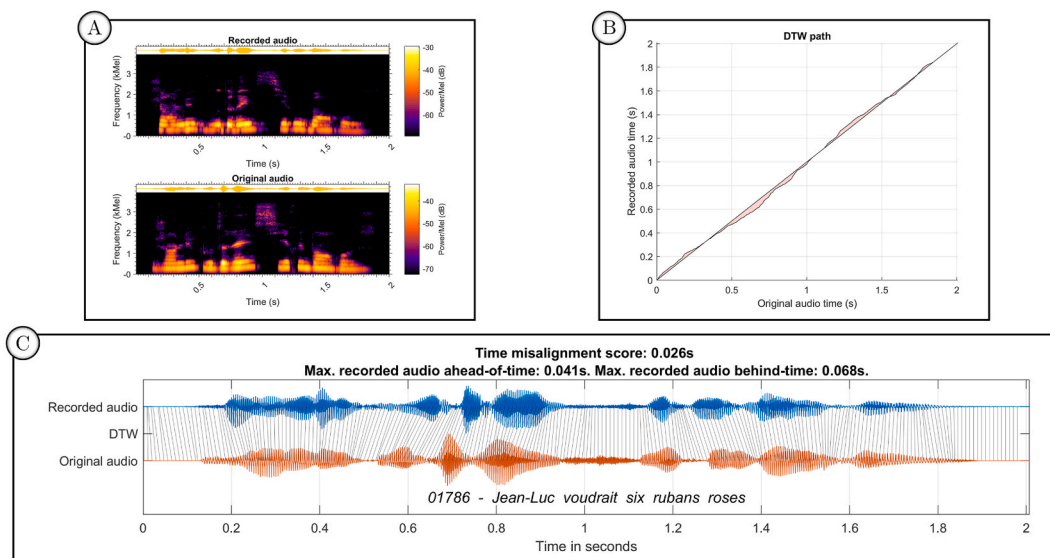


Fig. 2. Dynamic time wrapping (DTW) between an original audio and its most synchronous recording. Panel A shows the mel spectrograms of the two signals, which are used to compute the DTW. Panel B shows the DTW path between the mel spectrograms displayed in Panel A. The red area represents the time misalignment between the two spectrograms. It is used to compute the time misalignment score and maximum time misalignments (recorded audio ahead-of-time and behind-of-time) displayed in Panel C (top). Panel C shows the waveforms of the recorded audio (blue), the original audio (orange) and their temporal misalignment according to the frame-restricted DTW. The black lines in between the two signals, labelled as DTW, show the match between mel spectrogram frames. The sentence code and spoken words are shown below the signals (they are not temporally aligned with the spectrograms).

algorithm [26]. In technical terms, the DTW provides a warping path that can be interpreted as a time misalignment showing words spoken too early or too late and/or words spoken slower or faster when comparing two sentences. The time misalignment score allows selecting the best match between the recordings and the original sentences, enabling the selection of the most synchronized videos with the original audio (the new audio recordings are only used to select the most suitable videos). We computed the mel-spectrograms of the original and the dubbed sentences (example for one sentence on Fig. 2A) and applied the DTW algorithm (example for one sentence on Fig. 2B). We obtained the selection of the best aligned sentences. We secondarily constrained the algorithm to compare the mel spectrogram of one speech signal of the recorded sentences to a maximum of three frames of the matching original audio recording, in order to provide a more accurate time misalignment score. Fig. 2 reports the best final time misalignment score of one sentence, i.e. the average temporal difference between the reference and the best recorded sentence (go to <https://doi.org/10.5281/zenodo.8188917> for all sentences). Fig. 3 shows the distribution of the time misalignment scores of the original sentences against their final best selected sentences (left box plot: median 32.2 ms, mean 32.8 ± 9.5 ms, range: 16.4 to 59.3), against their four repetitions (middle box plot: mean 38.4 ± 14.6 ms) and against mismatched sentences (right box plot: mean 114.3 ± 96 ms). The ideal, but unrealistic, value would be a time misalignment score of 0 ms. The extreme case (no alignment) is displayed by the third box plot iteratively fitting mismatched sentences. The significant improvement in score between the three calculations ($p < 10^{-3}$) shows the relevance of the DTW method.

To summarize, the final audio-visual material consisted of the original validated auditory sentences of the FrMatrix and their best synchronized videos (selected on the smallest time misalignment scores between the two audio tracks). No participant reported any audio-visual lag during the test.

2.3. Validation of the AV material

2.3.1. Participants

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

The test was validated on a gender-balanced cohort of 35 normal-hearing participants. Two participants did not perform the retest session. Participants were 20–29 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). Their assessed mean pure tone average (PTA) was 1.22 dB HL (range: -4.25 ; 5).

2.3.2. Set up

The experiment was performed in an anechoic booth. Participants watched videos on a monitor at a viewing distance of ~ 80 cm

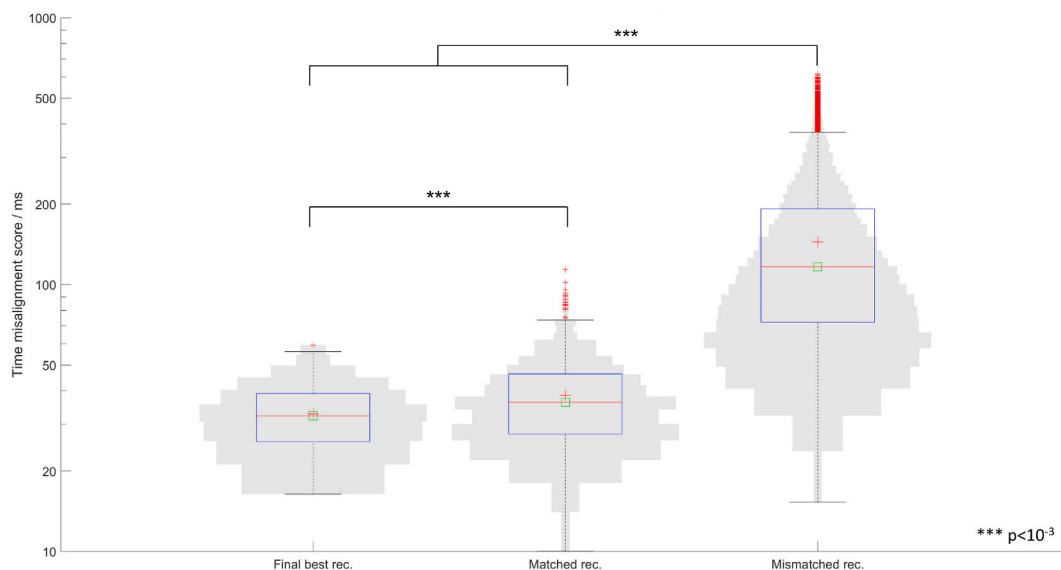


Fig. 3. Relevance and performance of the DTW method for selecting the best recorded audio tracks.

The left box plot displays the time misalignment scores between the original sentences and their best-matched recordings ($n = 150$ misalignment scores), the middle box plot displays the time misalignment scores between the original sentences and their 4 recordings ($n = 150 \times 4$ misalignment scores), the right box plot displays the time misalignment scores between the original sentences and the mismatched final best recordings ($n = 150 \times 149$ misalignment scores; the comparison between the sentence and its matching recording has been removed). The vertical axis is on a logarithmic scale. The large red cross represents the mean, the small red crosses represent outliers, and the red line represents the median. The lower whisker represents the 1st percentile, and the upper whisker represents the 99th percentile. The lower the score, the more aligned the audio signals

and listened to the speech sounds through binaural audiometric headphones. 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. The acoustic levels were calibrated using an artificial ear connected to a sound level meter.

2.3.3. Stimuli

Each video had a green background and started with a 500 ms still face. The participants were tested in quiet and in noise with the long-term stationary noise of the average speech spectrum (LTASS) from the original French MST [17], according to a usual adaptive procedure [20,21].

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

2.3.4. Experimental conditions

Three different modalities were tested, Audio only (Ao), Visual Only (Vo) and Audio-visual (AV). 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 noise to be more ecological 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. In the closed-set format, participants selected their responses on a matrix between all the 50 possible words and a "no response" option per word category.

For each condition, the participant was tested with one list of 20 sentences. The adaptive procedure was set to reach an individual Speech Reception Threshold of 80 % (SRT80) to avoid ceiling effects in the AV conditions. Indeed, some participants are able to understand more than 50 % of the content only relying on lipreading [13].

SRT80 is expressed in dB SPL in quiet conditions and in dB SNR (signal-to-noise ratio) in noisy conditions. For the Vo condition, we report the percentage of words correctly understood. No feedback was provided.

2.3.5. Procedure

A test-retest procedure (two sessions) was performed, spaced between 3 and 59 days apart.

All participants undertook a training phase before each session to control for the learning effect. This training phase was composed of four AV lists in noise during the first session, and one AV list during the retest session (Fig. 4). The response format (open-set or closed-set) of the training lists was randomly assigned to each participant. This response format was the same for both training phases. Participants started the session with the response format assigned during the training, and secondarily performed the same conditions with the other response format.

The order of the modalities (AV, Ao, Vo) and the environment (quiet and noise) was randomised for the two sessions. All

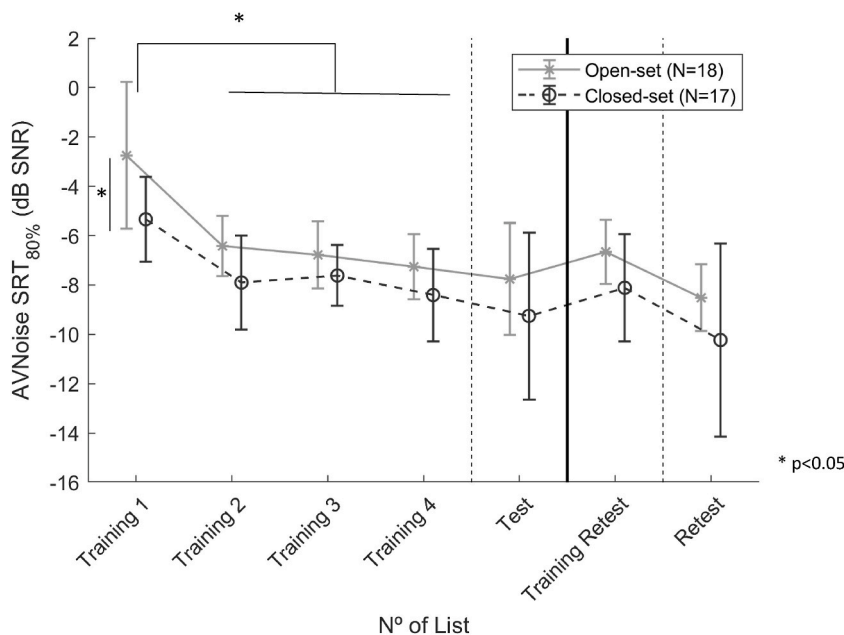


Fig. 4. Audio-visual training effect in noise during the test and retest sessions.

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

Results are expressed in mean SRT \pm 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.

participants performed the conditions in quiet and in noise.

We added another condition to the second session to compare our results with those described in Jansen et al. [17], i.e. SRT50 in Ao in noise and closed-set format. This measure was performed after the training list.

2.4. Statistical analysis

Results are expressed in means \pm 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 (Pearson) between the Vo scores (lipreading skills) and the AV benefit (AV – Ao scores).

3. Results

3.1. SRT50

In 2012, Jansen et al. [17] validated the French version of the MST with a mean SRT50 of -6 dB SNR (± 0.6). Here, we obtained a mean SRT50 of -7.75 dB SNR (± 0.79).

3.2. Evaluation of the training effect

The clinical use of the MST starts with a systematic training session based on two training lists to get used to the type of material and background noise [17]. For the AV version, we 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). The test and retest sessions were spaced 17 days apart on average (range: 3–59).

Fig. 4 shows the training effect (mean SRT80 % in dB SNR for the AV condition in noise in the response format performed) during the test and retest sessions. On average, the training effect between the first training list and the test was -4.5 dB SNR for the first session, and -1.93 dB SNR for the second session, regardless of the response format. 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 this effect came from the difference between the first list of the first session and the three other training lists ($p < 10^{-3}$, Fig. 4). Subsequent SRTs of the remaining training lists did not significantly differ. The analysis further showed a significant improvement during the second session between the training and the test ($F(1, 31) = 25$, $p < 10^{-3}$). 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 only ($F(1, 33) = 11.6$, $p = 0.002$). The interaction with the number of training lists was not significant ($F(2, 65.90) = 2.58$, $p = 0.084$).

3.3. Audio-only (ao) and audio-visual evaluation

Table 2 shows the mean SRTs (\pm 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.

The AV conditions allowed some participants to reach the SRT80 with speech presentation levels below 0 dB SPL in quiet (Fig. 5A) and SNRs below -20 dB SNR in noise (Fig. 5B). These auditory presentation levels are near the threshold of human hearing and speech detection [27]. For example, the speech detection threshold (SDT) of the female German MST has been evaluated at -16.9 dB SNR in the Ao condition in noise [28], a threshold that can be theoretically lowered by -3 dB when adding visual cues, as demonstrated by Bernstein et al. [29]. We believe that these participants reached these presentation levels because they were able to lipread more than 80% of the material, forcing the adaptive procedure to lower the SNRs or SPLs. As Llorach et al. [12] did, we considered that these presentations levels were not fully representative of audio-visual speech reception and limited their SRT scores to -20 dB SNR and 0 dB

Table 2

Mean Speech Reception Thresholds SRT80 % (\pm 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.

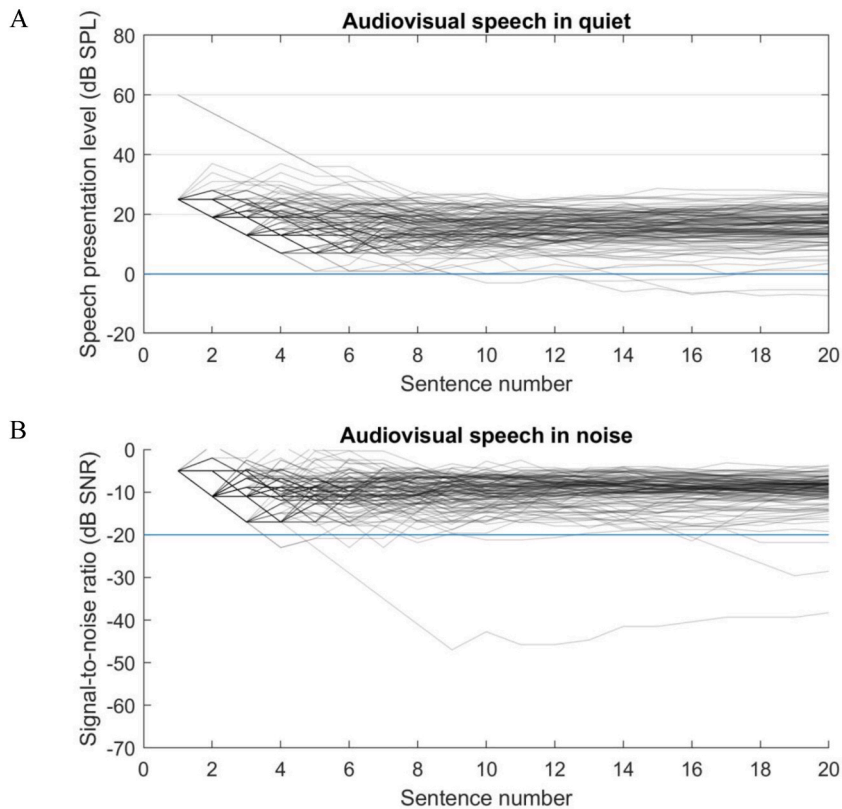


Fig. 5. Adaptive speech presentation levels for the Audio-visual conditions in quiet (a) and adaptive signal to noise ratios for the Audio-visual conditions in noise (b).

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 participant of the study who started with a presentation level of 60 dB.

SPL. This score limitation concerned 3 scores from 3 participants out of 445 SRTs (0.7 % versus 5 % in Llorach et al. [12]). It is to note that Ross et al. [30] and Sumbly and Pollack [31] found audio-visual speech reception thresholds below -20 dB SNR. However, the material used was different (noise and target sentences) making the comparison more difficult than with the German MST version that used similar testing conditions. Future work should consider measuring the real SDT of the French MST.

3.4. Lipreading (vo) and audio-visual benefit

Participants had a wide range of lipreading abilities (Fig. 6). 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. [12] 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 AV benefit ($p < 10^{-3}$) in all AV conditions (Fig. 7A in noise and 7B in quiet). The Pearson *r* ranged from -0.55 (in quiet and open-set format) to -0.74 (in noise and open-set format).

3.5. Test-retest reliability

To assess the test-retest reliability of the audio-visual FrMST, we compared the within-participant standard deviations (test minus retest) to the between-participant standard deviations. In a homogenous group of young normal-hearing, the within- and between-participant standard deviations are expected to be similar in the Ao condition of the MST [32]. In the AV and Vo conditions, a larger between-participant variability is expected [12]. To evaluate the ability of the test to discriminate participants, we used the ‘ 2σ criterion’ [32]. According to this criterion, the between-participant standard deviations should be double or more than the within-participant standard deviations to be able to discriminate between participants [32].

Fig. 8 shows the within- and between-participant standard deviations of the scores in the conditions tested (SRTs and percentage of correct words for the Vo condition). The black line represents the 2σ criterion, which is the double of the within-participant standard

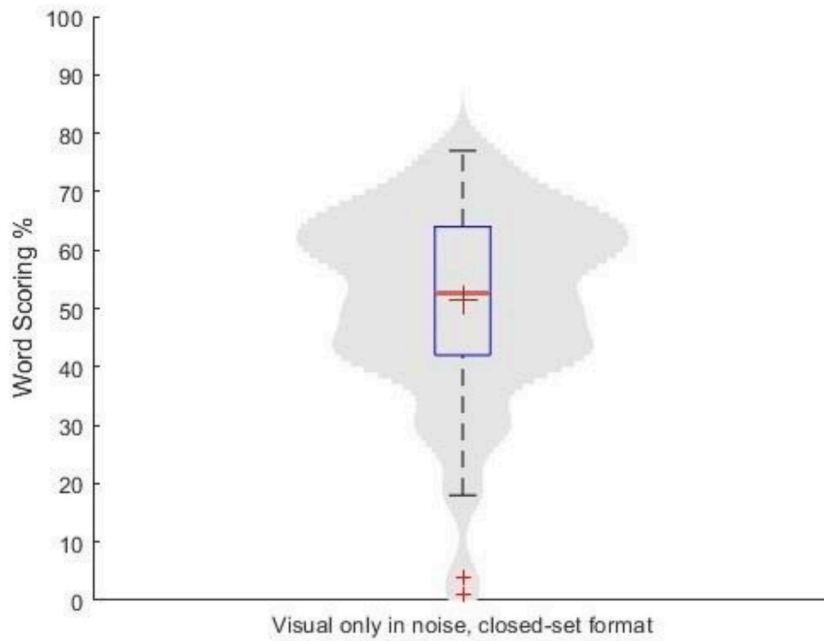


Fig. 6. Boxplot and distribution of the lipreading scores: Visual only in noise and closed-set format. The mean and the median are represented by a red cross and a red line, respectively. The outliers are represented by smaller red crosses. The boxplot is smoothed with the Matlab Kdensity function that returns a probability density estimate.

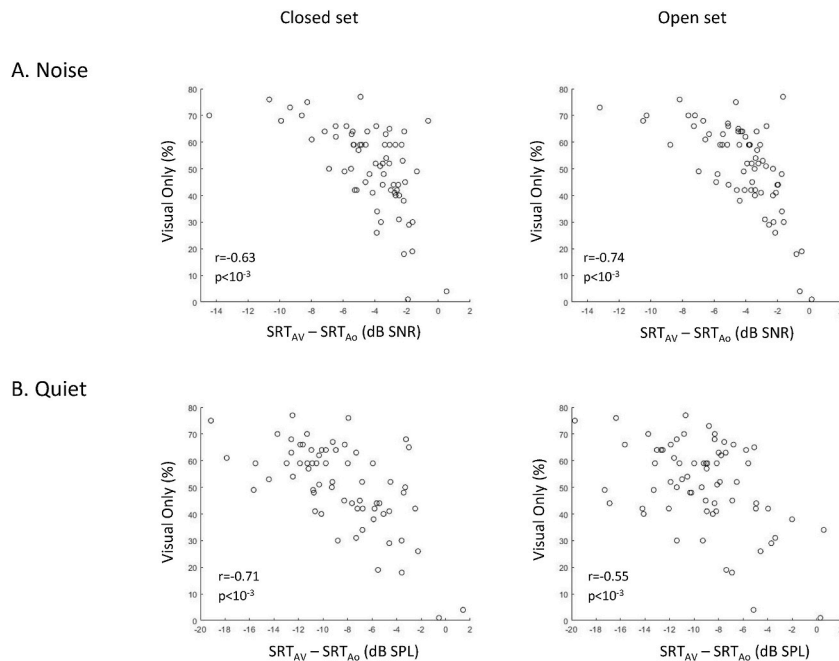


Fig. 7. Visual Only scores as a function of the audio-visual benefit A. in noise (in dB SNR) and B. quiet (in dB SPL), in closed-set (left column) and open-set (right column) formats (individual data points). The scores of the test and retest sessions are dissociated, represented as two open circles per participant. r and p stand for the results of the Pearson correlations.

deviation. Except for the Vo condition, none of the conditions exceeded the 2σ criterion, as found in Llorach et al. [12], who performed the same analysis. These results mean that it is possible to differentiate skilled from unskilled young normal-hearing participants in the Vo condition but not in the other conditions when using a 20-sentence list.

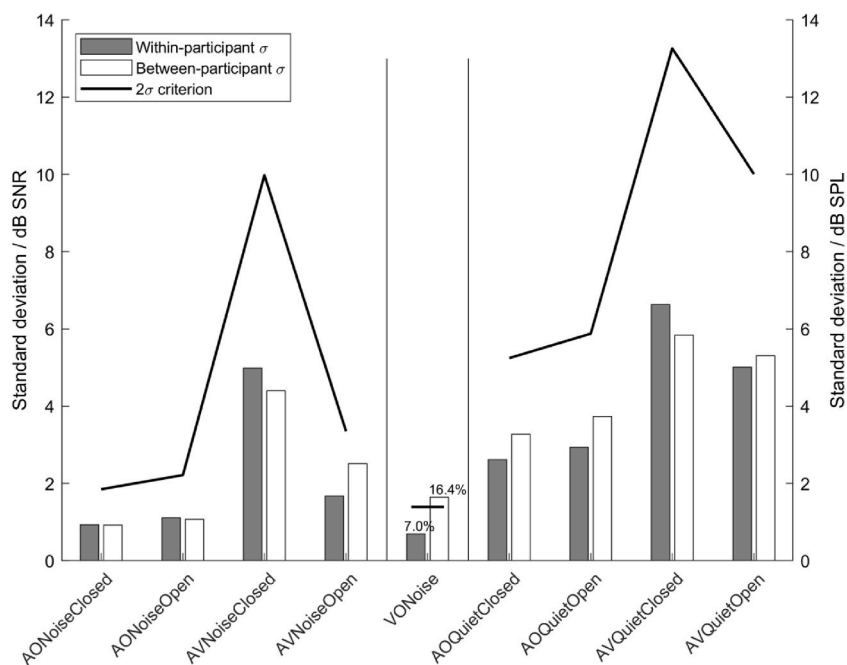


Fig. 8. Within-participant (grey bars) and between-participant (white bars) standard deviations (SD) for all conditions. The 2σ criterion is indicated as a thick black line. On the left, SDs of speech in noise conditions expressed in dB SNR; middle, SDs of the visual-only, expressed in percentage; right, SDs of speech in quiet, expressed in dB SPL.

4. Discussion

The present study designed and tested a standardized audio-visual material to evaluate lipreading skills and audio-visual benefit in a young normal-hearing sample. We used the French audio MST corpus validated by Jansen et al. [17], and the methodology described in Llorach et al. [12]. We compare our results with these two studies.

4.1. 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 participants than Llorach et al. [12] and 5 more participants than Jansen et al. [17]. They all had normal hearing.

4.2. SRT50

A difference of -1.75 dB SNR was observed between the current study (-7.75 ± 0.79 dB SNR) and Jansen et al. [17] (-6 ± 0.6 dB SNR) when measuring SRT50 in audio-only and closed-set format. This outperformance may be explained by differences in calibration and technical configuration (binaural *versus* monaural presentation), and/or timing of the SRT50 assessment during the second session (greater training).

4.3. Validation of the visual material

The absolute AV benefit (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 and 7.0 dB SPL in Llorach et al. [12], and 3 dB SNR in Van de Rijt et al. [13]. SRT differences in the AV and Ao conditions may show that the test was more difficult in French than in German. Similar differences between these languages have been mentioned before for SRT50 [33].

Although the AV benefit was in accordance with the literature and no participant reported AV lags, it is worth noting that our audio-visual material has inherent misalignments due to the recording method. The median and mean misalignment scores were however acceptable, around 32 ms. These pitfalls should be taken into account if audio-visual synchrony is paramount. Future work may use the warping path from the DTW to secondarily time-align the video recordings with the original sentences if more synchrony is required.

During the Visual-only condition, some participants scored up to 77 % of comprehension. The average was 51.3 % (± 16.2 %), similar to that of Llorach et al. [12] (50 % ± 21.4 %, range: 0%–84 %).

4.4. Training effect

An improvement of 1.8 dB SNR between the first and the sixth list is expected in the audio-only French MST, i.e., for SRT50 [17]. Here, we found a 4.5 dB SNR improvement for the SRT80 in the AV condition in noise, between the first training list and the test list. This greater effect probably arose from the participants learning to lipread the material. Indeed, Lander & Davies [34] demonstrated that lipreading performance increased overall with practice. 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. [12].

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. [12].

4.5. Test-retest reliability

We found a small within- and between-participants 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, in accordance with the literature [12,35,36].

The Visual-only scores were highly variable across participants, as expected [12,13,24]. According to the '2 σ criterion', the Vo test is a reliable choice to determine if a participant is proficient in lipreading. In other words, it is enough to test a 20-sentence list in the Vo condition to determine the lipreading capabilities of an individual (after training with two AV lists).

Unfortunately, this was not the case for the AV conditions. The '2 σ criterion' was not fulfilled for any of the AV conditions, meaning that one 20-sentence list was not enough to differentiate AV speech perception between young NH participants and that several lists might be required to obtain a more precise score. These results are in accordance with Llorach et al. [12], where inherent asynchronies in the AV conditions were also present. These asynchronies could explain the increase in the variability in the AV measurements. Alternatively, it is plausible that audio-visual integration operates as an independent process [37] exhibiting variability that is unrelated to the variability observed in the Ao and Vo conditions. To shed light on these hypotheses, future research should investigate the within-participant variability of audio-visual speech perception.

A larger between and within participants' variability was found in the AV conditions in the present study as compared to Llorach et al. [12]. This variability can be explained by the different individual lipreading abilities between the samples.

4.6. Response format

There was no difference between the response formats once participants had been trained, similarly to the audio-visual German MST [12]. For most languages, a closed-set response format means smaller SRTs (i.e., easier task), but not for German and Polish, as reported in Kollmeier et al. [33].

5. Conclusion

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

The SRT80 values for young normal hearing participants in the AV condition were -9.2 dB SNR in noise and 16.8 dB SPL in quiet. The absolute AV 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 %. These results are in accordance with those already published for AV MSTs in other languages.

One should take into account that the audio-visual test suffers from ceiling effects as some participants were able to recognize more than 50 % of the words in the Visual-only situation. Therefore, targeting SRT80 is recommended. Due to the within-participant variability of the AV test more than one list might be required to obtain a precise auditory-visual score when comparing young normal-hearing participants.

This audio-visual material can be used for clinical purposes, but in practice, it is not possible to carry out all the conditions described, especially in hearing-impaired participants. For clinical purposes, we suggest performing two AV training lists in noise (closed-set or open-set format, no significant difference in French), followed 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, in a randomised order.

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

Statements

The video recordings of the female French Matrix Sentence Test can be found free of charge online (<https://doi.org/10.5281/zenodo.8188917>). Please cite the present article and the following when using the material.

Funding statement

This work was funded by Fondation Pour l'Audition FPA RD-2020-10 (for LLR, TD, LHA, DL), Fondation d'entreprise Optic 2000 – Lissac – Audio 2000 (donation to DL and LHA) and by the Deutsche Forschungsgemeinschaft (for GL, DFG, Cluster of Excellence EXC

1077/1 “Hearing4all”, and SFB1330 Projects B1 and C4). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability statement

Data will be made available on request.

Additional information

No additional information is available for this paper.

CRediT authorship contribution statement

Loïc Le Rhun: Writing – original draft, Methodology, Investigation, Formal analysis. **Gerard Llorach:** Writing – review & editing, Software, Methodology, Formal analysis, Conceptualization. **Tanguy Delmas:** Writing – review & editing, Methodology, Investigation, Formal analysis. **Clara Suied:** Software, Resources. **Luc H. Arnal:** Supervision, Project administration, Funding acquisition, Conceptualization. **Diane S. Lazard:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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), Fondation d’entreprise Optic 2000 – Lissac – Audio 2000 (donation to DL and LHA) and by the Deutsche Forschungsgemeinschaft (for GL, DFG, Cluster of Excellence EXC 1077/1 “Hearing4all”, and SFB1330 Projects B1 and C4). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank one of the reviewers for relevant discussions about the DTW calculations, which helped improve our work.

References

- [1] P. Bottalico, S. Murgia, G.E. Puglisi, A. Astolfi, K.I. Kirk, Effect of masks on speech intelligibility in auralized classrooms, *J. Acoust. Soc. Am.* 148 (5) (2020), <https://doi.org/10.1121/10.0002450>. Article 5.
- [2] R. Sönnichsen, G. Llorach Tó, S. Hochmuth, V. Hohmann, A. Radeloff, How face masks interfere with speech understanding of normal-hearing individuals : vision makes the difference, *Otol. Neurotol.*: Official Publication of the American Otological Society, American Neurotology Society [and] European Academy of Otolaryngology and Neurotology 43 (3) (2022), <https://doi.org/10.1097/MAO.0000000000003458>. Article 3.
- [3] M. Tofanelli, V. Capriotti, A. Gatto, P. Boscolo-Rizzo, S. Rizzo, G. Tirelli, COVID-19 and deafness : impact of face masks on speech perception, *J. Am. Acad. Audiol.* 33 (2) (2022) 98–104, <https://doi.org/10.1055/s-0041-1736577>.
- [4] H. Yi, A. Pingsterhaus, W. Song, Effects of wearing face masks while using different speaking styles in noise on speech intelligibility during the COVID-19 pandemic, *Front. Psychol.* 12 (2021). <https://www.frontiersin.org/article/10.3389/fpsyg.2021.682677>.
- [5] L.H. Arnal, B. Morillon, C.A. Kell, A.-L. Giraud, Dual neural routing of visual facilitation in speech processing, *J. Neurosci.* 29 (43) (2009). Article 43.
- [6] M. Bourguignon, M. Baart, E.C. Kapnola, N. Molinaro, Lip-reading enables the brain to synthesize auditory features of unknown silent speech, *J. Neurosci.* 40 (5) (2020) 1053–1065, <https://doi.org/10.1523/JNEUROSCI.1101-19.2019>.
- [7] N.P. Erber, Auditory-visual perception of speech, *J. Speech Hear. Disord.* 40 (4) (1975) 481–492, <https://doi.org/10.1044/jshd.4004.481>.
- [8] L. Kessous, G. Castellano, G. Caridakis, Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis, *Journal on Multimodal User Interfaces* 3 (1–2) (2010) 33–48, <https://doi.org/10.1007/s12193-009-0025-5>.
- [9] H. Pimperton, A. Ralph-Lewis, M. MacSweeney, Speechreading in deaf adults with cochlear implants : evidence for perceptual compensation, *Front. Psychol.* 8 (2017). <https://www.frontiersin.org/article/10.3389/fpsyg.2017.00106>.
- [10] D.S. Lazard, A.-L. Giraud, Faster phonological processing and right occipito-temporal coupling in deaf adults signal poor cochlear implant outcome, *Nat. Commun.* 8 (1) (2017), <https://doi.org/10.1038/ncomms14872>. Article 1.
- [11] A. Dell’Arling, E. Adachi, A. Dell’Arling, Lip reading role in the hearing aid fitting process, *Brazilian journal of otorhinolaryngology* 73 (2007) 95–99, [https://doi.org/10.1016/S1808-8694\(15\)31129-0](https://doi.org/10.1016/S1808-8694(15)31129-0).
- [12] G. Llorach, F. Kirschner, G. Grimm, M.A. Zokoll, K.C. Wagener, V. Hohmann, Development and evaluation of video recordings for the OLSA matrix sentence test, *Int. J. Audiol.* 61 (4) (2022) 311–321, <https://doi.org/10.1080/14992027.2021.1930205>.
- [13] L.P.H. van de Rijt, A. Roye, E.A.M. Mylanus, A.J. van Opstal, M.M. van Wanrooij, The principle of inverse effectiveness in audiovisual speech perception, *Front. Hum. Neurosci.* 13 (2019) 335, <https://doi.org/10.3389/fnhum.2019.00335>.
- [14] J.-E. Fournier, A.P. Aubin, *Audiométrie vocale : Les épreuves d’intelligibilité et leurs applications au diagnostic, à l’expertise et à la correction prothétique des surdités*, Maloine, 1951.
- [15] J.C. Lafon, *Phonetic test, phonation, audition*, *JFORL. Journal français d’oto-rhino-laryngologie; audiophonologie et chirurgie maxillo-faciale* 21 (3) (1972) 223–229.
- [16] V. Zimpfer, G. Andéol, G. Blanck, C. Suied, T. Fux, Development of a French version of the modified Rhyme test, *J. Acoust. Soc. Am.* 147 (1) (2020), <https://doi.org/10.1121/10.0000559>. EL55-EL61.
- [17] S. Jansen, H. Luts, K.C. Wagener, B. Kollmeier, M. Del Rio, R. Dauman, C. James, B. Fraysse, E. Vormès, B. Frachet, J. Wouters, A. van Wieringen, Comparison of three types of French speech-in-noise tests : a multi-center study, *Int. J. Audiol.* 51 (3) (2012), <https://doi.org/10.3109/14992027.2011.633568>. Article 3.

- [18] F. Leclercq, C. Renard, C. Vincent, Speech audiometry in noise : development of the French-language VRB (vocale rapide dans le bruit) test, *European Annals of Otorhinolaryngology, Head and Neck Diseases* 135 (5) (2018), <https://doi.org/10.1016/j.anorl.2018.07.002>. Article 5.
- [19] S.J. Luck, *An Introduction to the Event-Related Potential Technique*, second ed., MIT Press, 2014.
- [20] T. Brand, S. Kissner, T. Jürgens, D. Berg, B. Kollmeier, mars 5). Adaptive Algorithmen zur Bestimmung der 80%-Sprachverständlichkeitsschwelle, 2011.
- [21] T. Brand, B. Kollmeier, Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests, *J. Acoust. Soc. Am.* 111 (6) (2002), <https://doi.org/10.1121/1.1479152>. Article 6.
- [22] G.A. O'Beirne, R.H. Trounson, A.D. McClelland, S.A. Jamaluddin, M.A. Maclagan, Development of an auditory-visual matrix sentence test in New Zealand English, *Journal of International Advanced Otolology* 11 (Supplement 1) (2015). Article Supplement 1.
- [23] R.H. Trounson, Development of the UC Auditory-Visual Matrix Sentence Test, 2012.
- [24] S.A. Jamaluddin, Development and Evaluation of the Digit Triplet and Auditory-Visual Matrix Sentence Tests in Malay, 2016.
- [25] S. Hochmuth, B. Kollmeier, T. Brand, T. Jürgens, Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests, *Int. J. Audiol.* 54 (sup2) (2015), <https://doi.org/10.3109/14992027.2015.1046502>. Article sup.2.
- [26] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, *IEEE Trans. Acoust. Speech Signal Process.* 26 (1) (1978), <https://doi.org/10.1109/TASSP.1978.1163055>. Article 1.
- [27] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, BRILL, 2012.
- [28] W. Schubotz, T. Brand, B. Kollmeier, S.D. Ewert, Monaural speech intelligibility and detection in maskers with varying amounts of spectro-temporal speech features, *J. Acoust. Soc. Am.* 140 (1) (2016), <https://doi.org/10.1121/1.4955079>. Article 1.
- [29] L. Bernstein, E. Auer, S. Takayanagi, Auditory speech detection in noise enhanced by lipreading, *Speech Commun.* 44 (2004) 5–18, <https://doi.org/10.1016/j.specom.2004.10.011>.
- [30] L.A. Ross, D. Saint-Amour, V.M. Leavitt, D.C. Javitt, J.J. Foxe, Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments, *Cerebr. Cortex* 17 (5) (2007) 1147–1153, <https://doi.org/10.1093/cercor/bhl024>.
- [31] W.H. Sumbly, I. Pollack, Visual contribution to speech intelligibility in noise, *J. Acoust. Soc. Am.* 26 (2) (2005) 212–215, <https://doi.org/10.1121/1.1907309>.
- [32] K.C. Wagener, T. Brand, Sentence intelligibility in noise for listeners with normal hearing and hearing impairment : influence of measurement procedure and masking parameters, *Int. J. Audiol.* 44 (3) (2005), <https://doi.org/10.1080/14992020500057517>. Article 3.
- [33] B. Kollmeier, A. Warzybok, S. Hochmuth, M.A. Zokoll, V. Uslar, T. Brand, K.C. Wagener, The multilingual matrix test : principles, applications, and comparison across languages: a review, *Int. J. Audiol.* 54 (Suppl 2) (2015) 3–16, <https://doi.org/10.3109/14992027.2015.1020971>.
- [34] K. Lander, R. Davies, Does face familiarity influence speechreadability? *Q. J. Exp. Psychol.* 61 (7) (2008) <https://doi.org/10.1080/17470210801908476>. Article 7.
- [35] G.F. Smoorenburg, Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram, *J. Acoust. Soc. Am.* 91 (1) (1992) 421–437, <https://doi.org/10.1121/1.402729>.
- [36] P.E. Souza, K.T. Boike, K. Witherell, K. Tremblay, Prediction of speech recognition from audibility in older listeners with hearing loss : effects of age, amplification, and background noise, *J. Am. Acad. Audiol.* 18 (1) (2007) 054-065, <https://doi.org/10.3766/jaaa.18.1.5>.
- [37] K.W. Grant, Measures of auditory-visual integration for speech understanding : a theoretical perspective, *J. Acoust. Soc. Am.* 112 (1) (2002) 30–33, <https://doi.org/10.1121/1.1482076>.