scientific reports

OPEN

# Pupillometry reveals effects of pitch manipulation within and across words on listening effort and short-term memory

Yue Zhang[1]✉, Anastasia Sares[2], Arthur Delage[3], Alexandre Lehmann[1] & Mickael Deroche[3]

**For individuals with hearing loss, even successful speech communication comes at a cost. Cochlear implants transmit degraded information, specifically for voice pitch, which demands extra and sustained listening effort. The current study hypothesized that abnormal pitch patterns contribute to the additional listening effort, even in non-tonal language native speaking normally hearing listeners. We manipulated the fundamental frequency (F0) within and across words, while participants listen and repeat (simple intelligibility task), or listen, repeat, and later recall (concurrent encoding task) the words. In both experiments, the F0 manipulations resulted in small changes in intelligibility but no difference in free recall or subjective effort ratings. Pupillary metrics were yet sensitive to these manipulations: pupil dilations were larger when words were monotonized (flat contour) or inverted (the natural contour flipped upside-down), and larger when successive words were organized into a melodic pattern. The most likely interpretation is that the natural or expected F0 contour of a word contributes to its identity and facilitate its matching and retrieval from the phonological representation stored in long-term memory. Consequently, degrading words' F0 contour can result in extra listening effort. Our results call for solutions to improve pitch saliency and naturalness in future development of cochlear implants' signal processing strategies, even for non-tonal languages.**

Hearing loss is recognized by the WHO as one of the top 5 leading causes of disability, and its effects will only grow as we live in an increasingly noisy world[1]. On top of the practical challenges of living with reduced hearing, hearing loss is often related to negative cognitive and psychosocial impact. For instance, hearing loss has been identified as the leading modifiable risk factor for dementia[2,3]. Though many challenges remain in addressing this issue, medical interventions do exist, such as hearing aids and, in cases of severe to profound sensorineural hearing loss, cochlear implants (CIs).

CI recipients experience good speech-in-quiet performance. However, individual differences in outcomes are substantial, both in speech performance and in life quality[4–7]. Specifically, most CIs on the market follow envelope-based coding strategies, and eliminate the fine structure necessary to extract a precise fundamental frequency (F0)[8–10]. The result is that, despite generally accurate timing cues, the quality of pitch encoding is severely degraded compared to that of normally hearing (NH) listeners. This significantly limits CI users' abilities in domains that rely on a fine sense of pitch, such as music, speech prosody and emotion, speaker identity, and segregating competing speakers in a conversation[11–15]. This problem is not unique to CI users: hearing impaired (HI) listeners have decreased temporal fine structure processing and broadened peripheral filters[16,17]. For young children, a fine representation of voice pitch information has shown to be primordial, as the exaggerated prosody of infant-directed speech is thought to play a key role in language acquisition[18–21]. Therefore, pitch is an important perceptual feature that is perturbed in multiple ways depending on the type of hearing loss, with deleterious effects on early language learning as well as cognitive aging.

For both individuals with hearing loss and NH listeners in non-optimal situations (for instance low-quality phone calls), degraded pitch information may not only affect hearing performance, but could also contribute to elevated effort and fatigue, even when speech recognition scores are seemingly unaffected[22–24]. The possible impact of degraded pitch information on listening effort can be understood using the ease of language understanding (ELU) model: when the speech signals are clean, perceived signals can be matched to long-term phonological representations automatically, without engaging explicit cognitive resources. However, when speech is distorted, for instance when the acoustic signal is degraded by masking noise or spectral smearing, a

[1]McGill University, Montreal, Canada. [2]Colorado State University, Fort Collins, CO, USA. [3]Concordia University, Montréal, Canada. ✉email: yue.zhang7@mail.mcgill.ca

mismatch will occur between the perceived signal and the template stored in the long-term memory. Resolving this mismatch requires extra processing and more cognitive resources, leading to more effortful listening[25–27]. Subsequently, this additional demand might leave fewer resources for other secondary tasks, such as recall and information synthesis. However, to date, this potential relation between degraded pitch information and listening effort has not been investigated. The paucity of evidence on this relation leads to an alternative hypothesis, namely that it is largely irrelevant for intelligibility in quiet. Even though pitch is important for suprasegmental informational transfer such as emotion, intonation, and prosody, pitch has not been assigned the same crucial importance to intelligibility compared to other phonemic components (i.e., vowels and consonants) specifically in non-tonal languages. For instance, manipulating average F0 and F0 contour between successive speech segments did not impair top-down phonemic restoration[28,29]. Flattening F0 contour had no significant impact on speech reception thresholds (SRTs) in speech-shaped noise[57]. Therefore, are abnormal F0 inflections generally irrelevant (in quiet or stationary noise) because listeners engage too easily in speech decoding mechanisms mostly concerned with articulation, or is pitch distortion genuinely costly to the matching of words with stored templates? Are these effects too subtle for intelligibility purposes and need more complex tasks or more sensitive neurophysiological measures to be revealed? Past studies have shown cases where perfect intelligibility can still come at additional cognitive processing cost[79,80]. Therefore, it is possible that distorted pitch information does not impair intelligibility (at least in a speech task that does not have other competitors of cognitive resources), but it will impact more sensitive measures that tap into the cognitive processing of speech, for instance pupillary change. If F0 degradation induced by current CI processing is not as detrimental to speech recognition and listening effort as other cues (for instance low spectral resolution[30]), then the research and development should focus on the most detrimental cues, and leave the pitch problem aside. Therefore, the current study aims to investigate whether distorted voice pitch information affected listening effort and aims to quantify its effect size to be later compared with other acoustic cues.

Answering this question requires a suitable experimental methodology. The behavioral cost of elevated listening effort has been demonstrated and quantified in past studies using dual-task paradigms. For instance, studies using the sentence-final word identification and recall (SWIR) test have shown that recall can be impaired in difficult listening conditions, even though intelligibility scores remained unchanged[31–33]. Therefore, it is likely that using a similar dual-task paradigm, degraded pitch information would increase the listening effort during speech decoding (i.e., primary task), leaving fewer resources for memorizing the items (i.e., secondary task). In addition, there is some evidence that certain pitch patterns in spoken material (across words) can improve immediate free recall for NH listeners[34–38]. It is unclear whether such phenomena would hold with abnormal pitch patterns (within or across words) but this represents an additional route by which pitch degradations could impair the secondary task in a dual-task paradigm.

Elevated listening effort has a physiological 'footprint.' Pupillometry has been shown to be a valid correlate of listening effort in different listening conditions[24,39–43]. Event-evoked pupillary responses typically increase with increased task demands, for instance low signal-to-noise ratio (SNR), reverberation, spectral degradation, etc., until an 'inversion point' where the participants 'give up'[44,45]. Therefore, event-evoked responses like peak pupil dilation (PPD), mean pupil dilation and peak latency are interpreted as indications of effort. Recently, more studies have started to investigate pupil responses during concurrent tasks, in an effort to understand pupil dynamics in ecological conditions[46–49], because rarely in real life are people challenged by one task at a time. For instance, Zhang et al., (2021)[49] found that the task complexity (listening, repetition, or learning) and the acoustic adversity (noise level) affected different elements of the pupil traces. While more difficult background noise was associated with higher PPD, accumulating memory load was associated with increasing baseline pupil response (plateauing in the second half of a list). Therefore, pupillometry is sensitive to different types of cognitive load and suitable for the purpose of the current study. Here, we replaced SNR (noise) manipulations with F0 manipulations, at both within-word or across-word levels to address the hypothesized impact of pitch distortion on listening effort and memory performance in a NH sample.

The first experiment focused on within-word pitch information by manipulating the F0 contour of each word presented. We hypothesized that *monotonized* words would be less accurately recalled, and lead to a greater pupil dilation, than *naturally intonated* words, because participants would need to engage explicit cognitive resources to resolve the mismatch between the somewhat artificial (robotic-like) voice quality of the words and their stored phonological representations. We also included words with an *inverted* F0 contour, which, despite having the same pitch range and formant changes as the naturally intonated words, violate the stored template for the words. Both monotonized and inverted versions should pose a conflict with the stored representation of these words, just like speech degraded by noise or spoken in accents[39,40,50]. On the contrary, exaggerated pitch contours may be processed with more ease than normally-intonated words, leading to better memory and smaller pupil dilation. While more speculative, we made this prediction based on the exaggerated intonation present in 'motherese speech', i.e. speech directed to infants and toddlers[51] as well as caricatures[52,53] where exaggerated vocal traits may help (not hinder) recognition despite the acoustic distortions they produce.

The second experiment focused on across-word ("suprasegmental") pitch sequences by manipulating mean F0 of successive words across the list of ten. Apart from contributing to the internal representation of speech sounds, pitch is also a strong across-time grouping cue that fuses different speech tokens with varied spectrum patterns together. Disrupting this grouping effect might impact both speech recognition and memory. We compared a fixed condition (identical to the monotonized condition of experiment 1) with a melodic condition, in which monotonized words were presented at different mean F0s ("notes"), making an arpeggio-like sequence of pitches. It was not known how this pitch pattern would affect recall, if at all: there are various studies showing pitch patterns to be either helpful, irrelevant, or even detrimental to short-term memory of words[34,35,37,38,54–56]. The pace at which these pitch sequences have been played (in these past studies) is likely a critical factor contributing to some of this discrepancy, but here it was constrained by pupillometry (i.e., waiting enough time for the pupil to return to baseline before the next word).

These manipulations will help us understand whether access to salient pitch information supports word decoding and storage, and whether it requires fewer explicit cognitive resources (i.e., lowering listening effort). Results will guide us in the future to improve the front-end speech processing strategies to improve acoustic information transfer for assistive hearing devices such as hearing aids and CIs.

## Results

### Experiment 1: within-word pitch manipulations

*Behavioral data: intelligibility*

The results are summarized in Table 1; Fig. 1 (top left). As hypothesized, there was a main effect of pitch manipulation on word recognition [$\chi^2(3) = 13.9$, $p = 0.003$], caused primarily by the exaggerated pitch condition leading to better intelligibility than the monotone and inverted conditions ($p < 0.010$). Bear in mind, however, that word recognition in quiet was close to ceiling, so this difference was negligible, with the difference amounting to about 2% (intelligibility was > 95% overall). The task (i.e. whether listeners simply repeated words or repeated while memorizing them) had no effect, but there was a main effect of position on word recognition [$\chi^2(1) = 12.6$, $p < 0.001$], reflecting that intelligibility varied throughout a list (bottom-left). We did not grant much credit to this position effect as a linear regression did not reach significance; the effect could be driven by a few occasional mistakes occurring earlier in the list. Finally, intelligibility did not correlate with the subject's age [$p = 0.687$] (Supplementary material 2).

*Behavioral data: recall*

The recall data demonstrated the expected primacy/recency effect (bottom-middle, Fig. 1), which was revealed as a quadratic main effect of position on word recall [$\chi^2(2) = 259.7$, $p < 0.001$]. But the pitch manipulation had no effect, nor did it interact with position. Participants recalled about 6.5 words per list (top-middle, Fig. 1). Average overall recall decreased with the subject's age [$R^2 = 0.17$, $p = 0.015$] (Supplementary material 2).

*Behavioral data: subjective rating*

As expected, participants found the recall task more effortful than the repeat-only task [$\chi^2(1) = 1111.9$, $p < 0.001$] (top-right, Fig. 1), but pitch manipulation had no role. Average effort was unrelated to the subject's age [$p = 0.427$] (Supplementary material 2).

| **Effect of adding fixed term to the model** | **Exp1: within-word F0 manipulation** | | | **Exp2: across-word F0 manipulation** | | |
|---|---|---|---|---|---|---|
| | **Intelligibility** | **Recall** | **Subjective** | **Intelligibility** | **Recall** | **Subjective** |
| *Behavioral data* | | | | | | |
| Recall | $\chi^2(1) = 0.7$, $p = 0.416$ | | $\chi^2(1) = 1111.9$, $p < 0.001$* | $\chi^2(1) = 2.4$, $p = 0.122$ | | $\chi^2(1) = 367.3$, $p < 0.001$* |
| Pitch | $\chi^2(3) = 13.9$, $p = 0.003$* | $\chi^2(3) = 0.3$, $p = 0.954$ | $\chi^2(3) = 0.5$, $p = 0.909$ | $\chi^2(1) = 3.2$, $p = 0.074$ | $\chi^2(1) = 1.5$, $p = 0.219$ | $\chi^2(1) = 1.2$, $p = 0.271$ |
| Position | $\chi^2(1) = 12.6$, $p < 0.001$* | $\chi^2(2) = 259.7$, $p < 0.001$* | | $\chi^2(1) = 1.4$, $p = 0.239$ | $\chi^2(2) = 114.3$, $p < 0.001$* | |
| Recall × pitch | $\chi^2(3) = 2.2$, $p = 0.528$ | | $\chi^2(3) = 0.2$, $p = 0.971$ | $\chi^2(1) = 3.5$, $p = 0.061$ | | $\chi^2(1) < 0.1$, $p = 0.811$ |
| Recall × position | $\chi^2(1) = 0.7$, $p = 0.395$ | | | $\chi^2(1) < 0.1$, $p = 0.999$ | | |
| Pitch × position | $\chi^2(3) = 6.3$, $p = 0.098$ | $\chi^2(6) = 5.0$, $p = 0.547$ | | $\chi^2(1) = 3.7$, $p = 0.055$ | $\chi^2(2) = 5.7$, $p = 0.059$ | |
| 3-way | $\chi^2(3) = 1.3$, $p = 0.738$ | | | $\chi^2(1) < 0.1$, $p = 0.810$ | | |
| **Effect of adding fixed term to the model** | **Exp1: within-word F0 manipulation** | | | **Exp2: across-word F0 manipulation** | | |
| | **Baseline** | **PPD** | **Peak latency** | **Baseline** | **PPD** | **Peak latency** |
| *Pupillary data* | | | | | | |
| Recall | $\chi^2(1) = 917.3$, $p < 0.001$* | $\chi^2(1) = 1.0$, $p = 0.319$ | $\chi^2(1) = 11.3$, $p < 0.001$* | $\chi^2(1) = 178.4$, $p < 0.001$* | $\chi^2(1) = 8.6$, $p = 0.003$* | $\chi^2(1) = 3.1$, $p = 0.080$ |
| Pitch | $\chi^2(3) = 2.0$, $p = 0.577$ | $\chi^2(3) = 12.2$, $p = 0.007$* | $\chi^2(3) = 5.4$, $p = 0.148$ | $\chi^2(1) = 1.8$, $p = 0.181$ | $\chi^2(1) = 6.8$, $p = 0.009$* | $\chi^2(1) = 0.5$, $p = 0.489$ |
| Position | $\chi^2(1) = 2.5$, $p = 0.114$ | $\chi^2(1) = 186.8$, $p < 0.001$* | $\chi^2(1) = 68.7$, $p < 0.001$* | $\chi^2(1) = 0.2$, $p = 0.634$ | $\chi^2(1) = 34.6$, $p < 0.001$* | $\chi^2(1) = 12.8$, $p < 0.001$* |
| Recall × pitch | $\chi^2(3) = 8.5$, $p = 0.037$* | $\chi^2(3) = 2.1$, $p = 0.543$ | $\chi^2(3) = 10.3$, $p = 0.016$* | $\chi^2(1) = 1.2$, $p = 0.267$ | $\chi^2(1) < 0.1$, $p = 0.961$ | $\chi^2(1) < 0.1$, $p = 0.818$ |
| Recall × position | $\chi^2(1) = 192.1$, $p < 0.001$* | $\chi^2(1) = 15.9$, $p < 0.001$* | $\chi^2(1) = 0.8$, $p = 0.384$ | $\chi^2(1) = 55.0$, $p < 0.001$* | $\chi^2(1) = 1.7$, $p = 0.191$ | $\chi^2(1) = 0.1$, $p = 0.7$ |
| Pitch × position | $\chi^2(3) = 3.3$, $p = 0.35$ | $\chi^2(3) = 1.5$, $p = 0.681$ | $\chi^2(3) = 9.2$, $p = 0.027$* | $\chi^2(1) = 0.2$, $p = 0.656$ | $\chi^2(1) = 1.1$, $p = 0.288$ | $\chi^2(1) = 1.7$, $p = 0.192$ |
| 3-Way | $\chi^2(3) = 2.1$, $p = 0.556$ | $\chi^2(3) = 1.1$, $p = 0.768$ | $\chi^2(3) = 6.3$, $p = 0.098$ | $\chi^2(1) < 0.1$, $p = 0.972$ | $\chi^2(1) < 0.1$, $p = 0.772$ | $\chi^2(1) = 0.5$, $p = 0.462$ |

**Table 1.** Results of the statistical analyses of behavioral and pupillary data in both experiments. Significant effects ($p < 0.05$) are marked with an asterisk *.

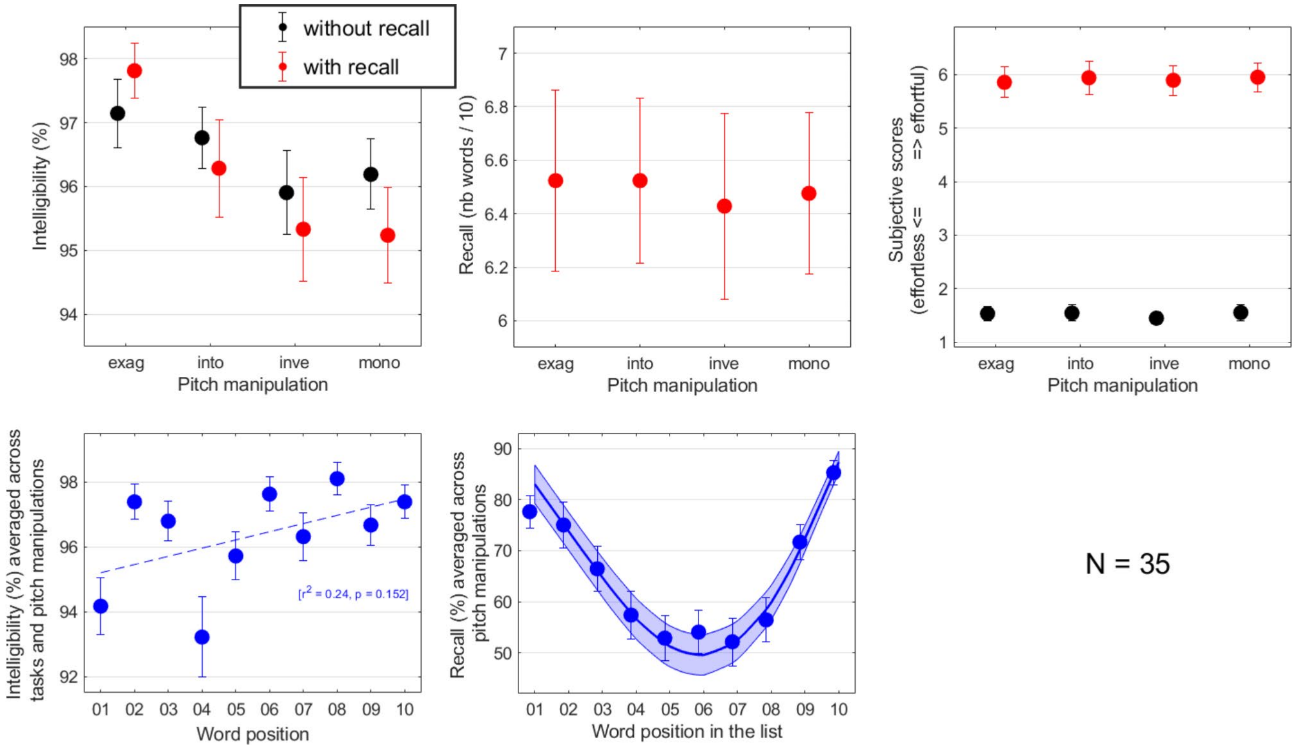

**Fig. 1**. Behavioral data of Exp.1, in which the F0 within individual words was manipulated across 4 conditions: exag = exaggerated pitch contour; into = normal (intact) intonation; inve = inverted contour; mono = monotonized contour.

To summarize the behavioral portion of this experiment, except for slight changes in intelligibility (questionable due to ceiling effects), within-word pitch manipulations had no impact on recall and were equally effortful. All other effects, e.g. recall exhibiting a typical U-shape, recall worsening with age, and recall being more effortful than the repeat-only task, were expected from the literature using the same paradigm[49].

*Pupillary data*

The pupil traces are shown in Figs. 2 and 3, aligned at the response onset, i.e. when subjects were instructed to repeat the word back to the experimenter. On this scale, the word presentation ended at -1 s, and it is clear that the pupil kept on dilating for another half a second after that (bottom panels). It is also apparent that the pupil traces were on average higher in the recall task than in the repeat-only task (top panels). Critically, it is not because participants happened to have different baseline to begin with: as illustrated in the most-top-left panel of Fig. 3, the pupil traces started from a similar range at the beginning of a list, but progressively departed from each other depending on the task. The pupil progressively relaxed throughout the list when subjects only repeated the words, while dilation was maintained at a high level (and progressively incremented) when subjects had to maintain the words in their mind in addition to repeating them. This also led to PPDs that quickly dwindled in amplitude, resulting in a different pattern as a function of position than in the repeat-only task.

To provide more quantitative assessments of these observations, three metrics were focused on: baseline, PPD amplitude, and PPD latency.

*Pupillary data: baseline*

Regarding the baseline of pupillary responses, the full results are displayed in Table 1. There was a main effect of task on baseline [$\chi^2(1) = 917.3$, $p < 0.001$] confirming that the pupil diameter was overall higher in the recall task than in the repeat-only task (top-left, Fig. 4). This was because the pupil kept on increasing slowly throughout the list in the recall task, whereas it progressively relaxed in the repeat-only task (top-right, Fig. 4), resulting in an interaction between task and position [$\chi^2(1) = 192.1$, $p < 0.001$].

There was no main effect of the pitch manipulation on pupil baseline, but it interacted modestly with the task [$\chi^2(3) = 8.5$, $p = 0.037$]. The effect of recall was more prominent with the *mono* than with the *inve* condition (top-left, Fig. 4).

Finally, the average baseline decreased for older subjects [$R^2 = 0.23$, $p = 0.004$] (by 2 mm across a 35-year range - see Supplementary material 2).

*Pupillary data: PPD*

For results regarding PPD of pupillary responses, there was a main effect of position [$\chi^2(1) = 186.8$, $p < 0.001$], confirming that PPD dropped in amplitude quickly from the first to the second word, and kept on decreasing

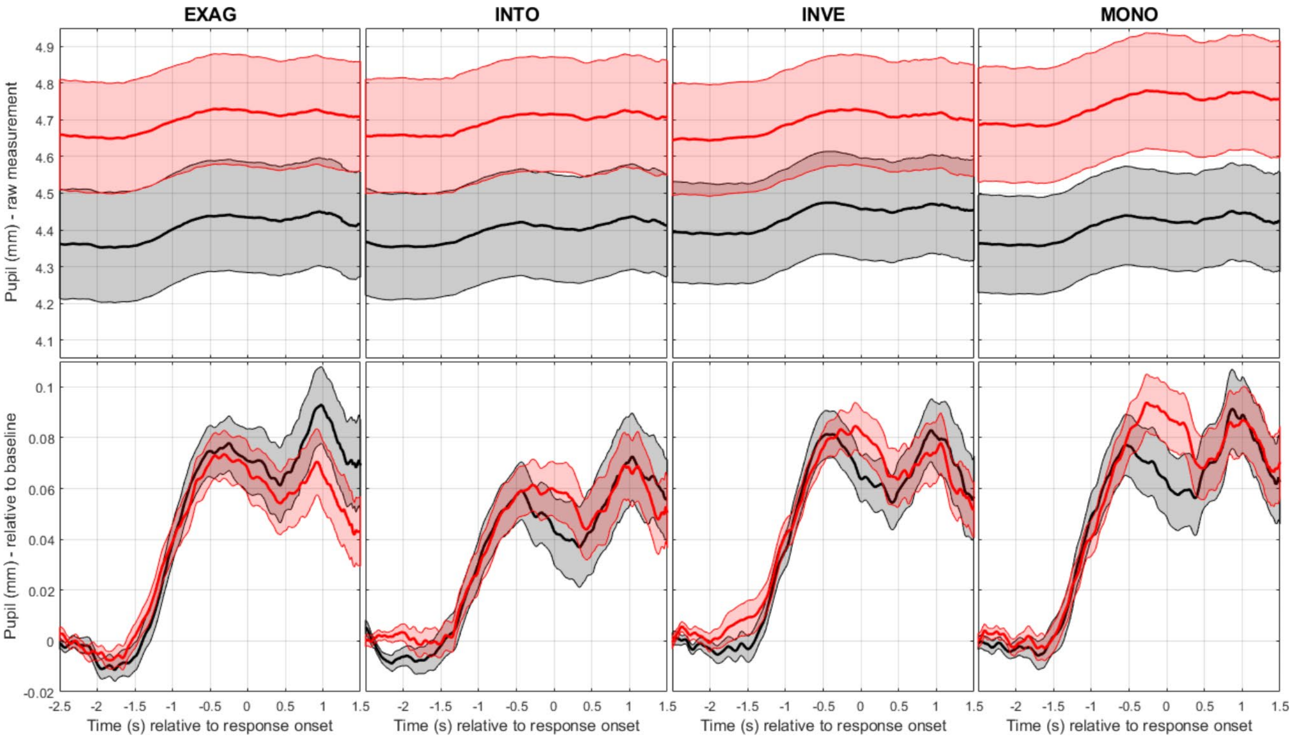

**Fig. 2**. Averaged pupil traces aligned at response onset, as a function of the pitch manipulation within a word: exag = exaggerated pitch contour; into = normal (intact) intonation; inve = inverted contour; mono = monotonized contour. Red traces are from the repeat & recall condition; gray traces are from the repeat-only condition. These traces are pooled across the 10 words of a list and three repetitions (with different lists), expressed in raw units to better appreciate the baseline diameter (top) or baseline-corrected to better appreciate the size and latency of the PPD (bottom). Lines represent the means and the areas reflect one standard error of the mean across subjects.

throughout the rest of the list. This effect was dependent on the task [$\chi^2(1) = 15.9$, $p < 0.001$], being exacerbated in the recall task (middle-right, Fig. 4).

One of the key results was that, despite there being no difference in behavioral recall for the different pitch conditions, within-word pitch manipulation did impact the size of the PPD [$\chi^2(3) = 12.2$, $p = 0.007$]. Post-hoc comparisons revealed that PPD was smaller for the *into* than the *exag* condition, itself smaller than both *inve* and *mono* conditions (which did not differ from each other). In other words, if we take the size of PPD as a metric of the effort engaged in decoding a given word, we would conclude that any degradation from the original pitch contours induced additional effort, and even more when these degradations were inconsistent with the directions of the original contour (middle-left, Fig. 4).

Averaged PPD amplitude did not correlate with subjects' age [$p = 0.306$] (Supplementary material 2).

*Pupillary data: peak latency*

A main effect of task on pupil peak latency [$\chi^2(1) = 11.3$, $p < 0.001$] suggested later pupillary peak latency in recall than in repeat-only (bottom-left, Fig. 4), but this occurred only for the two conditions that led to the largest PPDs, namely *inve* and *mono* conditions, resulting in an interaction with pitch [$\chi^2(3) = 10.3$, $p = 0.016$].

Position also led to a main effect [$\chi^2(1) = 68.7$, $p < 0.001$], reflecting earlier and earlier peaks from the beginning to the end of a list (bottom-right, Fig. 4), and this effect depended on the pitch manipulation [$\chi^2(3) = 9.2$, $p = 0.027$] (not shown). This latter interaction appeared quite complex and to a small degree also dependent on the task (the 3-way interaction approaching significance, $p = 0.098$). For simplicity (and also because this interaction with position did not occur in Exp2), we did not explore it further.

Averaged PPD latency did not correlate with subjects' age [$p = 0.548$] (Supplementary material 2).

## Experiment 2: across-word pitch manipulations

*Behavioral data: intelligibility*

No significant main effects or interactions were found (Table 1). There was a trend such that the *melodic* condition was slightly more intelligible than the *fixed* condition ($p = 0.074$) but the effect size was small: it was only about 1% and intelligibility was > 95% overall, as in experiment 1 (top-left, Fig. 5). Similarly, pitch tended to interact with task ($p = 0.061$) and with position ($p = 0.055$) in this narrow and close-to-ceiling range. Average intelligibility did not depend on the subject's age [$p = 0.638$] (Supplementary material 2).

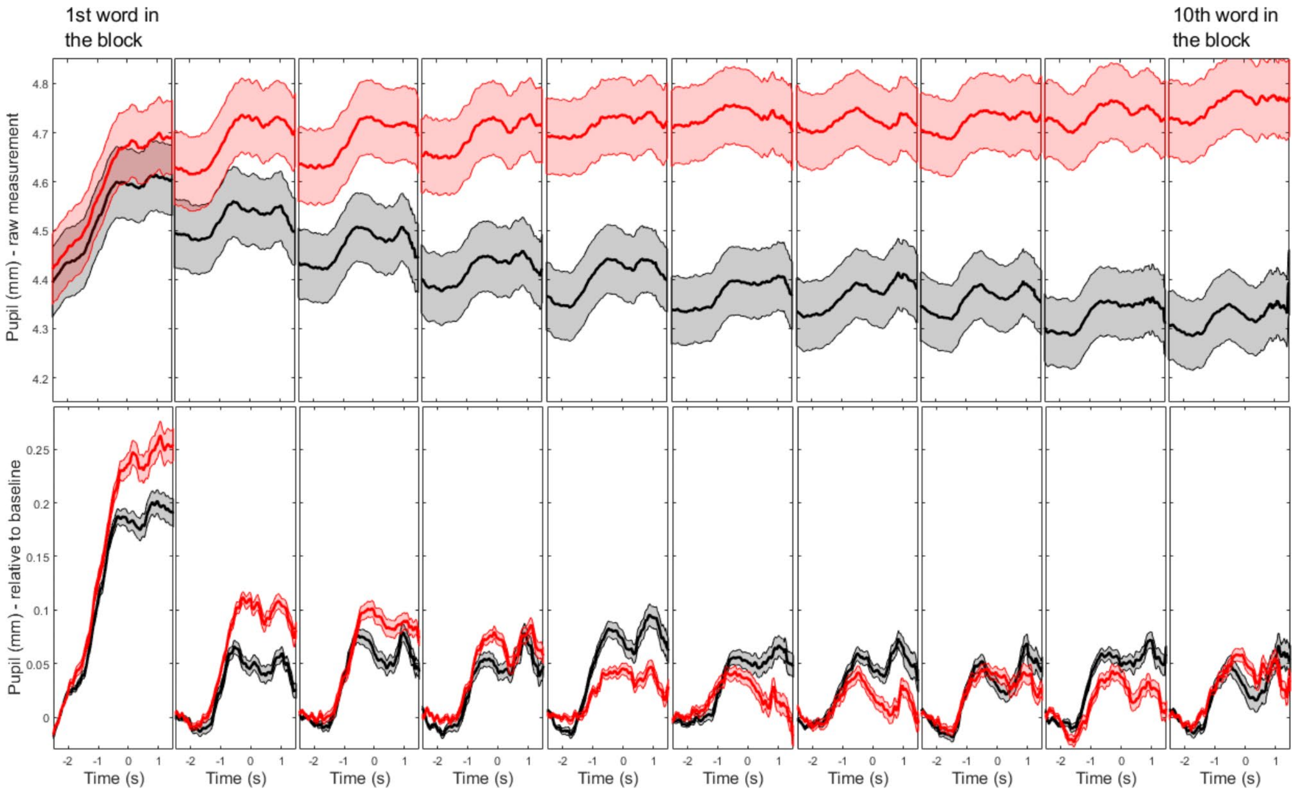

**Fig. 3**. Averaged pupil traces during word listening, aligned at response onset, as a function of the position of a word within a list. These traces are pooled across the 4 pitch manipulations, and expressed in raw units (top) or baseline-corrected (bottom). Red traces are from the repeat & recall condition; gray traces are from the repeat-only condition. Lines represent the means and the areas reflect one standard error of the mean across subjects.

*Behavioral data: recall*

The recall data demonstrated primacy and recency effects (bottom-middle, Fig. 5), yielding a main effect of quadratic position (Table 1). The pitch manipulation had no overall effect, and its interaction with position did not reach significance ($p = 0.059$). Subjects recalled about 6.5 words per list (top-middle, Fig. 5), and this performance decreased with age [$R^2 = 0.16$, $p = 0.045$] (Supplementary material 2).

*Behavioral data: subjective rating*

As in experiment 1, all subjects found the recall task more effortful than the repeat-only task [$\chi^2$ (1) = 367.3, $p < 0.001$] (top-right, Fig. 5), but pitch manipulation had no effect (Table 1). Average effort was not related to the participant's age [$p = 0.281$] (Supplementary material 2).

To summarize, except for slight changes in intelligibility (which never reached significance), the melodic manipulation had no impact. All other effects, e.g. recall exhibiting a typical U-shape, recall worsening with age, and recall being more effortful, replicated previous results[49].

*Pupillary data*

The averaged pupil traces are shown in Figs. 6 and 7. Similarly, the pupil diameter increased during the presentation of a word and kept on dilating for another half a second after that. This pattern is not so obvious when looking at the raw measurements in mm (top panels) but quite clear in the baseline-corrected traces (bottom panels). The raw measurements, on the other hand, illustrate clearly that the pupil was more dilated in the recall task than in the repeat-only task. Once again, this was not a spurious baseline difference between conditions: as illustrated in the most-top-left panel of Fig. 7, participants started roughly at the same baseline, but the pupil traces progressively departed from one another, depending on the task.

To provide quantitative assessments of these observations, three metrics were extracted.

*Pupillary data: baseline*

The full results are displayed in Table 1. There was a main effect of task on pupil baseline [$\chi^2(1) = 178.4$, $p < 0.001$] confirming that the pupil diameter was about 0.2 mm higher in the recall than in the repeat-only task (top-left, Fig. 8). As described earlier, this was because the pupil kept on increasing slowly throughout the list in the recall task, whereas it progressively relaxed in the repeat-only task (top-right, Fig. 8), resulting in an interaction between task and position [$\chi^2(1) = 55.0$, $p < 0.001$].

There was no effect of pitch manipulation or any interaction. The average baseline decreased for older subjects [r2 = 0.22, $p = 0.020$] (by about 2 mm across 30 years - Supplementary material 2).

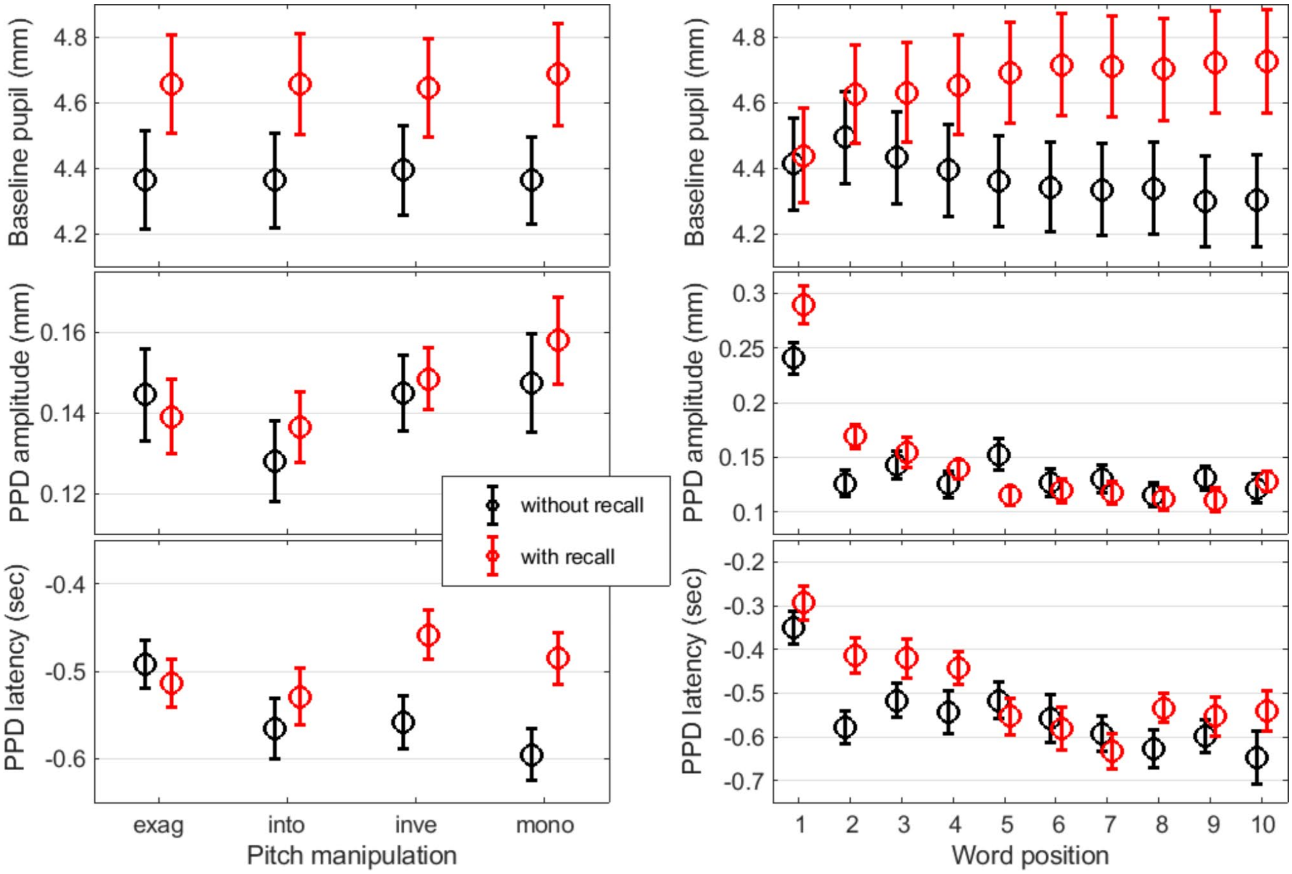

**Fig. 4**. Three metrics extracted from the pupil data in Exp.1: baseline (top), PPD amplitude (middle, i.e., maximum pupil diameter relative to baseline), and PPD latency (bottom, i.e., time between peak dilation time point and the response onset), shown as a function of the experimental manipulation (left panels) or as a function of the word position within a list (right panels). Symbols represent the means and error bars are one standard error from the mean across subjects. The 4 F0 manipulations were: exag = exaggerated pitch contour; into = normal (intact) intonation; inve = inverted contour; mono = monotonized contour.

*Pupillary data: PPD*

There was a main effect of position on PPD [$\chi^2(1) = 34.6$, $p < 0.001$] confirming that PPD diminished in amplitude very quickly from the first to the second position, but was relatively stable after that (middle-right, Fig. 8). In contrast to Exp.1, the effect of task was more homogeneous across the list, and so it appeared as a main effect (rather than an interaction with position) [$\chi^2(1) = 8.6$, $p = 0.003$]. PPD was on average 0.02 mm larger in the recall than in the repeat-only task. Note that this is 10 times smaller than its effect on the baseline.

More to the heart of this study, the pitch manipulation impacted PPD amplitude [$\chi^2(1) = 6.8$, $p = 0.009$]. The PPD was elevated in the *melodic* relative to the *fixed* condition (middle-left, Fig. 8). Also, averaged PPD amplitude tended to decrease with older subjects [r2 = 0.17, $p = 0.043$] (Supplementary material 2).

*Pupillary data: peak latency*

The main effect of task on peak latency did not reach significance [$\chi^2(1) = 3.1$, $p = 0.080$] but would have suggested the same direction as in Exp.1: about 50-ms later peaks in recall than in repeat-only (bottom-left, Fig. 8). Position also led to a main effect [$\chi^2(1) = 12.8$, $p < 0.001$], reflecting earlier and earlier peaks from the beginning to the end of a list (bottom-right, Fig. 8).

More importantly, pitch did not result in a main effect and did not interact with task or position. Average PPD latency did not correlate with subjects' age [$p = 0.105$] (Supplementary material 2).

## Discussion

Overall, we found that task-evoked pupillary response was sensitive to subtle effects generated by the F0 contour within and across words, even when there were only small or negligible effects on intelligibility (in ceiling range), recall or subjective effort ratings. Our results join the past literature to call for 're-visiting' the effect of classic acoustic factors from the perspective of cognitive hearing because intelligibility outcomes might not reveal the full cognitive process involved in speech communication.

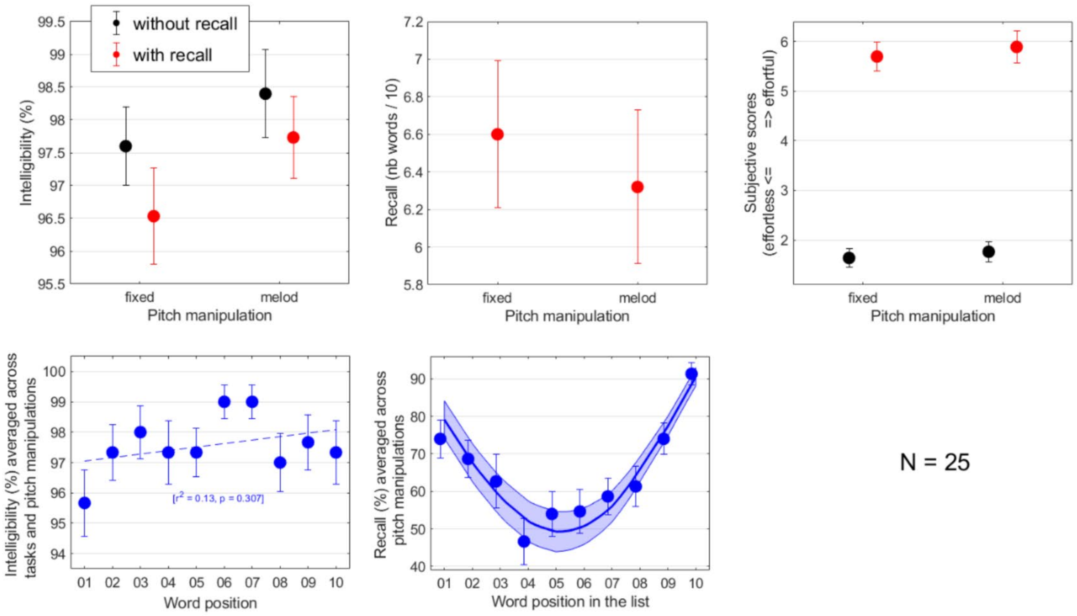

**Fig. 5**. Behavioral data of Exp.2, in which the F0 across individual words was manipulated to form a melody.

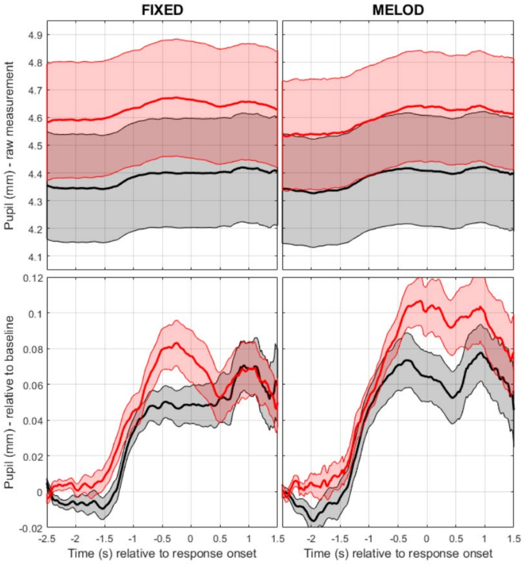

**Fig. 6**. Averaged pupil traces aligned at response onset, as a function of the pitch manipulation across words, to form either a steady pattern (*fixed*) or a melody (*melodic*) that was consistent throughout Experiment 2. Red traces are from the repeat & recall condition; gray traces are from the repeat-only condition. These traces are pooled across the 10 words of a list and three repetitions (with different lists), expressed in raw units to better appreciate the baseline diameter (top) or baseline-corrected to better appreciate the size and latency of the PPD (bottom). Lines represent the means and the areas reflect one standard error of the mean across subjects.

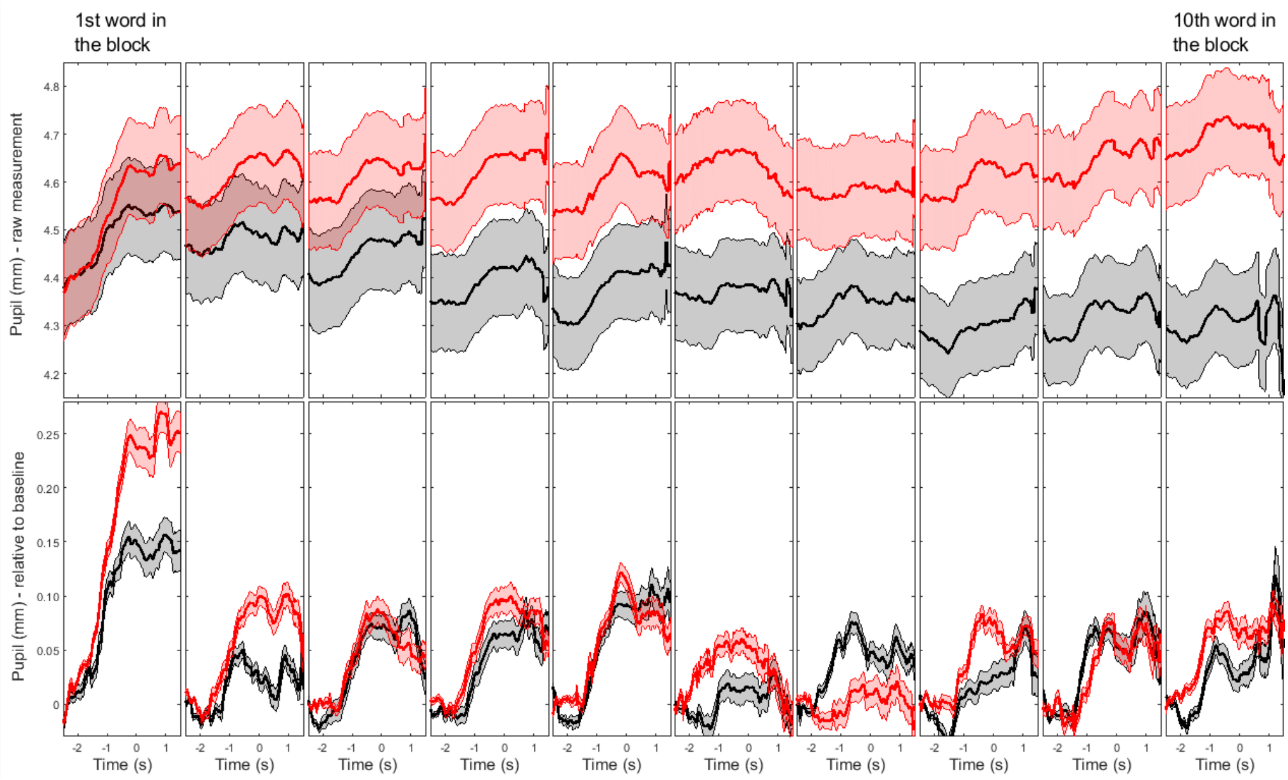

**Fig. 7**. Averaged pupil traces aligned at response onset, as a function of the position of a word within a list. Red traces are from the repeat & recall condition; gray traces are from the repeat-only condition. These traces are pooled across the 2 pitch manipulations, and expressed in raw units (top) or baseline-corrected (bottom). Lines represent the means and the areas reflect one standard error of the mean across subjects.

## Effect of F0 contour manipulation on listening effort

Consistent with our hypothesis for pupillometry outcomes, the PPD was larger in both monotonized and inverted conditions, suggesting that these conditions required extra resources to process. For behavioral outcomes, we replicated the poorer intelligibility of these two conditions as it has been found in sentences with noise[57,58] and interrupted speech[28,29], though the impairment was minimal in our current study because individual words in quiet can hardly be misunderstood (ceiling performance). Note that this was by design: had we presented the words in background noise (to lower intelligibility), the periodicity information would have been less well extracted (or would have served other purposes, e.g. segregation).

These results demonstrate that, even in NH non-tonal language native speaking individuals, pupillometry is sensitive to unnatural and distorted F0 inflections, and we interpret differences in response to reflect explicit investment of cognitive resources. Contrary to our hypothesis that exaggerated F0 contour might ease listening effort, results showed that all distortions from the original F0 contour induced additional effort, but more so when the distortions were inconsistent with the directions of the original contour (i.e., flat and inverted conditions). This would be explained by the ELU model in the following manner: the phonological representation of monotonized or inverted words becomes harder to match with a "template" of a known word in the listener's long-term memory. To make the match, explicit engagement of extra cognitive resources is necessary, and this extra engagement, can be revealed by the task-evoked pupillary response (even when behavioral response and subjective reflection do not capture it). Past study has shown that inverting F0 contour impaired phonemic restoration and interrupted speech recognition, while flattening F0 contour did not[28]. However, our results seem to suggest that in terms of listening effort in word recognition, flattening F0 contour was as effortful as inverting F0 contour. This might suggest that there are further complexities in F0 information integration throughout the duration of a sentence, than what were reported in past studies looking only at behavioral outcomes.

## Interaction between F0 contour manipulation and recall on pupillary response

In Experiment 1, pupil traces differed markedly depending on the task (recall vs. repeat), consistent with past studies that investigate the effect of recall on pupillary response. Although baseline dilation increased at the beginning of each list for both tasks, they diverged after the first 2–3 words. The repeat-only task was easier, and as a consequence the pupil relaxed, ending up 0.2 mm lower than where it started. In contrast, recalling as many words as possible demanded sustained cognitive resources, leading to a continued increase in pupil diameter. Initial PPD values were high, but these decreased over the course of the 10-word list, meaning that average PPD could be inordinately affected by the first word. These effects of task and position are interesting and generally warn us about the limitations of a single metric such as PPD[46,49,59].

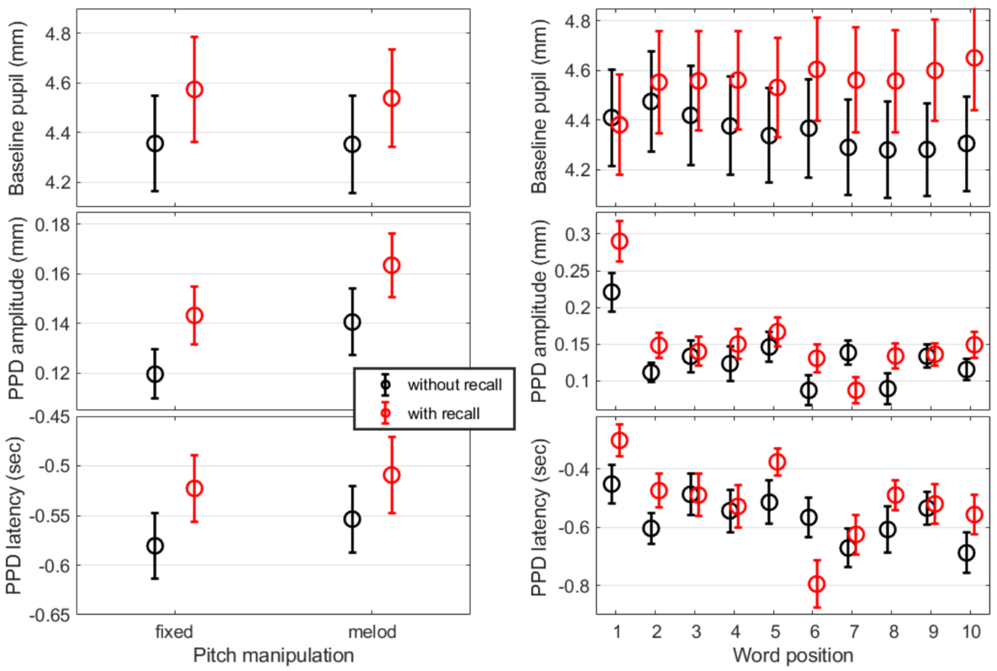

**Fig. 8**. Three metrics extracted from the pupil data in Exp.2: baseline (top), PPD amplitude (middle), and PPD latency (bottom), shown as a function of the pitch manipulation (left panels) or as a function of the word position within a list (right panels). Symbols represent the means and error bars are one standard error from the mean across subjects.

Additionally, we observed two interesting interactions between pitch and task conditions on another two pupillary responses: baseline and PPD latency. The latency effect is relatively easy to interpret since the effect of recall (delaying the peak dilation) was seen exclusively for the conditions that were already difficult to process, namely monotonized and inverted pitch contours. The effect of the task would therefore exacerbate any initial difficulty in word decoding. The baseline effect is not as easy to interpret: it is possible that the absence of pitch information (or more accurately the fact that it is uninformative) in monotonized words would make them even more difficult to match to stored templates than inverted words (which have some - but potentially misleading - pitch information). In turn, expecting extra difficulty in the whole monotonized condition would exacerbate the overall effort induced by recall[60]. In other words, both interactions might point to the idea that the recall task is more costly for listening situations that were already taxing resources.

In Experiment 2, we found again that pupillometry was sensitive to pitch manipulations, this time across words instead of within words. This is a different interpretation, no longer about the phonological representation of words but more about their encoding as a sequence forming a melody via their respective F0s. A complete understanding of this phenomenon is not trivial. Although these manipulations did not have behavioral consequences here, we know that they *can in principle* affect memory performance. Sares et al.[38] tested free recall performance after presenting word lists with different pitch sequences, and found that there was an improvement in free recall only when the sequences indicated a grouping (arpeggios). Since the *melodic* condition in this experiment had a similar pattern to the arpeggios in Sares et al. (2023), one might have expected the *melodic* condition to lead to better recall than the *fixed* condition. Not only was this not the case, but performance in the *melodic* condition tended to be worse. One important parameter differentiating the two designs is speed of presentation: while in Sares et al. (2023) the words were presented close in time (less than 1 s apart), in a way which could facilitate grouping and pitch pattern recognition, here the words were presented around 5 s apart (1s waitpeak + participants repeating back the words), a timescale which may be perceived quite differently[61]. The flattened profile of ten words in a sequence with short intervals makes it difficult to retain them in memory since items that are more distinct from one another could be easier to store and retrieve than items that share common features. For instance, several studies showed that, in NH listeners, the recall of words that rhyme is poorer than the recall of words that do not rhyme (Baddeley, 1966; Conrad & Hull, 1964; Nittrouer et al., 2013;

Salamé & Baddeley, 1986). Another possibility could be that on this larger timescale, the role of pitch is not to facilitate memory encoding, but rather to support participant's engagement with the task. Previous study has shown that pupillary response is sensitive to changes in task engagement even when behavioral performance is the same (similar in our case)[60]. However, we did not observe differences in baseline pupil diameter between two manipulations that were typically related with anticipating or mobilizing attention[66–68]. We did not have any other markers to estimate task engagement, nor a physiological correlate of arousal and stress (e.g. salivary cortisol, skin conductance, or heart rate recordings). The *melodic* condition resulted in larger PPDs but it remains unclear whether this is a sign of additional effort in processing and storing words, or whether it reflects a more enthusiastic engagement towards this condition.

### Comparing and synthesizing behavioural, objective and subjective outcomes

In summary, pitch manipulations applied in both Experiment 1 and Experiment 2 (within-word and across-word) had small effects on intelligibility (in ceiling region) and did not significantly affect immediate recall or subjective difficulty ratings. It is surprising that immediate recall did not show any significant difference, especially considering how similar dual-task behavioral paradigms have shown to be sensitive to different acoustic manipulations[31,32]. This lack of sensitivity could be due to the use of words rather than sentences. In our case, recall was likely tapping into a phonological loop of recently stored monosyllabic words, so the impact of acoustic manipulation could be more heavily influenced by the recency effect. When using longer and more complex stimuli, for instance the SWIR, the impact of an experimental manipulation might rely more on the primacy effect, where sentence-final words are transferred to long term memory. Additionally, the effect size of our acoustic manipulations might be smaller compared to other manipulations (i.e., SNR, noise reduction turned on and off), hence harder to observe with the current statistical power.

On the other hand, the *pupil responses did differ*, suggesting that pupillary responses might be sensitive to subtleties in the allocation of cognitive resources. These subtleties are meaningful within the ELU framework: the more matched the acoustic inputs with the stored template of pitch contour, the less need for explicit cognitive resources to resolve the mismatch, hence the smaller the pupillary response. This is presumably why the inverted and monotone pitch contours led to increased pupil dilation and response latency, especially during the recall task, as well as slightly decreased intelligibility. Experiment 2 showed increased pupil dilation and response latency for the melodic pitch condition, and intelligibility was slightly *increased* (though not significantly). This is the opposite relationship to Experiment 1.

The dichotomy between listening effort and task performance is seen elsewhere in past studies using pupillometry to quantify effort. For instance, in speech recognition tasks, elevated noise during listening can lead to greater pupil dilation and poorer intelligibility or recall, up to a point[40,45]. In memory tasks, greater pupil dilation is seen for words that are correctly recalled compared to those that are not recalled (see details of the replication results in Supplementary material 1) and also in previous work[59]. Thus, though pupillometry is a powerful and sensitive technique to register fluctuation in physiological responses to cognitive demands, it is difficult to interpret pupil signals in the absence of the task demands and behavioral outcomes. An increase in task-evoked pupillary response or cognitive resources expenditure is not necessarily a negative marker; it can reflect engagement and be followed by successful completion of a more complex task. Ultimately, it is individual differences in cognitive and motivational status that decide whether the expenditure of cognitive resources is perceived as negative or positive by the listeners (Carolan et al., 2022; Herrmann & Johnsrude, 2020; Pichora-Fuller et al., 2016).

Note that the current experiment was conducted in NH listeners. Future studies should investigate whether these findings extend to hearing impaired populations and specifically CI users. The auditory inputs from a CI contain less salient and sometimes distorted F0 cues (e.g. incomplete array insertion). Although many factors contribute to the challenging and effortful speech recognition in CI users, the importance of F0 saliency and fidelity cannot be ignored in speech recognition and associated listening effort. This is not only due to the importance of F0 in transmitting prosodic information (i.e., intonation, emotion, etc.), but also in decoding the words themselves, since a word's F0 contour is part of its identity. Whether CI users completely ignore pitch in their phonological representation, or whether they struggle with it because it lacks discriminative power, is an open question which we hope to address in a future study.

## Methods

### Participants

In Experiment 1, we recruited a group of 35 adults (12 male, 23 female) between the ages of 18 and 51 (mean ± SD = 25.7 ± 8.2). Participants had normal hearing); defined as having no history of audiological problems and having binaural pure tone thresholds better than 30 dB HL at 0.25, 0.5, 1, 2, 4, 8 kHz. All participants were native speakers of either French or English (the two most common languages in the city; 14 native French speakers). The experiment was always run in their native language (another study[81] investigated language background effects in such paradigms).

In Experiment 2, we recruited 25 adults (10 male, 15 female) between the ages of 18 and 51 with (mean ± SD = 24.6 ± 7.4) years. All but three had participated in experiment 1, but the two experiments were treated separately. No statistical comparison was made between the two experiments.

This work received ethical approvals from McGill University Faculty of Medicine Research Ethics Board (IRB) under the number A05-B11-18B. Both studies were performed in accordance with McGill University ethical guidelines and regulations. Prior to the experiment, participants were given enough time to read the information sheet and consent form approved by the ethics board. All gave written informed consent for their participation.

## Stimuli

Word stimuli were consonant-nucleus-consonant (CNC) words[73] recorded by a male native American English speaker, and Fournier words[74] recorded by a male French speaker. The STRAIGHT algorithm[75] was then used to manipulate the F0 of each word.

In experiment 1, there were 4 conditions: (1) a monotonized condition (*mono*) where F0 remained steady at a fixed value for all words (this value was 121 Hz as the mean F0 of the entire English material, or 101 Hz as the mean F0 of the entire French material); (2) a naturally intonated (*into*) condition where F0 was unprocessed; (3) an exaggerated (*exag*) condition where the fluctuations of the naturally intonated F0 pattern relative to the monotone were doubled on a logarithmic scale (since pitch is perceived logarithmically); (4) an inverted condition (*inve)* where the fluctuations of the naturally intonated F0 pattern were flipped upside-down around the mean F0 for that word. These corresponded to the parameter $k$ being 0, 1, 2, and $-1$, respectively for monotonized, intonated, exaggerated, and inverted, in the Eq. (1) below:

$$F0' = ref\ \times\ 2 \wedge (k \times log2\,(F0/F0m)) \tag{1}$$

where *F0'* is the manipulated F0 contour, *ref* the mean F0 of 121–101 Hz depending on the language material (due to different average F0 in English and French words), ***k*** the parameter differentiating the 4 conditions, *F0m* the mean F0 respective to each sentence, and F0 the original F0 contour.

Note that all these manipulations have been validated before, in the context of speech intelligibility for NH adults[57]. Finally, words were grouped into lists of ten. Word lists were randomized across block repetitions and participants.

In experiment 2, all words were first monotonized to a fixed value (*ref*). Then the F0 of these words was manipulated in 2 conditions: a fixed condition where F0 remains steady for all words; and a melodic condition where F0 of each word within the 10-word block follows the notes: do (*ref*), mi (4 semitones over *ref*), sol (7 semitones), re (2 semitones), fa (5 semitones), la (9 semitones), re# (3 semitones), fa# (6 semitones), si (11 semitones) and sol# (8 semitones).

## Procedure

Participants sat on a stable chair in a soundproof room, 2.5 m in front of a 35-inch screen monitor and wearing an infrared binocular eye tracker (Tobii Glasses Pro2, 50 Hz sampling rate). The room and screen luminance levels were adjusted to reach 80 lx (measured using a luxometer with the sensor positioned at the same height of the participants' left eye and facing the screen). The luminance levels were fixed throughout the experiments, to avoid changes in light level inducing task-unrelated pupillary response and to maximize task-related responses[76]. All audio stimuli were presented through a Beyer Dynamics DT 990 Pro headphone via an external soundcard (Edirol UA), calibrated at 65 dB SPL. Experiments were run in Matlab 2016b, using Psychtoolbox and custom scripts.

On each block, before listening, a word on the screen indicated that the participant should try to remember the words (RECALL, printed in red) or not (NO RECALL, printed in black). After the notification, there was a 3 s pause, and then participants listened to a list of 10 words. Participants were instructed to fixate on a light gray fixation cross displayed at the center of the dark gray screen during this time. 1.5 s before the onset of each word presentation served as time for pupil to restore from the previous trial (0.5s) and baseline measurement (1s). At the offset of the word presentation, participants were instructed to not respond immediately but keep fixating on the gray fixation cross for 1s to wait for the late-occuring pupil peak to appear ('waitpeak' period). At the end of the waitpeak stage, participants were prompted by a light gray circle at the center of the screen to repeat the word aloud. The spoken responses were scored by the experimenter based on whether the word was correctly repeated. In blocks requiring recall, at the end of the tenth word, participants were prompted by the word RECALL followed by a light gray circle on the screen to recall as many words as possible from the previous ten words, in any order. All the signs on the screen during the listening task were the same size and hue to avoid pupil responses caused by visual inputs. Again, the spoken responses were typed down by the experimenter. At the end of each block, participants were asked to rate how effortful the last block was from 1 to 10, with 10 being most effortful. Their subjective ratings were typed down and the program proceeded to the next block. Figure 9 illustrates the procedure. Altogether, each pitch condition was repeated for six blocks, three blocks requiring recall and three blocks not requiring recall (24 blocks in total in Experiment 1; 12 blocks in total in Experiment 2). Block sequences were randomized for each participant.

## Behavioral data analysis

There were three dependent variables (DVs) in the behavioral data: (1) word recognition, (2) word recall, and (3) subjective rating of effort. The first two were binary and the third a score from 0 to 10. Logistic mixed-effect models were fitted to the data in the first two cases (using the Matlab function fitglme with a binomial distribution), and linear mixed-effects models in the third case (using the Matlab function fitlme).

Note that language was not considered as a factor in the analysis. Although language background is potentially a factor of interest for isolated word processing including their short-term storage[81], its impact is subtle and less relevant when speech materials become artificially manipulated (i.e. non-ecological F0 contours). Such effects usually require a large sample size. In this study that was designed to investigate impact of pitch degradation, we did not have sufficient statistical power to examine this factor. Furthermore, the statistical models utilized had captured some of the language variability by using *by subject* as random factors (see further below).

For word recognition, we considered three fixed factors: pitch manipulation (either 4 levels in Exp1, or 2 levels in Exp2), task (i.e. repeat vs. recall), and position (01 to 10). We did not expect the position of the word within a

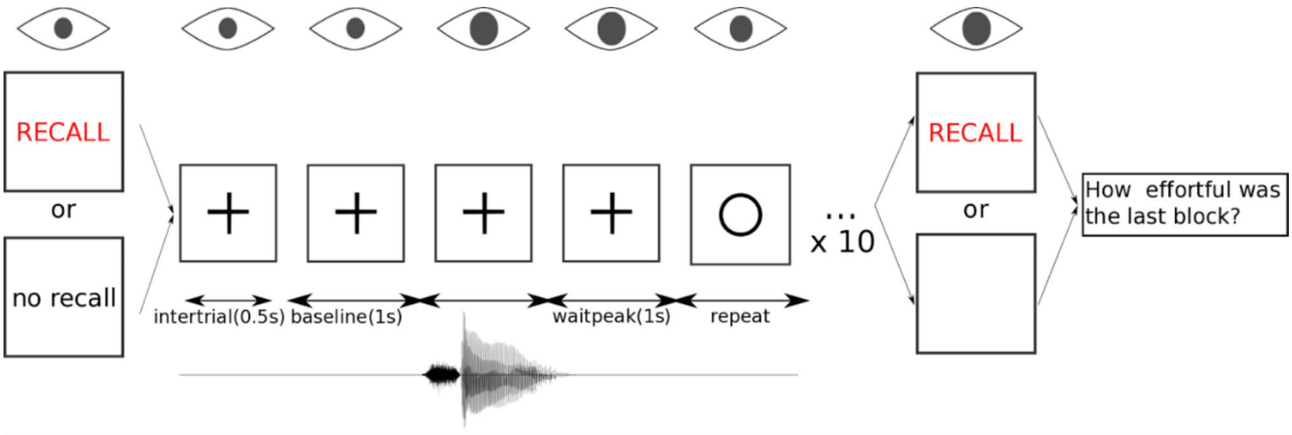

**Fig. 9**. Schematic illustration of the experimental procedure.

list to have much impact on intelligibility but kept it because of its role on recall (and on the pupil metrics). For word recall (only available in the recall task, not the repeat-only task), there were only two fixed factors: pitch manipulation and position (with linear and quadratic terms to account for primacy/recency effects). For effort rating, the DV existed only per list (not per position) so there were only two fixed factors: pitch manipulation and task.

Mixed effect models allow for controlling the variance associated with random effect factors without data aggregation. Therefore, using *by subject* and *by list* random effect factors in the model, we controlled for the variance in overall performance (random intercept) and dependency on other fixed factors (random slope) that were associated with the two factors. Random terms entered the model, and only remained in the model if they significantly improved the model fitting, using Chi-squared tests based on changes in deviance ($p<0.05$). Random terms *by list* and random slope *by subject* did not significantly improve the model fitting. On the other hand, the model fitting improved significantly with random intercepts *by subject*. We also considered random slopes by subject for the effect of pitch, task, and position, and this did improve the models but only for certain fixed factors, and not consistently across DVs. Since by-subject random slopes were not helpful for pupil data, we aimed for consistency across the different analyses and restricted random terms to intercepts by subject. This does not yield the most refined model for any given analysis, but it has the advantage of being systematic throughout this article (across the two experiments).

Thus, the final model for word recognition was:

$$\mathrm{DV} \sim \mathrm{task} * \mathrm{pitch} * \mathrm{position} + (1|\mathrm{subject})$$

The final model for word recall was:

$$\mathrm{DV} \sim \mathrm{pitch} * \mathrm{position}^2 + (1|\mathrm{subject})$$

The final model for effort rating was:

$$\mathrm{DV} \sim \mathrm{pitch} * \mathrm{task} + (1|\mathrm{subject})$$

All main effects and interactions were determined with a Chi-square test between the model with and without the term in question. As a supplementary analysis, averaged DVs were correlated against the age of participants.

## Pupil data analysis

Prior to analysis, pupil traces were cleaned using procedures consistent with previous studies[40,49,77]. Pupil diameter values below 3 standard deviations (SD) of the mean of the whole recording were coded as blinks. Traces within 25 data points before the start and after the end of the blink were cubically interpolated in Matlab, to decrease the impact of the obscured pupil from blinks. Trials that had over 20% of data points counted as blinks were excluded. All valid traces were then low-pass filtered at 10 Hz with a first-order Butterworth filter to preserve only cognitively relevant pupil size modulation[78]. Processed traces were aligned by the onset of the response prompt (the display of light gray circle to signal participants to repeat back the word) and aggregated per listener and per condition.

For the word recognition phase of each block, (listening and repeating the words), baseline pupil diameters were calculated as the averaged value of the pupil trace over 1 s, just before the start of each word. For the word recall section of the test, baseline pupil diameter was calculated as the mean of the ten previous baseline pupil diameters (see Supplementary material 1 for analysis). The rest of the pupil diameter indices were corrected by the corresponding baselines (i.e., baseline pupil size was subtracted from the trace). This event-related trace was used to analyze pupil size change evoked by word recognition and recall.

Over the presentation of the 10 words of a list, three DVs of pupil response were obtained: pupil baseline, PPD amplitude, and PPD latency. They were extracted within the time window between the onset of word and onset of verbal response, thus excluding any period where the pupil could have fluctuated for articulatory reasons.

Three linear mixed effects models were fitted, one for each DV, with an identical model of the form:

$$\mathrm{DV} \sim \mathrm{task} * \mathrm{pitch} * \mathrm{position} + (1|\mathrm{subject})$$

All main effects and interactions were determined with a Chi-squared test between the model with and without the fixed term in question. During the model building, random term *by list* did not significantly improve the model fitting. We also calculated the correlation between each DV and the age of participants.

Finally, to examine the replicability of recently published results using similar experimental paradigms, two additional analyses were conducted, and their methods and results are presented in Supplementary Material 1. Although comparisons of task performance and task-evoked pupillary responses are of central interest in testing our scientific hypothesis, these additional analyses and results can help us to understand better whether previous findings are spurious or can be observed in other independent studies.

## Data availability

The datasets generated during and/or analysed during the current study are available in the Open Science Framework repository, https://osf.io/74mpy/.

Received: 9 August 2023; Accepted: 16 September 2024
Published online: 30 September 2024

## References

1. World Health Organization. Challenges facing ear and hearing care. *World Rep. Hear.* 139–198 (2021).
2. Livingston, G. et al. Dementia prevention, intervention, and care. *Lancet*. **390**, 2673–2734 (2017).
3. Livingston, G. et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet (London England)*. **396**, 413 (2020).
4. Blamey, P. et al. Factors affecting auditory performance of Postlinguistically Deaf adults using cochlear implants: An update with 2251 patients. *Audiol. Neurotol.* **18**, 36–47 (2012).
5. Holden, L. K. et al. Factors affecting Open-Set Word Recognition in adults with cochlear implants. *Ear Hear.* **34**, 342 (2013).
6. Dorman, M. F. & Gifford, R. H. Speech understanding in Complex listening environments by listeners fit with cochlear implants. *J. Speech Lang. Hear. Res.* **60**, 3019–3026 (2017).
7. Vermeire, K. et al. Quality-of-life benefit from cochlear implantation in the elderly. *Otol Neurotol.* **26**, 188–195 (2005).
8. Loizou, P. C. Mimicking the human ear. *IEEE Signal. Process. Mag.* **15**, 101–130 (1998).
9. Shannon, R. V. Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics. *Hear. Res.* **11**, 157–189 (1983).
10. Zeng, F. G. Temporal pitch in electric hearing. *Hear. Res.* **174**, 101–106 (2002).
11. Gfeller, K. et al. Musical backgrounds, listening habits, and aesthetic enjoyment of adult cochlear implant recipients. *J. Am. Acad. Audiol.* **11**, 390–406 (2000).
12. Kong, Y. Y., Cruz, R., Jones, J. A. & Zeng, F. G. Music perception with temporal cues in Acoustic and Electric hearing. *Ear Hear.* **25**, 173–185 (2004).
13. Peng, S. C., Lu, N. & Chatterjee, M. Effects of cooperating and conflicting cues on Speech Intonation Recognition by Cochlear Implant users and normal hearing listeners. *Audiol. Neurotol.* **14**, 327–337 (2009).
14. Fu, Q. J., Chinchilla, S., Nogaki, G. & Galvin, J. J. Voice gender identification by cochlear implant users: the role of spectral and temporal resolution. *J. Acoust. Soc. Am.* **118**, 1711–1718 (2005).
15. Stickney, G. S., Assmann, P. F., Chang, J. & Zeng, F. G. Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences. *J. Acoust. Soc. Am.* **122**, 1069–1078 (2007).
16. Moore, B. C. J. The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *JARO - J. Assoc. Res. Otolaryngol.* **9**, 399–406 (2008).
17. Bernstein, J. G. W. & Oxenham, A. J. The relationship between frequency selectivity and pitch discrimination: sensorineural hearing loss. *J. Acoust. Soc. Am.* **120**, 3929–3945 (2006).
18. Jusczyk, P. W. Narrowing the distance to language: one step at a time. *J. Commun. Disord.* **32**, 207–222 (1999).
19. Soderstrom, M., Seidl, A., Kemler Nelson, D. G. & Jusczyk, P. W. The prosodic bootstrapping of phrases: evidence from prelinguistic infants. *J. Mem. Lang.* **49**, 249–267 (2003).
20. Thiessen, E. D., Hill, E. A. & Saffran, J. R. Infant-directed speech facilitates word segmentation. *Infancy*. **7**, 53–71 (2005).
21. Deroche, M. L. D., Lu, H. P., Limb, C. J., Lin, Y. S. & Chatterjee, M. Deficits in the pitch sensitivity of cochlear-implanted children speaking English or Mandarin. *Front. Neurosci.* **8**, 103393 (2014).
22. Hällgren, M., Larsby, B., Lyxell, B. & Arlinger, S. Speech understanding in quiet and noise, with and without hearing aids (2009). http://dx.doi.org/10.1080/14992020500190011 44, 574–583
23. Nachtegaal, J. et al. Hearing status, need for recovery after work, and psychosocial work characteristics: results from an internet-based national survey on hearing  (2009). http://dx.doi.org/10.1080/14992020902962421 48, 684–691
24. Winn, M. Rapid Release from listening effort resulting from semantic context, and effects of Spectral Degradation and Cochlear implants. *Trends Hear.* **20**, (2016).
25. Rönnberg, J. et al. The ease of Language understanding (ELU) model: theoretical, empirical, and clinical advances. *Front. Syst. Neurosci.* **7**, 48891 (2013).
26. Pichora-Fuller, M. K. et al. Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). *Ear Hear.* **37**, 5S–27S (2016).
27. Rönnberg, J. et al. Hearing impairment, cognition and speech understanding: exploratory factor analyses of a comprehensive test battery for a group of hearing aid users, the n200 study. *Int. J. Audiol.* **55**, (2016).
28. Clarke, J. et al. Effect of F0 contours on top-down repair of interrupted speech. *J. Acoust. Soc. Am.* **142**, EL7–EL12 (2017).
29. Clarke, J., Gaudrain, E., Chatterjee, M. & Başkent, D. T'ain't the way you say it, it's what you say – perceptual continuity of voice and top–down restoration of speech. *Hear. Res.* **315**, 80–87 (2014).
30. Winn, M. B., Edwards, J. R. & Litovsky, R. Y. The impact of Auditory Spectral Resolution on listening Effort revealed by Pupil Dilation. *Ear Hear.* **36**, e153 (2015).

31. Sarampalis, A., Kalluri, S., Edwards, B. & Hafter, E. Objective measures of listening effort: effects of background noise and noise reduction. *J. Speech Lang. Hear. Res.* **52**, 1230–1240 (2009).
32. Ng, E. H. N., Rudner, M., Lunner, T., Pedersen, M. S. & Rönnberg, J. Effects of noise and working memory capacity on memory processing of speech for hearing-aid users (2013). https://doi.org/10.3109/14992027.2013.776181. *776181* 52, 433–441 (2013).
33. Lunner, T., Rudner, M., Rosenbom, T., Ågren, J. & Ng, E. H. N. using speech recall in hearing aid fitting and outcome evaluation under ecological test conditions. *Ear Hear.* **37**, 145S–154S (2016).
34. Frankish, C. Perceptual Organization and Precategorical Acoustic Storage. *J. Exp. Psychol. Learn. Mem. Cogn.* **15**, 469–479 (1989).
35. Frankish, C. Intonation and auditory grouping in immediate serial recall. *Appl. Cogn. Psychol.* **9**, S5–S22 (1995).
36. McElhinney, M. & Annett, J. M. Pattern of Efficacy of a musical mnemonic on recall of familiar words over several presentations (1996). http://dx.doi.org/10.2466/pms.1996.82.2. *395* 84, 395–400.
37. Savino, M., Winter, B., Bosco, A. & Grice, M. Intonation does aid serial recall after all. *Psychon Bull. Rev.* **27**, 366–372 (2020).
38. Sares, A. G. et al. Grouping by Time and Pitch facilitates free but not cued Recall for Word lists in normally-hearing listeners. *Trends Hear.* **27**, (2023).
39. Granholm, E., Asarnow, R. F., Sarkin, A. J. & Dykes, K. L. Pupillary responses index cognitive resource limitations. *Psychophysiology*. **33**, 457–461 (1996).
40. Zekveld, A. A. & Kramer, S. E. Cognitive processing load across a wide range of listening conditions: insights from pupillometry. *Psychophysiology*. **51**, 277–284 (2014).
41. Zekveld, A. A., Kramer, S. E. & Festen, J. M. Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. *Ear Hear.* **32**, 498–510 (2011).
42. Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A. & Kramer, S. E. The pupil response is sensitive to divided attention during speech processing. *Hear. Res.* **312**, 114–120 (2014).
43. Koelewijn, T., Zekveld, A. A., Festen, J. M. & Kramer, S. E. Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear Hear.* **33**, 291–300 (2012).
44. Ohlenforst, B. et al. Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. *Hear. Res.* **365**, 90–99 (2018).
45. Ohlenforst, B. et al. Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hear. Res.* **351**, 68–79 (2017).
46. Zekveld, A. A., Kramer, S. E., Rönnberg, J. & Rudner, M. In a concurrent memory and auditory perception Task, the Pupil Dilation response is more sensitive to memory load than to auditory stimulus characteristics. *Ear Hear.* **40**, 272–286 (2019).
47. Micula, A. et al. The effects of Task Difficulty predictability and noise reduction on Recall performance and pupil dilation responses. *Ear Hear.* **42**, 1668 (2021).
48. Bönitz, H. et al. How do we allocate our resources when listening and memorizing Speech in noise? A Pupillometry Study. *Ear Hear.* **42**, 846–859 (2021).
49. Zhang, Y., Lehmann, A., & Deroche M. Disentangling listening effort and memory load beyond behavioural evidence: Pupillary response to listening effort during a concurrent memory task. *PLoS One*. **16**, e0233251 (2021).
50. Zekveld, A. A., Koelewijn, T. & Kramer, S. E. The Pupil Dilation Response to Auditory Stimuli: Current state of knowledge. *Trends Hear.* **22**, (2018).
51. Singh, L., Nestor, S., Parikh, C. & Yull, A. Influences of infant-directed speech on early word recognition. *Infancy*. **14**, 654–666 (2009).
52. López, S. et al. Vocal caricatures reveal signatures of speaker identity. *Sci. Rep. 2013*. **31** (3), 1–7 (2013).
53. Schweinberger, S. R. Eiff, C. I. Enhancing socio-emotional communication and quality of life in young cochlear implant recipients: perspectives from parameter-specific morphing and caricaturing. *Front. Neurosci.* **16**, 956917 (2022). von.
54. Racette, A. & Peretz, I. Learning lyrics: to sing or not to sing? *Mem. Cognit.* **35**, 242–253 (2007).
55. Purnell-Webb, P. & Speelman, C. P. Effects of Music on Memory for Text. (2008). http://dx.doi.org/10.2466/pms.106.3. *927-957* 106, 927–957.
56. Savino, M., Bosco, A. & Grice, M. Intonational cues to item position in lists: Evidence from a serial recall task. *Speech Prosody* 708–712 (2017).
57. Binns, C. & Culling, J. F. The role of fundamental frequency contours in the perception of speech against interfering speech. *J. Acoust. Soc. Am.* **122**, 1765–1776 (2007).
58. Meister, H., Landwehr, M., Pyschny, V., Grugel, L. & Walger, M. Use of intonation contours for speech recognition in noise by cochlear implant recipients. *J. Acoust. Soc. Am.* **129**, EL204–EL209 (2011).
59. Micula, A. et al. A glimpse of memory through the eyes: Pupillary responses measured during encoding reflect the likelihood of subsequent memory recall in an Auditory Free Recall Test. *Trends Hear.* **26**, (2022).
60. Micula, A., Rönnberg, J., Zhang, Y. & Ng, E. H. N. A decrease in physiological arousal accompanied by stable behavioral performance reflects task habituation. *Front. Neurosci.* **16**, 876807 (2022).
61. Buhusi, C. V. & Meck, W. H. What makes us tick? Functional and neural mechanisms of interval timing. *Nat. Rev. Neurosci.* 6, 755–765 (2005). (2005).
62. Nittrouer, S., Caldwell-Tarr, A. & Lowenstein, J. H. Working memory in children with cochlear implants: Problems are in storage, not processing. *Int. J. Pediatr. Otorhinolaryngol.* **77**, 1886–1898 (2013).
63. Baddeley, A. D. Short-term memory for word sequences as a function of Acoustic, semantic and formal similarity (1966). https://doi.org/10.1080/14640746608400055 18, 362–365
64. Conrad, R. & Hull, A. J. Information, acoustic confusion and memory span. *Br. J. Psychol.* **55**, 429–432 (1964).
65. Salamé, P. & Baddeley, A. Phonological factors in STM: Similarity and the unattended speech effect. *Bull. Psychon Soc.* **24**, 263–265 (1986).
66. Hopstaken, J. F., van der Linden, D., Bakker, A. B. & Kompier, M. A. J. The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics. *Biol. Psychol.* **110**, 100–106 (2015).
67. Ayasse, N. D. & Wingfield, A. Anticipatory baseline pupil diameter is sensitive to differences in hearing thresholds. *Front. Psychol.* **10**, 504013 (2020).
68. Seropian, L. et al. Comparing methods of analysis in pupillometry: application to the assessment of listening effort in hearing-impaired patients (2017). https://doi.org/10.1016/j.heliyon.2022.e09631
69. Herrmann, B. & Johnsrude, I. S. A model of listening engagement (MoLE). *Hear. Res.* **397**, 108016 (2020).
70. Carolan, P. J., Heinrich, A., Munro, K. J. & Millman, R. E. Quantifying the effects of Motivation on listening effort: a systematic review and Meta-analysis. https://doi.org/10.1177/23312165211059982
71. He, A., Deroche, M. L., Doong, J., Jiradejvong, P. & Limb, C. J. Mandarin tone identification in cochlear implant users using exaggerated pitch contours. *Otol Neurotol.* **37**, 324–331 (2016).
72. Meyer, M., Steinhauer, K., Alter, K. & Friederici, A. D. Von Cramon, D. Y. brain activity varies with modulation of dynamic pitch variance in sentence melody. *Brain Lang.* **89**, 277–289 (2004).
73. Ilse Lehiste. *Some Acoustic Characteristics of Dysarthric Speech* (S. Karger, 1965).
74. Fournier, J. E. *Audiométrie vocale: les épreuves d'intelligibilité et leurs applications au diagnostic, à l'expertise et à la correction prothétique des surdités*Maloine,. (1951).
75. Kawahara, H. & Morise, M. Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *Sadhana - Acad. Proc. Eng. Sci.* **36**, 713–727 (2011).

76. Zhang, Y., Malaval, F., Lehmann, A. & Deroche, M. L. D. Luminance effects on pupil dilation in speech-in-noise recognition. *PLoS One*. **17**, e0278506 (2022).
77. Winn, M. B., Wendt, D., Koelewijn, T. & Kuchinsky, S. E. Best practices and advice for using pupillometry to measure listening effort: an introduction for those who want to get started. *Trends Hear.* **22**, (2018).
78. Klingner, J., Kumar, R. & Hanrahan, P. Measuring the task-evoked pupillary response with a remote eye tracker. *Eye Track. Res. Appl. Symp.* **69-72**. https://doi.org/10.1145/1344471.1344489 (2008).
79. Winn, M. B. & Teece, K. H. Listening effort is not the same as speech intelligibility score. *Trends Hear.* **25**, (2021).
80. Winn, M. B. & Teece, K. H. Effortful listening despite correct responses: the cost of mental repair in sentence recognition by listeners with cochlear implants. *J. Speech Lang. Hear. Res.* **65**, 10 (2022).
81. Lew, E. et al. Differences between French and English in the use of suprasegmental cues for the short-term recall of word lists. *J. Speech Lang. Hear. Res.* https://doi.org/10.1044/2024_JSLHR-23-00655 (2024).

## Acknowledgements

We are grateful to all participants for their time and effort. We thank Florian Malaval for assistance with running the experiment with native French-speaking participants.

## Author contributions

YZ, AL, MD designed the experiment. YZ, AS, AD and MD acquired data and managed participants' appointments and follow-ups around the study. YZ and MD created scripts and software used in the experiments. YZ, AS and MD analyzed the data. YZ, AS, and MD wrote the manuscript. AL provided critical revisions to the submitted manuscript. All authors contributed to the article and approved the submitted version.

## Declarations

## Competing interests

This research was supported by a grant from the Quebec government (Mitacs Accelerate https://www.mitacs.ca/en/programs/accelerate) in collaboration with an industrial partner Oticon Medical Canada (https://www.oticonmedical.com/) [grant number IT10517]. The funding was issued to Dr. Alexandre Lehmann and Dr. Mickael Deroche, for the postdoctoral work of Dr. Yue Zhang. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The rest of authors declare that no competing interests exist.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-73320-z.

**Correspondence** and requests for materials should be addressed to Y.Z.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.