Say It Like You Mean It: Using Spoken Responses to Estimate Learning Success

E.C.C. (Noor) Velthuizen

s4550293

Department of Psychology, University of Groningen

PSB3N-BT15: Bachelor Thesis

Group number: 19

Supervisor: Dr. A. (Tassos) Sarampalis

Second evaluator: Dr. O.C. (Olaf) Dimigen

In collaboration with: Yordanka Daskalova, Edith Fernández Amurrio, Aleksandra Jarczynska, Anne Lukassen, and Thijs Peters

June 26th, 2025

A thesis is an aptitude test for students. The approval of the thesis is proof that the student has sufficient research and reporting skills to graduate, but does not guarantee the quality of the research and the results of the research as such, and the thesis is therefore not necessarily suitable to be used as an academic source to refer to. If you would like to know more about the research discussed in this thesis and any publications based on it, to which you could refer, please contact the supervisor mentioned.

Abstract

Adaptive learning systems rely on behavioural data such as accuracy scores and reaction times to estimate item-level memory strength. Previous research has shown that prosodic speech features can improve model estimations in adaptive learning systems beyond measures of accuracy and reaction times. The present study investigated whether prosodic features (intensity, speaking speed, average pitch and pitch change) can serve as indirect indicators of memory strength and subjective confidence during a sentence-based language learning task. Forty-eight native Dutch speakers studied and verbally retrieved subject-verb sentences in both Dutch and Italian. After each trial, participants rated their level of confidence about the response. Correlational analyses revealed that speaking speed showed the strongest associations with both accuracy and confidence, whereas intensity (loudness) was primarily associated with accuracy alone. The strength of these directions varied depending on whether participants responded in their native Dutch or in the unfamiliar Italian language. Low confident responses were associated with a rising pitch contour at the end of the utterance, a pattern also observed in incorrect responses. This same pattern was found for the incorrect responses. Notably, these findings were only observed in the native Dutch condition. The findings of this study suggest that prosodic speech features reflect both cognitive (memory strength) and metacognitive (subjective confidence) processes in language learning. This highlights the potential of incorporating real-time prosodic analysis into adaptive learning systems to improve the accuracy of memory strength estimations.

Keywords: adaptive learning system, confidence, memory strength, native language, non-native language, prosody, speech features

Introduction

When we speak, we not only rely on the words we choose to convey our message; the way in which we articulate them also plays an important role. This manner of speaking is shaped by suprasegmental features such as rhythm, intonation, loudness and timing, which are collectively referred to as prosody. Prosodic features of speech can convey emotional states, uncertainty and emphasis (Behrman & Finan, 2023; Cutler et al., 1997). Intonation, for instance, is reflected in the pitch contour of an utterance (Behrman & Finan, 2023). For example, we often raise our pitch at the end of a question, as in: “Do you know when the Beatles released *Let It Be*?”. Similarly, when uncertain, speakers tend to use a rising intonation at the end of the utterance. Imagine answering the previous question with uncertainty: “Maybe… late 60s?”. This suggests that prosodic features, such as the pitch contour, can serve as indicators of a speaker's confidence.

Goupil & Aucouturier (2021) examined the relationship between prosody and speaker confidence in a visual perception task. Participants fixated on a cross while a pseudo-word briefly appeared at the bottom of a screen. After a masked presentation they verbally chose the correct word from two similar alternatives and rated their confidence on a 1–4 scale. Analysis of the verbal responses showed that high-confident answers typically showed a rise-and-fall pattern of pitch, while doubtful responses displayed the opposite fall-and-rise pattern. Moreover, confident responses were longer in duration and higher in intensity. Loudness was found to best reflect accuracy of the response, with correct answers being louder on average. These findings align with earlier research by Kimble & Seidel (1991), who demonstrated that more confident answers to trivia questions were louder and had a faster reaction time. These studies indicate that prosody plays an important role in signalling memory retrieval and expressing confidence.

Adaptive learning systems optimize learning schedules by estimating how well the learner has memorized a certain item, reflecting the memory strength. The system determines what items need more practice than others (Lindsey et al., 2014; Pavlik & Anderson, 2008). Memory strength is a latent variable that cannot be observed directly but is inferred from measurable behaviours, with accuracy scores being the most straightforward measure. A correct response suggests that less repetition of the item is necessary, whereas an incorrect response indicates the need for further revision (Lindsey et al., 2014; Pavlik & Anderson, 2008; Settles & Meeder, 2016). In addition to accuracy scores, reaction times have also been shown to provide insights into memory strength (Mettler et al., 2011; Sense & Van Rijn, 2022). It assumes an item is consolidated more strongly when the learner responds more quickly since reaction times have shown to be correlated to accuracy scores (Benjamin & Bjork, 1996). Even when given answers are correct, the reaction times provide additional information about how well the learner knows the item and therefore the system can adapt to the learner who makes little to no mistakes (Sense & Van Rijn, 2022).

Building on these insights into how memory strength can be inferred from accuracy and response time, several adaptive learning systems have been developed that apply these principles in practice. These systems rely on learning theories to structure and personalize study schedules. *StudyGo* (formerly *WRTS*), for example, quizzes vocabulary via flashcards and is built on the testing effect. This describes the finding that repeated testing results in better long-term recall (Karpicke & Roediger, 2008; ‘StudyGo Onderzoek – Effectief leren en betere cijfers’, n.d.). *Duolingo*, a widely used language-learning platform, incorporates both the testing effect and the spacing effect. The spacing effect demonstrates that spreading study sessions over time enhances retention compared to last-minute crammed study sessions (Settles & Meeder, 2016). Both *StudyGo* and *Duolingo* infer memory strength from accuracy scores only, whilst the *SlimStampen* system also uses reaction times to gain more information (van Rijn et al., 2009). Its algorithm incorporates the spacing and testing effect to optimize the learning schedule based on learners' reaction times and accuracy scores (Sense et al., 2021).

Wilschut et al. (2023) worked with this adaptive scheduling system incorporating both accuracy scores and reaction times. They investigated whether prosodic features could further improve adaptive learning systems. In a language-learning task, participants studied English translations of Swahili words through spoken rehearsal. The results showed that a falling pitch, faster speaking speed and higher intensity significantly predicted correct answers. Conversely, rising pitch, slower speaking speed and lower loudness were associated with incorrect responses. This suggests that prosodic information provides predictive value beyond reaction time and accuracy alone. The findings imply that real-time analysis of prosodic features could improve adaptive learning systems by identifying items that need rehearsal. However, it is not clear from the results *why* specific prosodic patterns are associated with memory strength. Wilschut et al. (2023) proposed two explanations: prosody might directly reflect memory strength, similar to reaction times, or they might relate to the learner's subjective confidence. Since confidence was not measured in this study, the distinction could not be made.

To address this, Wilschut et al. (2025) investigated how prosody relates to both memory strength and subjective confidence in a vocabulary learning task. Participants studied Lithuanian-English words, produced spoken retrieval attempts and rated their confidence after each trial. The study found that pitch slope and speaking speed correlated with confidence ratings: a more negative pitch slope (rise-and-fall pattern) characterized higher confidence. Conversely, loudness was more strongly related to recall performance, as shown from the reaction times and accuracy scores. These findings suggest that Wilschut et al. (2025) were able to partially disentangle the roles of prosodic features in relation to memory strength and confidence. Specifically, it appears that some prosodic features (such as pitch slope) reflect subjective confidence, while others (such as loudness) reflect memory strength more directly. Against expectations, speaking speed showed a negative correlation with confidence. This contrasts with earlier research demonstrating a positive association between speaking speed and confidence (Jiang & Pell, 2017; Scherer et al., 1973).

Jiang & Pell (2017) examined perceived confidence by having participants rate utterances of sentences produced with varying intended confidence levels: confident, close-to-confident, unconfident and neutral. Perceived confidence ratings increased with intended confidence and acoustic analyses revealed that confident utterances were louder and faster, while unconfident utterances had higher average pitch and slower speaking speed. Wilschut et al. (2025) and Goupil & Aucouturier (2021) both found the direction going the other way: speaking speed correlated negatively with confidence. Wilschut et al. (2025) suggests that these differences in findings may be explained by the type of material used across studies. Both Wilschut et al. (2025) and Goupil and Aucouturier (2021) used single-word items, whereas studies showing a positive relationship between speaking speed and confidence used longer utterances with (multiple) sentences (Jiang & Pell, 2017; Scherer et al., 1973). Wilschut et al. (2025) expected that multiple-word trials may allow for pausing and pacing within an utterance, which possibly stabilizes the speaking speed as a confidence indicator.

The present study aims to extend this line of research by investigating how prosody relates to memory strength (measured via accuracy scores and reaction times) and subjective confidence during a subject-verb sentence learning task. Native Dutch speakers will study Italian subject-verb sentences in both directions (Italian-to-Dutch and Dutch-to-Italian) by verbally producing the translation. After each trial, participants will rate their confidence in the accuracy of their response. Additionally, this study examines whether the relationships between prosodic features, memory strength and confidence differ depending on whether participants speak in their native or non-native language. Understanding these patterns is crucial for identifying reliable prosodic indicators to improve adaptive learning systems in language learning contexts.

Method

Participants

In total, 48 Dutch native speakers (*M* = 19.9 years, *SD* = 1.6) took part in the experiment. Most participants were first-year Psychology students who received course credit for their contribution to the study. Others were not rewarded for their involvement. Participants' ages ranged from 18 to 25 years, with 70.8% identifying themselves as women and 29.2% as men. All participants indicated that they had no hearing or speaking problems and had never taken lessons in Italian. The study was approved by the Ethics Committee of the Faculty of Behavioural and Social Sciences at the University of Groningen (study code: PSY-2223-S-0257).

Materials

Forty subject-verb sentences were created collaboratively by the experimenters, each with corresponding Dutch and Italian translations. Four sentences were used in the practice block and thirty sentences were used in the experimental learning blocks. The sentences were randomly assigned to two sets. Participants received each of the set of sentences either in the Italian to Dutch condition or in the Dutch to Italian condition. The sets and conditions were counterbalanced across participants. This resulted eventually in four possible combinations of direction and set, thereby controlling for potential differences in sentence difficulty. The text-to-audio generator *Narakeet* (*Narakeet*, n.d.) was used to provide the participants with the right pronunciation of the sentences.

Software and Hardware

The experiment was built with JavaScript and HTML5 using the *jsPsych* online experiment library (De Leeuw, 2015). Participants completed the experiments in quiet individual cubicles, where the program was presented on a monitor. They wore *Nedis Xyawyon GHST100BK* over-ear headphones with built-in microphones. Responses were recorded and transcribed in real time to text using the Google Web Speech API.

Design and Procedure

Participants signed consent forms for the use of their data and audio recordings. In the background questionnaire that followed they reported their age, gender, first language, whether they had experienced any hearing or speaking difficulties and whether they had followed lessons in any language other than Dutch. A randomly generated participant ID was assigned to each participant to ensure anonymity of the data.

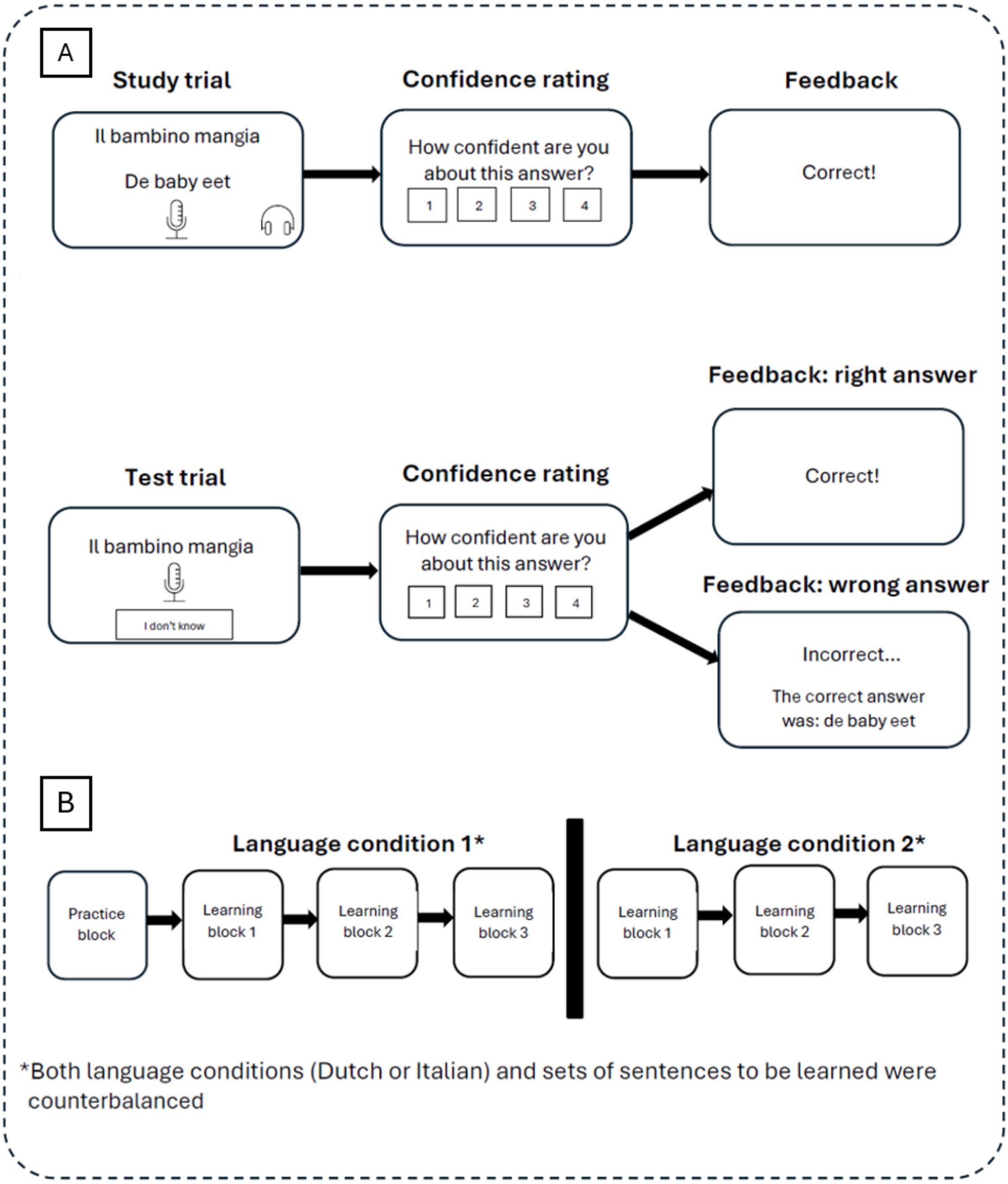
Participants were instructed to study subject-verb sentences and verbally provide the translation in either Dutch or Italian (depending on the condition they were in at that moment). After each trial (i.e., one sentence), they rated their subjective confidence on a 4-point Likert scale (1 = not confident, 2 = slightly confident, 3 = moderately confident, 4 = confident). Feedback was then provided indicating whether the answer was correct or incorrect. Responses with a Levenshtein's edit distance of two or lower were marked as correct and if the API registered the answer as incorrect, the correct translation was shown. Reaction time was defined as the time elapsed between item presentation and the onset of speech.

During the study trial, which was defined as the first time a new subject-verb sentence was presented, participants read and listened to the translation and were asked to repeat it out loud (Figure 1A). After studying five new sentences, participants completed test trials in which they retrieved the correct translation. They could skip a trial by clicking the 'I don't know' button, which they were instructed to use exclusively when they genuinely had no idea what the correct answer was. Each sentence was tested four times within a block.

The full experiment consisted of seven blocks (Figure 1B). The first block was a practice block with four sentences to familiarize participants with the task. Participants were instructed to ask questions or report errors after this practice phase. The actual experiment then began, consisting of three blocks in one language direction (either Dutch-to-Italian or Italian-to-Dutch as the first condition to complete), followed by a break. After the break, participants completed the remaining three blocks in the opposite language direction. This within-subject design made participants complete both language conditions. The experiment followed a fixed schedule: all participants were presented with the same sentences in the same order and the system did not adapt to their performance.

Figure 1

Overview of the Experimental Design and Structure of the Study and Test trials



Note. Panel A shows the study and test trials. In the study trials, participants repeated the correct translation and rated their confidence. In the test trials. Participants retrieved the translation and subsequently rated their confidence. Feedback was provided in both trial types. Panel B shows the structure of the experiment, which included two language conditions. The practice block consisted of four sentences. Each language condition comprised three learning blocks, each containing five sentences to be studied and tested. The order of language conditions and sentence sets was counterbalanced across participants.

Extracting Prosodic Speech Features from Audio Files

Praat version 6.2.7 (Boersma & Weenink, 2023) was used to extract the acoustic properties from the audio recordings. This provided data for the duration of the response, the speaking speed (syllables per second), the average intensity and the average pitch for twenty equal segments separately.

Data Analysis

Further analysis was conducted in *R* version 4.4.2 (R Core Team, 2024) to calculate overall average pitch across the twenty segments and the change in pitch over the utterance, used as an estimate for pitch slope. The pitch change was calculated by subtracting the average pitch of the first five segments from that of the last five segments.

To examine the differences in dependent variables across trial types (study vs. test) and language directions (responding in Dutch vs. Italian), a new dataset was derived from the original by-trial dataset. For each participant, mean scores were calculated for each of the four combinations: Dutch study, Italian study, Dutch test and Italian test trials. This by-participant dataset included average values for accuracy (proportion correct), response time, subjective confidence, speaking speed, intensity, average pitch and change in pitch by participant. Only correct responses were included in these analyses to ensure comparability across conditions and trial types, except in the calculations for accuracy. Including incorrect responses would have confounded effects of language direction task difficulty, as accuracy differed substantially between conditions and could have indirectly influenced variables such as confidence. Using this by-participant data set, split violin plots were created with the *tidyverse* version 2.0.0 (Wickham, 2016) and *introdataviz* version 0.0.0.9003 (Nordmann et al., 2021) *R* packages. The repeated measures ANOVAs were conducted in *JASP* version 0.19.3 (JASP Team, 2024).

In addition to the comparisons across trial types and language directions, correlational analyses were conducted using the original by-trial data set to examine associations between prosodic features and behavioural outcomes. Speaking speed, intensity and pitch measures were standardized within participants to account for individual differences (e.g., in natural pitch between male and female voices). Subsequently, Pearson correlation coefficients were computed between standardized prosodic features (speaking speed, intensity, average pitch, pitch change) and behavioural variables (accuracy, reaction times, confidence ratings). Correlation analyses and visualisations were performed in *R* using the *Hmisc* version 5.2.3 (Harrell Jr, 2003) and *corrplot* version 0.95 (Wei & Simko, 2010) packages. Finally, standardized pitch dynamics were further analysed by plotting it across the twenty segments using the *tidyverse* and *ggplot2* version 3.5.2 (Wickham et al., 2007) packages in *R*.

Results

Differences in Behavioural Variables and Prosodic Speech Features Across Conditions and Trial types

To gain an initial understanding of how behavioural variables (accuracy, reaction time and confidence) and prosodic features differed across conditions and trial types, we first averaged the trials per participant for each of the seven dependent variables. Only correct responses on test trials were included in the averages, with the exception of accuracy, which incorporated both correct and incorrect responses. The seven variables included accuracy, reaction time, confidence, speaking speed, intensity, average pitch and pitch change. Pitch change for a trial was calculated by subtracting the average pitch of the first five segments (out of a total of twenty) from that of the last five segments of the utterance. We created split violin plots using the means per participant to visualize the distributions (Figure 2). The split violin plots present the data organised by language (answering in Dutch vs. Italian) and trial type (study vs. test). To evaluate whether the observed differences between languages and trial types were statistically significant, repeated measures ANOVAs were conducted.

Differences Across Conditions and Trial types

Accuracy (proportion correct) (Figure 2A) was highest for the Dutch study trials (*M* = 0.92, *SD* = 0.09), as expected given that Dutch was the native language of all participants. In addition, the study trials provided participants with both the correct answer and its pronunciation, making the task relatively straightforward. Accuracy was lower in the Italian study trials (*M* = 0.75, *SD* = 0.15), likely due to participants' unfamiliarity with the Italian language and possible challenges in reproducing the correct pronunciation. A similar pattern was observed in the Italian test trials, which showed lowest overall accuracy (*M* = 0.24, *SD* = 0.15). In contrast, performance in the Dutch test trials remained relatively high (*M* = 0.81, *SD* = 0.13), which may reflect participants' unfamiliarity with the Italian language and the increased difficulty of reproducing unfamiliar words. The repeated measures ANOVA revealed significant main effects of both language ($F(1, 47) = 595.28, p < .001$) and trial type ($F(1, 47) = 315.37, p < .001$) on accuracy, as well as a significant interaction between the two ($F\ (1, 47) = 235.56, p < .001$).

Reaction times (in milliseconds) (Figure 2B) were fastest for the Dutch test trials (*M* = 1,758.14, *SD* = 323.29) and slowest for the Italian study trials (*M* = 2,869.27, *SD* = 616.17). One possible explanation for the slower reaction times in study trials is that participants encountered novel items that required more initial processing. Test trials, however, involved repeated retrieval attempts, which may have reduced reaction times over time, resulting in a faster average reaction time. The standard deviations for the Italian study, Dutch study and Italian test trials were relatively large, indicating substantial variability in reaction times across participants. The ANOVA showed significant main effects of language ($F(1, 47) =$

204.68, *p* < .001) and trial type (*F*(1, 47) = 52.41, *p* < .001), as well as a significant interaction effect (*F* (1, 47) = 37.93, *p* < .001).

Subjective confidence ratings (Figure 2C) followed a similar pattern to accuracy: confidence was highest during the Dutch study trials (*M* = 3.83, *SD* = 0.26) and lowest for the Italian test trials (*M* = 2.50, *SD* = 0.73). Confidence was rated on a 1 – 4 Likert scale (1 = not confident, 4 = confident). The results suggest that participants were aware of their decreased performance when retrieving Italian items. The relatively high spread of confidence ratings observed in the Italian study and test trial reflects the individual differences in how confident participants felt when responding in an unfamiliar language. The ANOVA indicated significant main effects of language (*F*(1, 47) = 204.68, *p* < .001) and trial type (*F*(1, 47) = 11.49, *p* = .001), as well as a significant interaction (*F*(1, 47) = 6.19, *p* = .020).

Speaking speed (in number of syllables per second) (Figure 2D) was lowest in the Dutch study trials (*M* = 1.36, *SD* = 0.24) and highest in the Italian study trials (*M* = 1.71, *SD* = 0.30). A possible explanation for this pattern is that participants may have attempted to reproduce the Italian audio examples, as the language was unfamiliar to them. In contrast, participants in the Dutch trials may have relied more on their personal and natural speaking speed leading to a slower average speaking speed. The repeated measures ANOVA revealed a significant main effect of language (*F*(1, 47) = 64.09, *p* < .001), but the main effect of trial type did not reach significance. The interaction effect was significant (*F*(1, 47) = 16.76, *p* < .001).

Intensity (in dB) (Figure 2E) remained relatively stable across all conditions. The mean intensity ranged from 69.04 dB (*SD* = 1.83) in the Dutch study trials to 70.15 dB (*SD* = 2.97) in the Italian testing trials. This overall consistency may reflect the participants' tendency to remain a relatively constant vocal intensity independent of language or trial type.

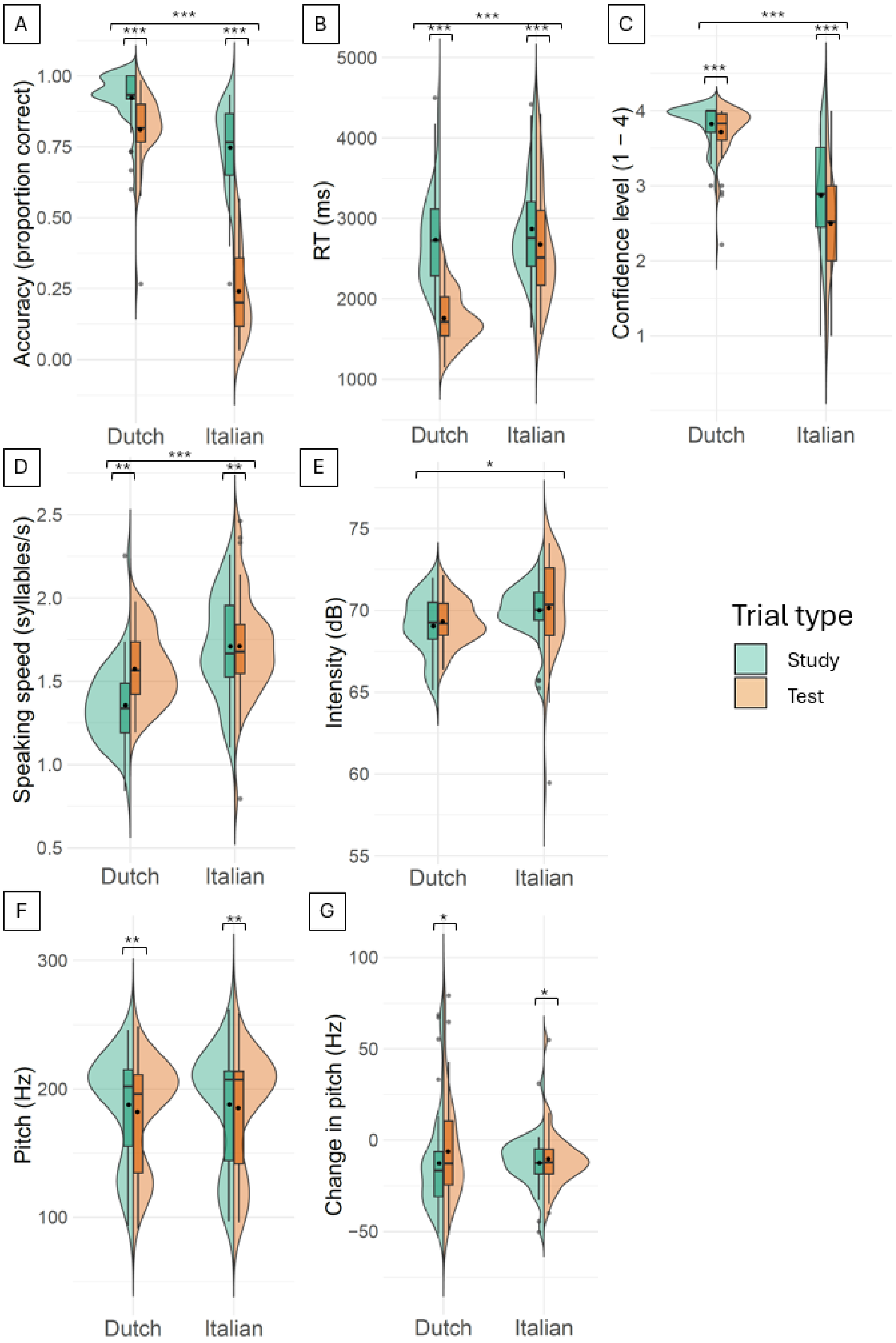
The ANOVA showed a significant main effect of language (*F*(1, 38) = 5.09, *p* = .030), but no significant effect of trial type, nor a significant interaction effect.

Average pitch (in Hz) (Figure 2F) shows the same pattern as intensity and remained relatively stable across all conditions. The group means for average pitch ranged from 182.11 Hz (*SD* = 40.91) in the Dutch testing trials to 187.78 Hz (*SD* = 45.23) in the Italian study trials. Again, this overall consistency may reflect the participants' tendency to remain a relatively constant average pitch independent of language or trial type. The repeated measures ANOVA only showed a significant effect of trial type (*F*(1, 38) = 8.04, *p* = .007).

Pitch change (in Hz) (Figure 2G) over the utterances shows substantial variability across conditions, with high standard deviations observed in all trial types. Standard deviations ranged from 14.08 Hz in the Italian study trials up to 28.36 Hz in the Dutch study trials. This variability limits the strength of any conclusions based on group means alone. In line with the results of the ANOVA for the average pitch, only the main effect of trial type was significant (*F*(1, 38) = 4.19, *p* = .048).

Figure 2

Split Violin Plots for each Dependent Variable by Condition and Trial Type



Note. Split violin plots were computed using data from correct answers on test trials only, except for accuracy. Accuracy plots were computed using both incorrect and correct answers on test trials. Graphs show boxplots and their kernel density estimates of the distribution. Violin shapes may slightly extend beyond the theoretical bounds for accuracy (0 – 1) and confidence (1 – 4), due to smoothing. The figure shows the significant effects of language or trial type (*$p < .05$; **$p < .01$, ***$p < .001$).

Correlations between Accuracy, RT, Confidence and Prosody Speech Features

To account for differences in voice characteristics between participants, prosodic features were standardized within participants, resulting in z-scores. Pearson correlation coefficients were then calculated on the trial level for reaction time, accuracy, confidence, standardized speaking speed, standardized intensity, standardized average pitch, and standardized pitch change including test trials only to investigate the relationship between prosody and memory strength. Throughout this section, references to prosodic features (e.g., pitch, intensity) refer to their standardized values. Correlation matrices were plotted separately for the responses on test trials in Dutch and Italian conditions to analyse potential language-specific patterns (Figure 3).

Correlations in the Dutch Condition

The left panel in Figure 3 shows the matrix with correlations between the dependent variables for the Italian to Dutch test trials. Accuracy correlated negatively with reaction times ($r = -.36$, $p < .001$), suggesting participants were more accurate when they responded more quickly. Accuracy correlated positively with confidence ($r = .49$, $p < .001$), which suggests that participants were aware of their performance. More confident responses were also associated with faster reaction times ($r = -.46$, $p < .001$).

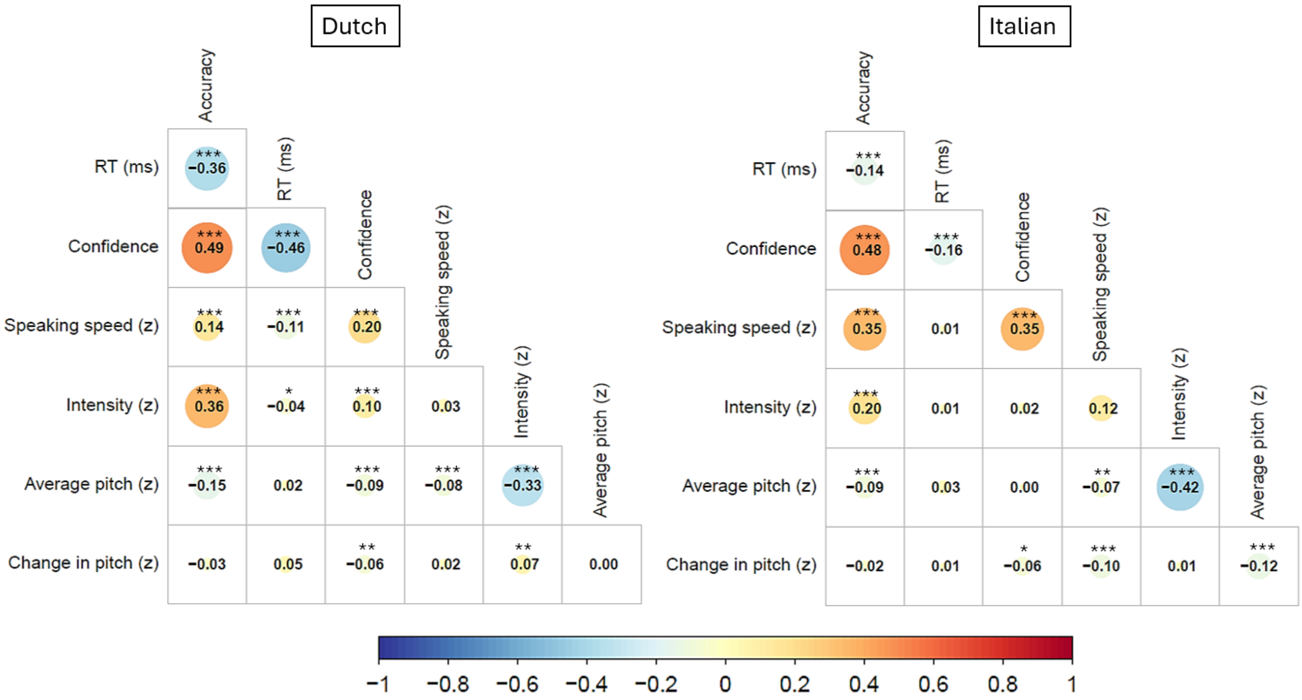
More accurate responses were associated with a faster speaking speed ($r = .14$, $p < .001$) and higher intensity ($r = .36$, $p < .001$). Average pitch correlated negatively with accuracy ($r = -.15$, $p < .001$). Change in pitch did not correlate significantly with accuracy.

Reaction times correlated negatively with standardized speaking speed ($r = -.11$, $p < .001$) and standardized intensity ($r = -.04$, $p = .039$) and positively with standardized change in pitch ($r = .05$, $p = .019$). The correlation between reaction times and average pitch was non-significant.

Subjective confidence of the participants showed positive correlations with standardized speaking speed (*r* = .20, *p* < .001), standardized intensity (*r* = .10, *p* < .001) and negative correlations with standardized average pitch (*r* = -.09, *p* < .001) and standardized change in pitch (*r* = -.06, *p* = .005).

Figure 3

Correlation Matrices with Pearson Correlation Coefficients for the Responses on Test Trials in Dutch (left panel) and in Italian (right panel)



Note. **p* < .05; ***p* < .01; ****p* < .001.

Correlations in the Italian Condition

The right panel of Figure 3 shows the matrix with correlations between the dependent variables for the Dutch to Italian test trials. Accuracy correlated negatively with reaction times (*r* = -.14, *p* < .001) and positively with confidence (*r* = .48, *p* < .001). Participants were on average faster and more confident when they were correct. More confident responses were also associated with shorter reaction times (*r* = -.16, *p* < .001).

More accurate responses were associated with a higher speaking speed (r = .35, p < .001) and higher intensity (r = .20, p < .001). Accuracy also correlated with standardized average pitch (r = -.09, p < .001), but not with the standardized change in pitch.

Confidence correlated only significantly with standardized change in pitch (r = -.06, p = .014), and not with any other prosodic features. Reaction times showed no significant correlations with any of the prosodic features in the Italian condition.

Comparing Correlations between Conditions

Reaction times correlated more strongly with confidence in the Dutch condition (r = -.46, p < .001) compared to the Italian condition (r = -.16, p < .001). A similar pattern was observed for the correlation between reaction times and accuracy, which was again stronger in Dutch (r = -.36, p < .001) than in Italian (r = -.14, p < .001). The correlations between confidence and accuracy are similar for Dutch (r = .49, p < .001) and Italian (r = .48, p < .001).

Among the prosodic features, speaking speed correlated significantly with reaction times in the Dutch condition (r = -.11, p < .001) in the Dutch condition, but not in the Italian condition (r = .01, p = .680). In contrast, standardized speaking speed was more strongly associated with confidence in the Italian condition (r = .35, p < .001) than in the Dutch condition (r = .20, p < .001). A similar pattern was observed for the relationship between speaking speed and accuracy, which was stronger in Italian (r = .35, p < .001) than in Dutch (r = .14, p < .001).

In the Dutch condition, standardized intensity showed higher correlations with both confidence (r = 0.10, p < .001) and accuracy (r = .36, p < .001) compared to the Italian conditions (r = .02, p = .437; r = .20, p < .001, respectively).

In the Italian condition, average pitch correlated negatively with accuracy (*r* = -.09, *p* < .001). In the Dutch condition, average pitch was also negatively correlated with accuracy (*r* = -.15, *p* < .001) and with confidence (*r* = -.09, *p* < .001).

Change in pitch correlated significantly with confidence in both the Dutch condition (*r* = -.06, *p* = .005) and the Italian condition (*r* = -.06, *p* = .014) with similar Pearson correlation coefficients. In the Dutch condition, change in pitch correlated additionally with reaction times (*r* = .05, *p* = .019), whilst this relationship was non-significant in the Italian condition (*r* = .01, *p* = .750).

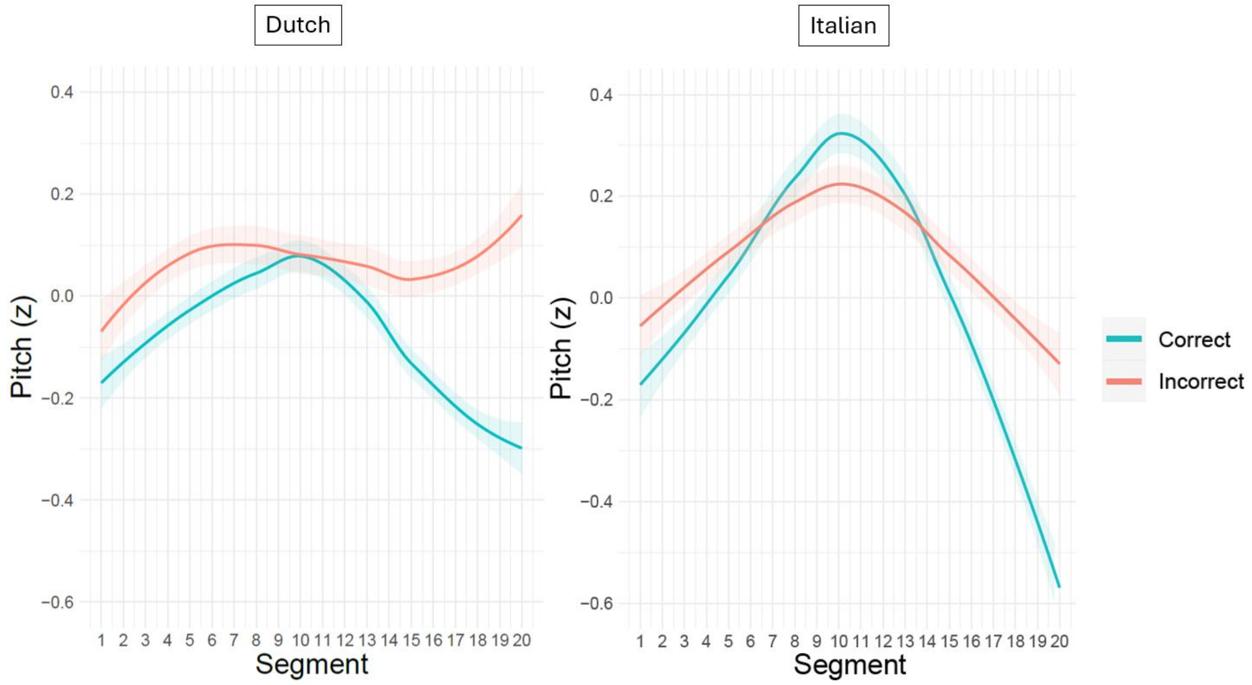
Pitch (z) Over Segments as Function of Language Direction, Confidence Levels and Accuracy

To more precisely examine the trajectory of pitch over utterances, the average standardized pitch was plotted across twenty segments, as a function of response accuracy (Figure 4) and accuracy (Figure 5) for both languages. As in the previous section, standardized pitch will be referred to as "pitch" for ease of reading.

Pitch shows a rise-and-fall pattern with the exception for the incorrect responses in Dutch (Figure 4), where pitch rises toward the end of the utterances. Although this effect is less pronounced in the Italian responses, incorrect responses still end with a higher pitch than correct responses.

Figure 4

Pitch (z) Across Segments: Dutch (left) vs. Italian (right), as a Function of Accuracy



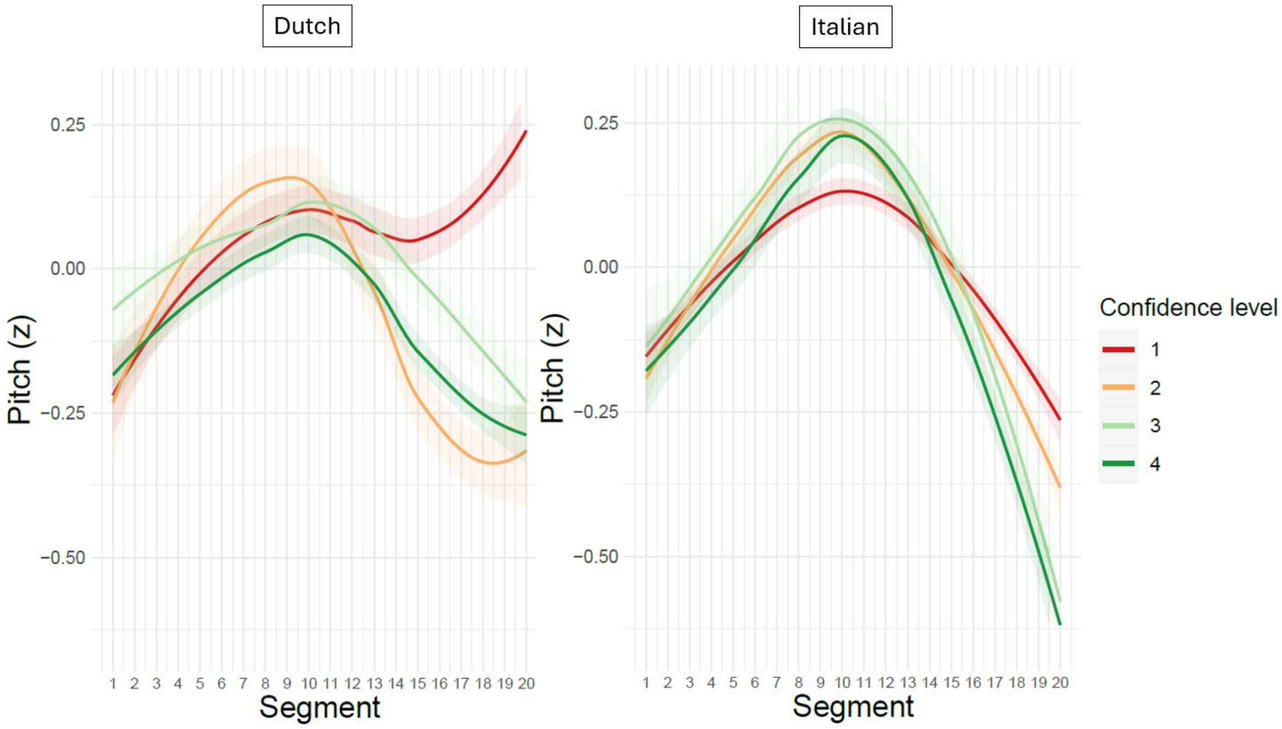
Note. Plotted data included testing trials from the learning phase only. Shaded areas represent 95% confidence intervals around LOESS-smoothed mean pitch (z) estimates, plotted for correct and incorrect responses separately. In the Dutch condition, 2,384 (81.1%) trials were marked as correct and 556 (18.9%) as incorrect. In the Italian condition, 725 (24.6%) trials were marked as correct and 2,220 (75.4%) as incorrect.

Figure 5 displays pitch (z) across segments for each confidence level. In the Dutch condition, pitch rises over the final segments for responses with the lowest confidence level. The same effect was previously described for the incorrect Dutch responses. Confidence level 2 also shows a slight rise in pitch at the end of the utterances in Dutch.

In the Italian condition, only rise-and-fall patterns can be observed. Notably, responses with the lowest confidence levels end with the highest pitch, whereas those with the highest confidence levels end with the lowest pitch.

Figure 5

Standardized Pitch Across Segments: Dutch (left) vs. Italian (right), Confidence Levels (1 – 4)



Note. Standardized pitch for Dutch responses (left panel) and Italian responses (right panel). Plotted data included testing trials from the learning phase only. Shaded areas represent 95% confidence intervals around LOESS-smoothed mean pitch (z) estimates, plotted separately for each level of confidence ranging from 1 (not confident) to 4 (confident). Number of trials per confidence level (1–4): Dutch = 135 (4.9%), 137 (5.0%), 424 (15.3%), 2,071 (74.9%); Italian = 834 (40.0%), 626 (30.1%), 436 (20.9%), 187 (9.0%).

Discussion

The main aim of this study was to investigate how prosodic speech features relate to markers of memory strength (accuracy and reaction time) and subjective confidence, and whether these relationships differ when responding in the native or non-native language. Confidence and accuracy were consistently associated across both language conditions: responses with higher confidence ratings were generally more likely to be correct. In the Italian condition, speaking speed showed the strongest correlations with accuracy and confidence amongst all prosodic features. These correlations were of moderate strength and suggest that speaking speed can be a potential indicator of strength of the memory trace. In the Dutch condition, intensity showed the strongest correlation with accuracy. Speaking speed was most strongly related to confidence, although this correlation was of a weaker magnitude compared to the relationship in the Italian condition. Overall, speaking speed appears to have potential in being a prosodic indicator of both accuracy and confidence, while intensity may provide additional information about the accuracy of the response.

Wilschut et al. (2025) found, contrary to their hypothesis and earlier findings (Jiang & Pell, 2017; Scherer et al., 1973), that more confident responses were spoken with a slower speaking speed. This finding aligned with the results of Goupil & Aucouturier (2021), and Wilschut et al. (2025) proposed that the discrepancy might be due to differences in the types of materials used. Specifically, both Wilschut et al. (2025) and Goupil & Aucouturier (2021) used word-level items, while Jiang & Pell (2017) and Scherer et al. (1973) used sentence-level items. The present study supports this explanation: using sentence-level items, we observed a positive relationship between speaking speed and confidence. According to Wilschut et al. (2025), sentence-level responses allow for more natural variation in pausing and pacing, which may make speaking speed a more reliable indicator of confidence. Isolated single-words allow little possibilities for pausing or pacing and speakers possibly articulate single-words more intentionally when confident, resulting in slower speech. These findings offer valuable guidance for the development of speech-based adaptive learning systems, as they clarify when and how speaking speed can reliably reflect learner confidence.

An additional methodological difference in our study and that of Wilschut et al. (2025) concerns the confidence rating scale. While we employed a four-point Likert scale that required participants to make an explicit choice before proceeding, Wilschut et al. (2025) used a continuous slider, which was by default presented in the centre of the scale. This may have resulted in participants leaving the slider untouched, particularly as repetition induced task fatigue, resulting in a clustering of responses around the centre. Indeed, their density plot illustrates this central concentration. In contrast, our confidence data shows a more widely distributed density (Figure 2C), suggesting that participants made more thoughtful and differentiated confidence ratings. Notably, in the Dutch condition there appears to be a ceiling effect, with a high density of responses for the higher confidence levels. This is likely due to the fact that participants were responding in their native language. These well-informed confidence ratings likely contributed to the stronger and more consistent correlations we observed between confidence and prosodic features, such as speaking speed. Based on these findings, we recommend future studies to consider using a Likert scale rather than a continuous slider to obtain more informative confidence ratings, and thereby improve the reliability of correlations with other dependent variables.

Yet, even with more informative confidence ratings, stronger correlations were not observed across all dependent variables. One exception was pitch slope, which was more strongly associated with confidence in Wilschut et al. (2025) than the change in pitch did in our study. This could possibly have to do with the way pitch slope was computed. They calculated the pitch slope by using a least squares linear regression on the final eight pitch segments of the utterances (out of a total of 15 segments), while our study used a simplified estimate of the pitch slope. We calculated the change in pitch over the utterance by subtracting the average pitch of the first five segments from that of the last five segments (out of a total of 20 segments). This method may have oversimplified the pitch contour and failed to capture the nuances in the patterns. After realizing this limitation, we conducted a more detailed follow-up analysis by plotting standardized pitch across all segments, separately for correct vs. incorrect responses and for different confidence levels. Visual inspection of the plots revealed that the simplified estimate discarded meaningful variation across the utterance, particularly in the middle segments where pitch followed a parabolic curve. We therefore recommend that future studies use more refined methods that capture the pitch contour more accurately. For instance, by applying a least squares linear regression to the final portion of the pitch segments, as done by Wilschut et al. (2025). This method is relatively simple to implement and highlights the variation of pitch near the end of the utterance, which is typically the most informative part.

A key insight from the standardized pitch across segments plots was that utterances rated with low confidence in the Dutch condition ended with a higher pitch, on average. This rising pitch trajectory was not only associated with lower confidence but also with incorrect responses. This suggests that pitch contour may serve as an indicator for subjective confidence and accuracy in the native language. Low-confidence responses were underrepresented in the Dutch condition and the results should therefore be interpreted with caution. Still, the observed pattern is consistent with previous research showing negative correlations between pitch slope and both confidence and accuracy (Goupil & Aucouturier, 2021; Wilschut et al., 2023, 2025). This alignment with earlier findings suggests that pitch slope at the end of the utterance may reflect confidence, which itself is closely related to accuracy. It indicates a possible indirect link between pitch contour and accuracy.

It is interesting to note that this pattern was not observed in the Italian condition. A possible explanation is that participants were less able to apply prosodic features effectively in an unfamiliar language. Previous research by van Maastricht (2018) suggests that language learners often experience difficulties in appropriately applying prosodic features. It is therefore likely that participants were more skilled at expressing (un)certainty through prosody in their native language than in the unfamiliar Italian language.

Together, these findings illustrate that prosodic speech features can serve as indirect markers of memory strength and subjective confidence. Features such as speaking speed, intensity and pitch slope appear to convey information about the both the accuracy of responses and the confidence with which they are given. This study contributes to our understanding of how prosody reflects cognitive states during memory retrieval in both native and non-native language contexts. It also extends previous research by analysing prosodic features not only in the native language, but also in an unfamiliar language with no prior experience. This is unlike earlier studies that focused on prosody in familiar second languages such as English (Wilschut et al., 2023, 2025) or pseudo-words (Goupil & Aucouturier, 2021). The current findings also clarify the relationship between speaking speed and confidence, which appears to be influenced by the nature of the materials used. In our study using subject-verb sentences, speaking speed was positively related to confidence, whereas previous studies using isolated words reported the opposite pattern.

From a practical perspective, these insights can help inform the development of speech-based adaptive learning systems. Real-time prosodic analysis could be used to estimate leaner's confidence and accuracy without requiring explicit rating of subjective confidence. Our findings provide support for earlier work showing that prosodic input can improve the predictive power of memory models (Wilschut et al., 2023).

Despite the contributions of this study for both theoretical understanding and practical application, several limitations should be acknowledged. First, transcription confidence from the Google Web Speech API was lower in the Dutch condition (*M* = 0.80, *range* = 0.65-0.88) than in the Italian condition (*M* = 0.96, *range* = 0.93-0.97), which may have affected feedback accuracy in the Dutch condition. This difference may reflect in how well the programme performs across languages, depending on the amount and quality of the training the system had. A challenge in scaling speech technologies to many languages, is the amount of data available for so-called low-resource languages (Bartelds et al., 2023; Zhang et al., 2023). This could possibly explain the difference between the API confidence for Dutch and Italian. No trials were excluded based on low API confidence scores, which may have affected feedback and accuracy scores, particularly in the Dutch condition. Several participants reported that they had been confident about their response in Dutch, but received feedback indicating it was incorrect. This suggests that transcription errors may have affected the accuracy scores. However, no unexpected patterns were observed in the overall accuracy differences between the Dutch and Italian condition. As expected, participants were most accurate in their native language, which indicates that the lower API confidence is unlikely to have significantly influenced the overall accuracy pattern. Nonetheless, future studies may improve data quality by reviewing low-confidence transcriptions and considering excluding them from analysis.

Second, performance in the Italian condition was relatively low, with 75.4% of test trials marked as incorrect and 70.1% of responses receiving a confidence rating of 1 or 2 on the 4-point scale. This suggests that the task of recalling Italian subject-verb sentences may have been too demanding for participants without prior knowledge of the language. One possible explanation for the difference in task difficulty between both conditions is that participants likely relied more on recognition in the Dutch condition, where the language was familiar. Meanwhile, the Italian condition required more active recall and production, which may have increased task difficulty and contributed to the lower accuracy scores. Another possible explanation is that the API had difficulties with transcribing the Italian responses of the participants. Non-native speakers often appear to struggle with language-specific intonation and rhythm patterns, which is crucial for intelligibility (van Maastricht, 2018). The low performance and subjective confidence in the Italian condition could have influenced the results and represents a potential confound when comparing the two conditions.

Lastly, this study only examined differences in the dependent variables between trial types and language directions for correct responses. However, incorrect responses may reveal different patterns across conditions. For completeness, future studies should also explore the differences in dependent variables using incorrect responses only to provide a more comprehensive understanding of their patterns.

Another interesting direction for future research would be to examine how a gradual increase in task complexity affects participants' accuracy scores, reaction times, confidence ratings and prosodic expression. Instead of presenting participants with full sentences from the beginning, future studies could implement a scaffolding approach in which sentences are constructed from previously learned words. This design could offer insights how accuracy scores, subjective confidence ratings and prosodic features evolve as task demands increase. Additionally, it would better reflect real-world learning contexts, where language use typically develops from simple to more complex structures.

To conclude, this study showed that prosodic speech features, particularly speaking speed, intensity and pitch slope, can serve as indirect markers of memory strength and subjective confidence across both native and unfamiliar languages. The strength of these associations depended on whether participants responded in the native Dutch or unfamiliar Italian language. These findings contribute to our understanding of how cognitive states, such as confidence, are reflected in prosody and highlight the potential of prosodic features as indicators of confidence and accuracy in speech-based adaptive learning systems. Building on these findings, future research could explore more dynamic task structures and test the effectiveness of adaptive learning systems that use real-time analysis of prosody.

References

Bartelds, M., San, N., McDonnell, B., Jurafsky, D., & Wieling, M. (2023). *Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2305.10951

Behrman, A., & Finan, D. (2023). Prosody. In *Speech and voice science* (Fourth edition). Plural Publishing, Inc.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval Fluency as a Metacognitive Index. In *Implicit Memory and Metacognition*. Psychology Press.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, *40*(2), 141–201. https://doi.org/10.1177/002383099704000203

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Goupil, L., & Aucouturier, J.-J. (2021). Distinct signatures of subjective confidence and objective accuracy in speech prosody. *Cognition*, *212*, 104661. https://doi.org/10.1016/j.cognition.2021.104661

Harrell Jr, F. E. (2003). *Hmisc: Harrell Miscellaneous* (Version 5.2.3) [Computer software]. https://CRAN.R-project.org/package=Hmisc

JASP Team. (2024). *JASP* (Version 0.19.3) [Computer software]. https://jasp-stats.org/

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, *88*, 106–126. https://doi.org/10.1016/j.specom.2017.01.011

Karpicke, J. D., & Roediger, H. L. (2008). The Critical Importance of Retrieval for Learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kimble, C. E., & Seidel, S. D. (1991). Vocal signs of confidence. *Journal of Nonverbal Behavior*, *15*(2), 99–105. https://doi.org/10.1007/BF00998265

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving Students' Long-Term Knowledge Retention Through Personalized Review. *Psychological Science*, *25*(3), 639–647. https://doi.org/10.1177/0956797613504302

Mettler, E., Massey, C. M., & Kellman, P. J. (2011). *Improving Adaptive Learning Technology through the Use of Response Times*.

Narakeet. (n.d.). [Computer software]. https://www.narakeet.com/app/text-to-audio/?projectId=257f6179-9524-4d3b-a8d6-d2d3c217dd02

Nordmann, E., McAleer, P., Tovio, W., Paterson, H., & Lisa, D. (2021). *Data visualisation using R, for researchers who don't use R* (Version 0.0.0.9003) [Computer software]. https://psyteachr.github.io/introdataviz

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*(2), 101–117. https://doi.org/10.1037/1076-898X.14.2.101

Peter Boersma & David Weenink. (2023). *Praat: Doing phonetics by computer* (Version 6.2.7) [Computer software]. http://www.praat.org/

R Core Team. (2024). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Scherer, K. R., London, H., & Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*, *7*(1), 31–44. https://doi.org/10.1016/0092-6566(73)90030-5

Sense, F., & Van Rijn, H. (2022). *Optimizing Fact-Learning with a Response-Latency-Based Adaptive System*. PsyArXiv. https://doi.org/10.31234/osf.io/chpgv

Sense, F., Velde, M. van der, & Rijn, H. van. (2021). Predicting University Students' Exam Performance Using a Model-Based Adaptive Fact-Learning System. *Journal of Learning Analytics*, *8*(3), Article 3. https://doi.org/10.18608/jla.2021.6590

Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1848–1858. https://doi.org/10.18653/v1/P16-1174

StudyGo Onderzoek – Effectief leren en betere cijfers. (n.d.). *StudyGo*. Retrieved 22 May 2025, from https://studygo.com/nl/studygo-onderzoek/

van Maastricht. (2018). *Second language prosody: Intonation and rhythm in production and perception*. Tilburg University.

van Rijn, D., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving Learning Gains by Balancing Spacing and Testing Effects. *Proceedings of the 9th International Conference on Cognitive Modeling*, 108–114.

Wei, T., & Simko, V. (2010). *corrplot: Visualization of a Correlation Matrix* (Version 0.95) [Computer software]. https://CRAN.R-project.org/package=corrplot

Wickham, H. (2016). *tidyverse: Easily Install and Load the 'Tidyverse'* (Version 2.0.0) [Computer software]. https://CRAN.R-project.org/package=tidyverse

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., & Van Den Brand, T. (2007). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics* (Version 3.5.2) [Computer software]. https://CRAN.R-project.org/package=ggplot2

Wilschut, T., Sense, F., Scharenborg, O., & Van Rijn, H. (2023). Improving Adaptive Learning Models Using Prosodic Speech Features. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, & V. Dimitrova (Eds.), *Artificial Intelligence in*

Education (Vol. 13916, pp. 255–266). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-36272-9_21

Wilschut, T., Sense, F., & van Rijn, H. (2025). *Cognitive and Metacognitive Markers of Memory Retrieval Performance in Speech Prosody*.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., Meng, Z., Hu, K., Rosenberg, A., Prabhavalkar, R., Park, D. S., Haghani, P., Riesa, J., Perng, G., Soltau, H., … Wu, Y. (2023). *Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2303.01037