# Learning from Speaking: Helping Adaptive Learning Systems learn from Speech Prosody.

Yordanka Daskalova

S4778243

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group 19

Supervisor and Examiner: dr. A. Sarampalis

Second evaluator: prof. dr. O. C. Dimigen

In collaboration with: Edith Fernández Amurrio, Alexanda Jarczynska, Thijs Peters, Anne Lukassen, Noor Velthuizen.

27th of June, 2025

**Abstract:**

Speech prosody can convey meaningful information about cognitive and metacognitive states, and can be a useful tool for improving Adaptive learning systems. Previous research has shown that prosodic cues in speech can serve as predictors of confidence and memory strength. In the current study, we mainly examine 1) how speaking speed, intensity and pitch as prosodic speech features relate to accuracy and reaction time, as markers of memory strength , and subjective confidence during a language learning task, and 2) whether prosodic information differs between native and foreign language speech. Participants completed a computer task where they learnt a total of 40 simple sentences in Italian, and had to speak out loud the correct translation either in Dutch (their Native language) or Italian. The design was bidirectional, with the response language switched halfway through the experiment, requiring participants to respond in both Italian and Dutch across the separate blocks. Subsequently they reported their subjective confidence after each response. The results showed that speaking speed and intensity were correlated to accuracy and confidence more strongly than pitch/intonation were. Yet, all the three prosodic features we tested showed to be significantly correlated with both accuracy and subjective confidence. Additionally, the correlations between the prosodic features, accuracy and confidence were higher on average in the Dutch response condition, which may suggest that native language is richer in prosodic information. Our findings not only contribute to a deeper understanding of how prosodic features may serve as a predictor of memory strength and subjective confidence, but also indicate that the prosodic markers differ between utterances in native and newly acquired foreign languages, revealing potential applications for improving adaptive learning systems in the context of language learning.

*Keywords: speech, prosody, Adaptive learning systems, memory trace, language learning task.*

**Learning from Speaking: Helping Adaptive Learning Systems learn from Speech Prosody.**

Cognitive science research has demonstrated that human memory is susceptible to rapid forgetting, even under optimal learning conditions. That raises the need for strategies that can promote durable learning and retention. While many such strategies have been explored, with various degrees of success, spacing of study, i.e. spacing out learning and retrieval of the given material, was shown to be effective for better long-term knowledge preservation. This strategy can be optimal when personalized to the learner's behavior and needs (Khajah et al., 2014; Lindsey et al., 2014, van Rijn et al., 2009). MemoryLab (i.e. SlimStampen) is an adaptive fact-learning system (ALS) which is found on a similar learning strategy. It continuously adjusts the spacing of learning trials within sessions, based on each learner's estimated memory strength, aiming at optimizing retention by scheduling repetitions at the most effective intervals. By modelling forgetting and practice dynamics, ALSs show how adaptive learning technologies can support individualized learning trajectories and strengthen memory trace, i.e. how well an item was remembered (Sense et al., 2021). Notably, such learning systems can be significantly beneficial even in short learning sessions, as many vocabulary learning sessions are (Sense & van Rijn, 2022).

To assess the strength of memory encoding, researchers commonly refer to the concept of a memory trace, i.e. how effectively a piece of information has been processed and stored in the long-term memory. And how adaptive learning systems determine the memory trace strength, is by constructing a single memory strength number each time the learner responds to a given tested item. Mostly this memory strength number will be determined by "behavioral proxies", most commonly reaction time and response accuracy, yet relying solely on accuracy in response for predicting memory trace strength was shown to have some limitations. Accuracy-based

models struggle to account for time-related forgetting, as early frameworks often assumed that once an item was answered correctly, it was permanently remembered. Moreover, accuracy does not differentiate between strong and weak correct responses, which reduces predictive precision and often requires a large number of incorrect responses to adjust performance estimates (Wilschut et al., 2023). Alternatively, the memory strength number for memory trace estimation can also be inferred from alternative sources, such as learner's speech and its patterns, which may enhance the ALS's ability to adapt to individual learners' needs in a more efficient way (Wilschut et al., 2021, 2023).

A person's speech can convey a lot of meaningful information that is beyond what is conveyed by the words alone. That type of information is referred to as the prosody of the speech, which can be considered as the variation of acoustic dimensions such as pitch, intensity, rhythm, and duration in one's speech (Xu, 2011). More precisely, speech prosody is the *suprasegmental* properties of speech, arising from time-varying acoustic properties, such as fundamental frequency (F0), amplitude, and the duration of speech segments (Cole, 2015). Wilschut's 2021 study formally showed that within learning systems such as SlimStampen, speech-based reaction times (RTs), compared to typing-based RTs, offer a more sensitive measure of memory retrieval strength. Furthermore, as stated earlier in this paper, information from speech is expected to add more valuable information for estimating a more accurate memory strength number (Wilschut et al., 2023). This information, used for estimating accuracy and confidence, will be useful at home settings as well, as it was found that these patterns in prosody are evident even in the absence of social interaction or a human observer (Goupil & Aucouturier, 2021). In the current study we will focus on using pitch, intensity, and speaking speed as prosodic cues, as we are aiming at replicating Wilschut et al.'s 2025 study, and because

of the differentiation of intensity and pitch as prosodic markers between speech in native and foreign language (van Maastricht et al., 2016). In addition, subjective confidence rating can be a good estimation of objective accuracy (Wilschut et al., 2025), which would be an effect of interest in our study. Further, the pattern of intonation ending with a falling pitch was shown to be significantly correlated with subjective confidence (Goupil & Aucouturier, 2021). Thus we will incorporate subjective confidence as a dependent variable in our study.

Additionally, prosodic information is less reliably informative in speech while speaking a foreign language (van Maastricht et al., 2016). Thus we focus on extending on Wilschut's 2025 study as we (1) recruit native Dutch speakers to control for prosodic variations that may be present when delivering utterances in a foreign known language, and (2) shift from single word cues to simple subject-verb sentences to allow for more prosodic variation. We choose to test native Dutch speakers on sentences in the Italian language for several reasons. First, Dutch language is chosen as a representative of speech in a native language. Second, the Italian language is a Roman language, which is expected to be easier for Dutch speakers to learn, while it is not commonly studied in a Dutch school environment, reducing the chance of them having significant prior knowledge of this language. Additionally, Italian language is classified as non-plastic language, while Dutch is classified as plastic, meaning that in Italian prosody is less sensitive or less flexible in signaling information structure (i.e. Italian speakers would accent both words in a spoken sentence, regardless of their discourse status), whereas in Dutch, prosodic features (accent placement and pitch prominence) are flexibly adjusted to reflect discourse-level functions

(Swerts et al., 2002). This differentiation in language direction ,which was not observed in previous studies, may convey important cross-linguistic differences in how prosody encodes communicative intent and may have implications for future studies examining the cognitive and metacognitive contents of speech. Thus, we may encounter some prosodic differences between the two conditions (Dutch → Italian; Italian → Dutch). Our bidirectional design is expected to account for such language-specific encoding differences.

In this paper we will explore how prosodic speech features, in particular intensity, pitch, and speaking speed, relate to memory trace in a language learning task. As we are building on Wilschut's (2025) work, we will focus on whether simple subject-verb sentences contain more prosodic information than a single word, and whether this information provides more informative input for the API. Additionally, we will investigate in what ways prosodic information (i.e. as being substantive for the API) may differ between native versus foreign spoken language in a language learning task. The difference of performance and task difficulty between the different language conditions will also be explored.

## 2 Methods

### 2.1. Participants

The required sample size was estimated with reference to Wilschut et al.'s paper (2025), which this study aims to replicate. Based on an expected effect size, similar to that reported in their findings, and assuming our desired significance level of $\alpha = .05$ and statistical power of $1 - \beta = .80$, a minimum of 40 participants was determined to be necessary. A total of 50 participants were recruited through the SONA university platform or volunteered. As 2 of the participants'

data was missing, the sample used for the statistical analysis was 48, aged between 18 and 25 (m =19.9, sd =1.65), 14 were male and 34 were female. All of the participants were Bachelor or Master students in the University of Groningen, and native Dutch speakers with no prior knowledge of Italian, as reported by each participant on the pre-screen questionnaire. None of them had reported hearing and/or speech impairments. Participants recruited from SONA received course credit for participation. The study was approved by the ethical committee of the department of Psychology at the University of Groningen (study approval code: 172 PSY-2223-S-0257). Written informed consent was obtained from all participants before the start of the experiment.

**2.2. Design and Procedure**

Upon arrival, each participant was asked to fill in an on-paper background questionnaire together with the informed consent. The background questionnaire consisted of 6 demographic-related questions on age, gender, native language, whether the participant has any speech or hearing problems that they are aware of, and participants' prior knowledge of Italian, academic and non-academic exposure, as well as their proficiency in other Romance languages (Spanish, French, Romanian, Portuguese), specifying self-reported proficiency levels (i.e., High-School Level). We choose Italian as a second language, as it is typically not offered within the standard Dutch high school curriculum, thus there should be a lower on average general familiarity of Dutch participants with the Italian language and the set's psychometric properties (i.e., balanced difficulty and answer frequency). Afterwards participants were asked to sit in the cubicle where they were provided with USB headphones and a computer.

The whole experiment consisted of 7 blocks in total and three phases in total, including a practice phase which served to accustom participants to the task. The first phase consisted of

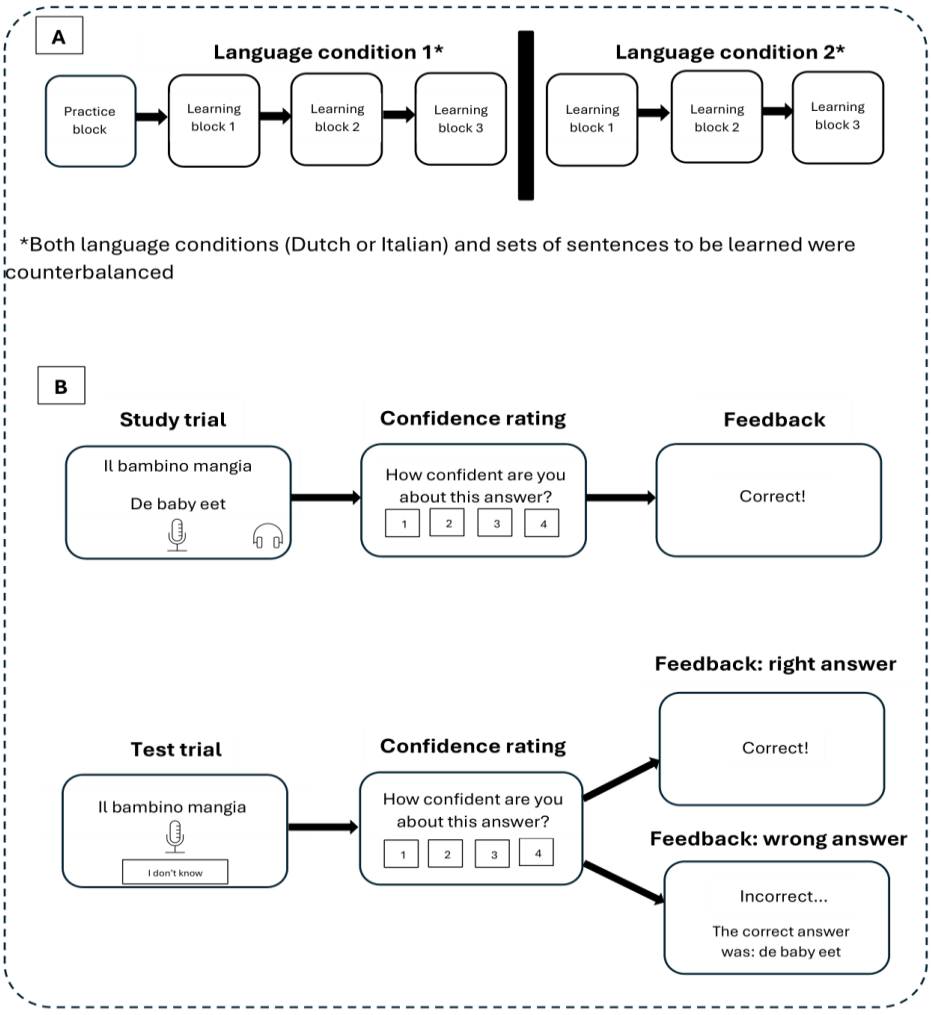
only one practice block, which consisted of 4 sentences repeated 3 times each. The following two blocks were the actual experiment, and they consisted of 5 sentences each. Each sentence was studied 1 time and tested 4 times. For the full structure of the experiment, refer to Figure 1. After the first study phase, participants could take a self-paced break followed by the second study phase. The test trials were identical to the learning trials, besides that they contained different items, with all items presented once, and were followed by feedback. In each block participants were presented with individual items of the Dutch-Italian vocabulary set, either starting with Dutch to Italian (Condition 1), or starting with Italian to Dutch (Condition 2), on a semi-random principle. Every initial presentation of an Italian/Dutch item also presented its Dutch/Italian translation. Two sets of 15 sentences were presented in the study, either in the Dutch or the Italian condition. They were presented in the same order for each participant (i.e. fixed scheduling).

Participants were asked to give the correct translation of the item by speaking it out loud. The spoken utterance was automatically transcribed to text, using Google's Text-to-Speech Assistant, to provide real-time feedback to the learners. Subsequently, participants were asked to rate their subjective confidence in the accuracy of the response, using a Likert scale (‘1 = not confident’, ‘2 = slightly confident’, ‘3 = moderately confident’, ‘4 = confident’). After the confidence rating, feedback was provided in the middle of the screen: ‘Goed!’ if the transcribed speech signal matched the given translation, and "Fout", together with the expected correct answer. Response times were determined by the time elapsed since the presentation of the item and the speech onset. Participants were instructed to directly speak when the microphone icon is present on the screen, without pressing any button. At a later stage of data collection, as we found that there may be an issue, participants were asked to speak only when they are ready to

produce a meaningful word, to avoid any “umm” sounds, which the API detected as a separate response and automatically gave feedback on. Reaction Time was calculated from the very onset of the utterance from each participant. For accuracy estimation Levenshtein distance of 2 was used as a marker/threshold for determining whether an answer is incorrect.

**Figure 1**

*Experiment Design for both Study and Testing trials*



*Note: Panel A shows how the language conditions were constructed. Panel B shows the structure of each trial.*

### 2.3. Materials

Simple subject-verb sentences in either Dutch or Italian, were used for each trial of the experiment. All the sentences were manually created and selected on the basis of their frequency score, calculated with the SUBTLEX-NL database (Keuleers et al., 2010). Each sentence consisted strictly of a subject (e.g. *The child*) and a verb in the simple present tense (e.g. *runs*).

The experiment was built with JavaScript and HTML5 using the Jatos online experiment platform and was conducted in a quiet, sound-insulated lab room. Text and audio items were presented to each participant on a computer screen using the MemoryLab adaptive learning algorithm software (van Rijn et al, 2009) with fixed scheduling of presenting the items. We used Nedis Xyawyon GHST100BK headphones with a built-in microphone, provided to each participant to play and record audio, as each headphone set was set to record/reproduce at a fixed volume of 30. Each spoken response was transcribed by the Google Speech-to-Text system (link to the transcribed responses: data_cleaned_answers_transcriptions). Participants' speech was recorded and saved for later processing, and transcribed to text in real time using the Google Web Speech API. The prosodic markers from each recording were extracted and analyzed using Praat 6.2 (Boersma, 2007). Narakeet (https://www.narakeet.com/) was the platform we used to generate the spoken versions of each of the sentences.

## 3 Results

Our aims were to examine the relationship between prosodic speech features and memory trace in a learning task, and whether the prosodic information will significantly differ from the

native language to the foreign language conditions. We started with examining the behavioral data of the participants, i.e. accuracy, reaction time, and subjective confidence for each response throughout the experiment. The confidence of the API was also inspected, as to control for inaccurately assessed accuracy of response. Next, the acoustic data were analyzed separately and further added to the behavioral data. Participants' performance and prosodic information were compared between Dutch and Italian response conditions. To assess our first hypothesis, we correlated all the 7 dependent variables to each other (i.e. accuracy, RT, subjective confidence, speaking speed, intonation and pitch). A Repeated Measures ANOVA was performed to assess mainly our second hypothesis.

### 3.1. Preliminary analysis and Behavioural Data

Throughout the whole study we used Jasp (Version 0.19.3.0; JASP Team, 2025) and R (v.4.3.1, R Core Team, 2021) to conduct our analysis. Prior to conducting statistical analysis for our dependent variables, the automatic speech recognition system's confidence estimates (i.e., the system's estimate of the probability that a given response was correctly transcribed and classified) were examined by calculating the means and standard deviations for this variable in JASP. This analysis was conducted as the API performance could be a potential confound between accuracy measures and actual performance, thus the need to account for potential limitations in the automated accuracy evaluation. Confidence values were calculated for each participant as to show an average per condition (i.e. Language). The API confidence was computed for the study trials only. The results showed that mean API confidence was higher in the Italian response condition ($M = 0.961$) than in the Dutch response condition ($M = 0.795$), suggesting that the API was better at recognising and correctly classifying responses as correct/incorrect in the condition where participants spoke in Italian. There were nine

participants for which the mean API confidence on Dutch transcriptions was lower than 0.75, primarily due to a small number of trials, for which the API reported very low confidence scores. However, we did not discard those trials, as they constituted only a small proportion relative to the overall number of trials for which the API was high in confidence.

Next, we analyzed the behavioural data separately from the acoustic parameters by calculating means for subjective confidence, reaction time, and accuracy, to estimate participants' overall performance and engagement with the task. As expected, reaction time was longer in the Italian condition for the testing trials than in the Dutch condition. Yet, the study trials for both language conditions had higher average RT, which may be attributed to the process of taking time to actually learn the items. Accuracy was higher in the Dutch condition, and very low in the Italian test trials (m = 0.24). Subjective confidence was rated on a scale from 1 to 4 from each participant. The highest average of reported confidence was for the Dutch study (m = 3.83) and testing (m = 3.72 ) trials. Interestingly, confidence and accuracy were significantly correlated in both the Dutch (r = 0.49) and Italian (r = 0.48) response conditions (see Figure 5).

All the variables, aside from accuracy, were standardised prior to the ANOVA analysis to account for any variation between participants, specifically sex-specific differences in pitch use, as it was suggested by previous research that males and females may have distinct prosodic strategies to convey confidence (Jiang & Pell, 2017).

**3.2 Acoustic data**

To analyze the acoustic data and answer our first hypothesis, we created correlation matrices for both language conditions, including all the seven dependent variables (see Figure 2). Additionally, the variables were analyzed, by conducting a 2x2 RM-ANOVA in JASP (with

language and trial type as factors). Violin boxplots for all the dependent variables between the two language conditions were created in R for the correct trials only (see Figure 1). A significant main effect of language type was found for speaking speed ($F(1, 47) = 64.09$, $p < .001$) and intensity ($F(1, 38) = 5.09$, $p = .030$), but not for average pitch and the change in pitch.

**Intensity**

The mean intensity varied between 69.31 dB (SD = 1.83) in the Dutch testing trials and 70.15 dB (SD = 2.97) in the Italian testing trials. The numbers were similar in the study trials of both language conditions. Mean intensity showed a moderate correlation with accuracy (r = .36, p < .001), and weak correlations with reaction time and confidence. This would suggest that louder utterances were more likely to be correct. Interestingly, a significant negative correlation was found between average pitch and intensity in both language conditions (see Figure 5).

**Speaking Speed**

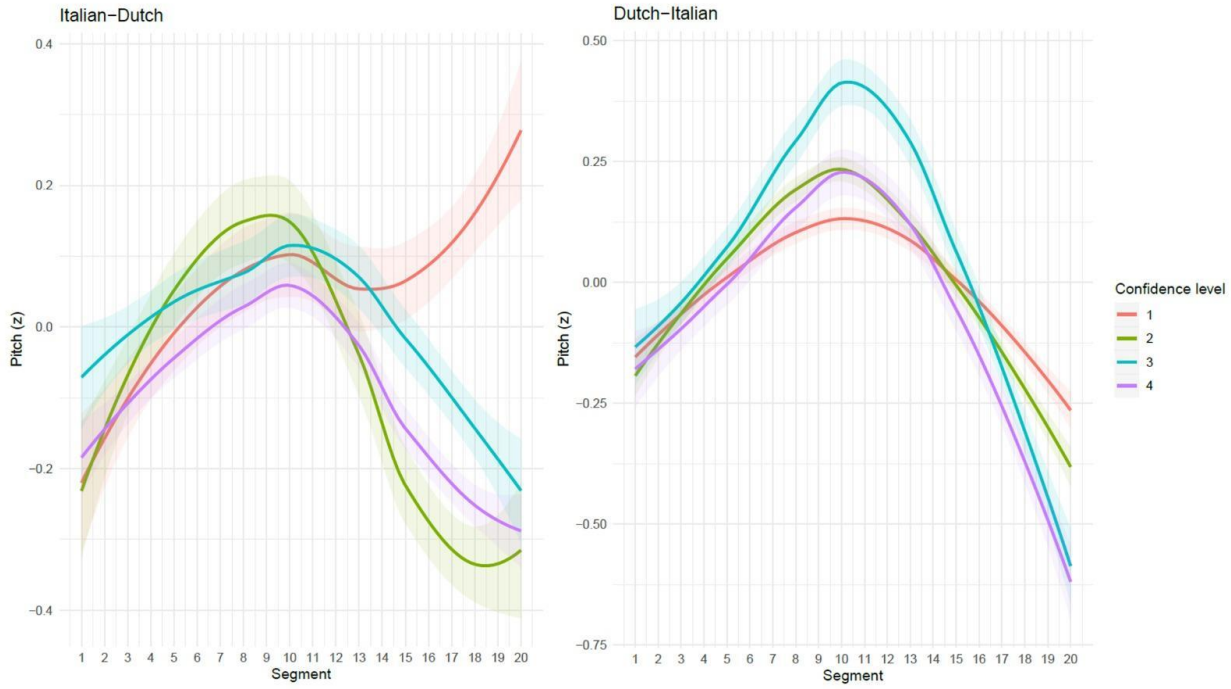
The mean of this variable was 1.57 syl/s (SD = 0.21) in the Dutch testing trials, and 1.71 syl/s (SD = 0.30) for all the Italian trials. The speech rate was significantly lower in the Dutch response condition ($M = 1.47$, $SE = 0.03$), compared to the Italian response condition (M = 1.71, SE = 0.03). Utterances were significantly slower during study trials (M = 1.53, SE = 0.03) than during testing trials (M = 1.64, SE = 0.03). Speaking speed was significantly positively associated with accuracy and confidence ratings in both conditions, negatively associated with pitch average in both conditions, and with pitch change/slope in the Italian response condition. The association with reaction time was negative and significant in the Dutch response condition (see Figure 5).

**Pitch Slope and Average Pitch**

To estimate pitch variation, we had to calculate its slope and change. The slope/change was calculated as we subtracted the average of the last 5 segments, from the average of the first 5 segments of the signals. The average pitch was calculated for the pitch plotted across 20 segments (see Figures 2 and 3). Average pitch (see Figures 4F and 4G) displayed a similar pattern to intensity and was relatively consistent across all conditions. The group means for average pitch ranged from 182.11 Hz (SD = 40.91) in the Dutch testing trials to 185. 03 Hz (SD = 44.82) in the Italian testing trials. Average pitch values correlated highest with intensity in both language conditions. Change in pitch correlations were significant with subjective confidence in both language conditions. (refer to Figure 5). For this variable in particular, we plotted the segments of the slope in both language conditions for both subjective confidence (Figure 2) and accuracy (Figure 3), to observe whether there are some significant patterns. Notably, for the Italian response condition both accuracy and confidence exhibited raising-falling pitch patterns. For the Dutch response condition the patterns for confidence and accuracy were similar, as incorrect responses followed the pattern of the lowest subjective confidence rating, further suggesting a correlation between low accuracy and low confidence.

**Figure 2**

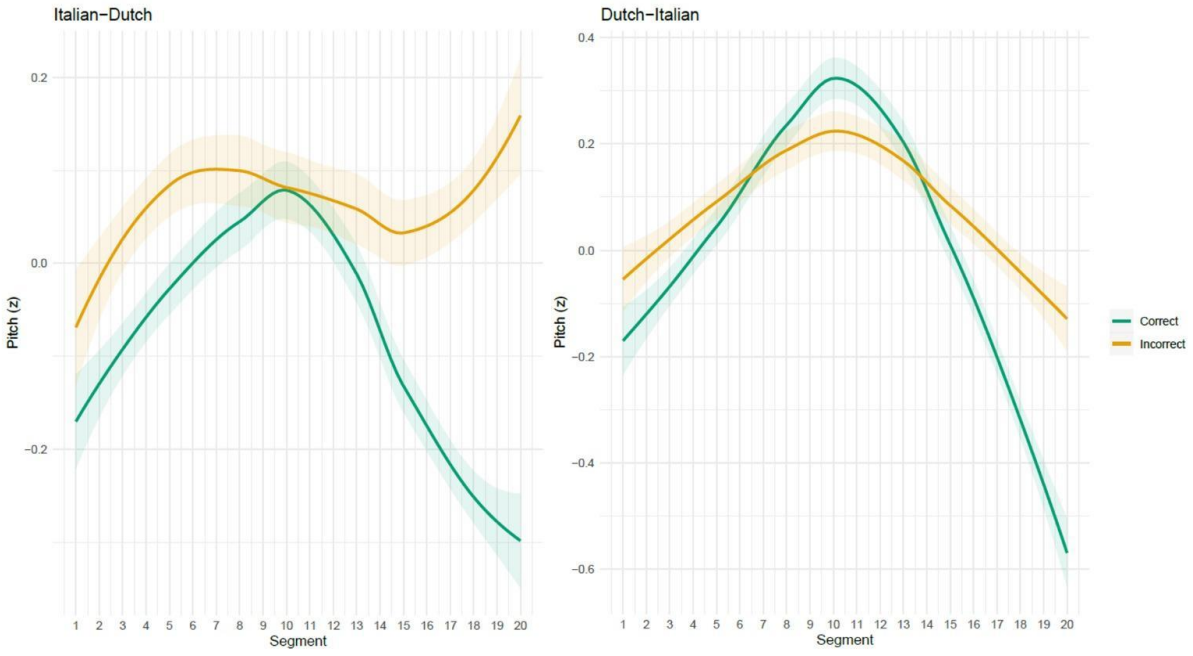
*Pitch (z) Across Segments: Dutch (A) vs. Italian (B), as a function of Confidence Levels*




*Note. Panel A shows standardized pitch for Dutch responses and panel B for Italian responses. Plotted data included testing trials from the learning phase only. Shaded areas represent 95% confidence intervals around the mean pitch (z) estimates, plotted separately for each level of confidence ranging from 1 (not confident) to 4 (confident).*

**Figure 3**

*Pitch (z) Across Segments: Dutch (A) vs. Italian (B), as a function of Accuracy*




*Note. Panel A shows standardized pitch for Dutch responses and panel B for Italian responses. Plotted data included testing trials from the learning phase only. Shaded areas represent 95% confidence intervals around the mean pitch (z) estimates, plotted for correct and incorrect responses separately.*

### 3.3. Repeated-Measures ANOVA: Language and Trial-Type Effects

Repeated-measures ANOVAs were conducted to assess whether language of responses and trial type (study or test) produced significant differences in behavioural and prosodic measures.

Significant main effects of language emerged for accuracy – $F(1, 47)=595.28$, $p<.001$, reaction time $F(1, 47)=59.64$, $p<.001$, confidence $F(1, 47)=204.68$, $p<.001$, speaking speed $F(1, 47)=64.09$, $p<.001$; and intensity $F(1, 38)=5.09$, $p=.030$. In contrast, average pitch and pitch change showed no language effect. Accuracy ($F(1, 47)=315.37$, $p<.001$), reaction time

($F(1, 47)=52.41$, $p<.001$) and confidence ($F(1, 47)=11.49$, $p = .001$) all varied significantly with trial type, whereas intensity did not ($F(1, 38)=0.35$, $p=.557$). Unlike the language factor, trial type did influence average pitch ($F(1, 38) = 8.04$, $p =.007$) and pitch change ($F(1, 38 )= 4.19$, $p =.048$). Significant language and trial-type interaction effect appeared for accuracy $F(1, 47 )= 235.56$, $p<.001$; reaction time $F(1, 47) = 37.93$, $p<.001$; confidence $F(1, 47) = 6.19$, $p = .020$; and speaking speed $F(1, 47) = 16.76$, $p<.001$. Interaction effect was not significant for intensity, average pitch, or pitch change.

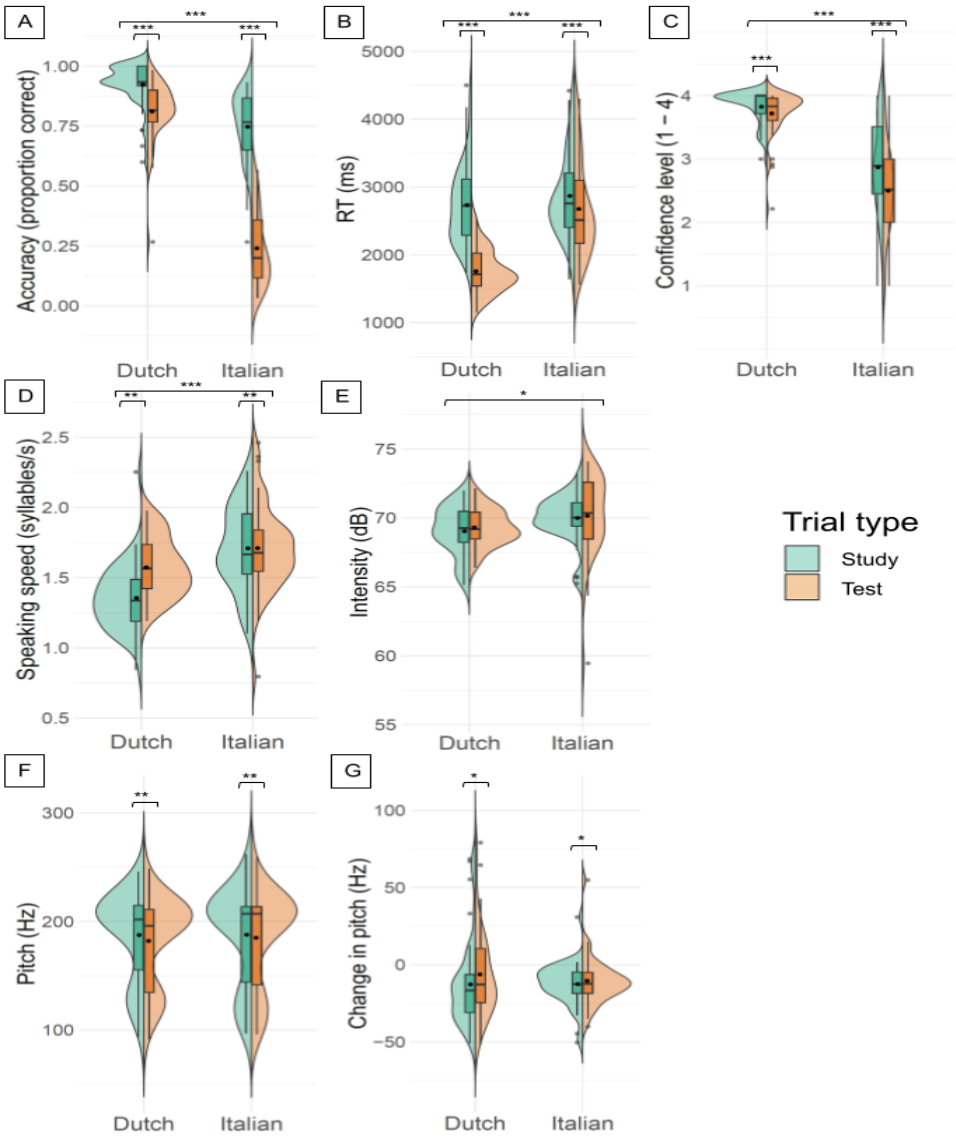
**3.4. Differences between the language conditions and the trial types.**

The aforementioned RM-ANOVA test and correlation matrices (see Figure 5) revealed several differences between the two language conditions. First, the performance of the participants was lower in the Italian response condition, as characterized by behavioural measures, i.e. their accuracy scores, reaction times, and evaluated subjective confidence, which may suggest they experienced greater difficulty in this condition. Further, outliers were mostly present in the Dutch response conditions, rather than the Italian (see Figure 4), which may be connected to the lower API confidence in this condition ($M=0.795$), yet no participants data were removed on that basis. Reaction time was fastest for the Dutch testing condition, and slowest for the Italian study condition, and accuracy was highest for the Dutch condition (see Figure 4). Confidence ratings were higher in the Dutch response condition as well, while in the Italian these ratings were more dispersed. Speaking speed, intensity and pitch average did not show any significant variations between the language conditions, yet pitch change/pitch slope shows slightly more variation in the Dutch response condition, compared to the pattern in the Italian condition (see Figure 4). The correlations between the behavioral measures (accuracy, confidence, and RT) and the acoustic data were overall higher in the Dutch response condition,

compared to the Italian response condition. An exception were speaking speed and accuracy (see Figure 5).

**Figure 4**

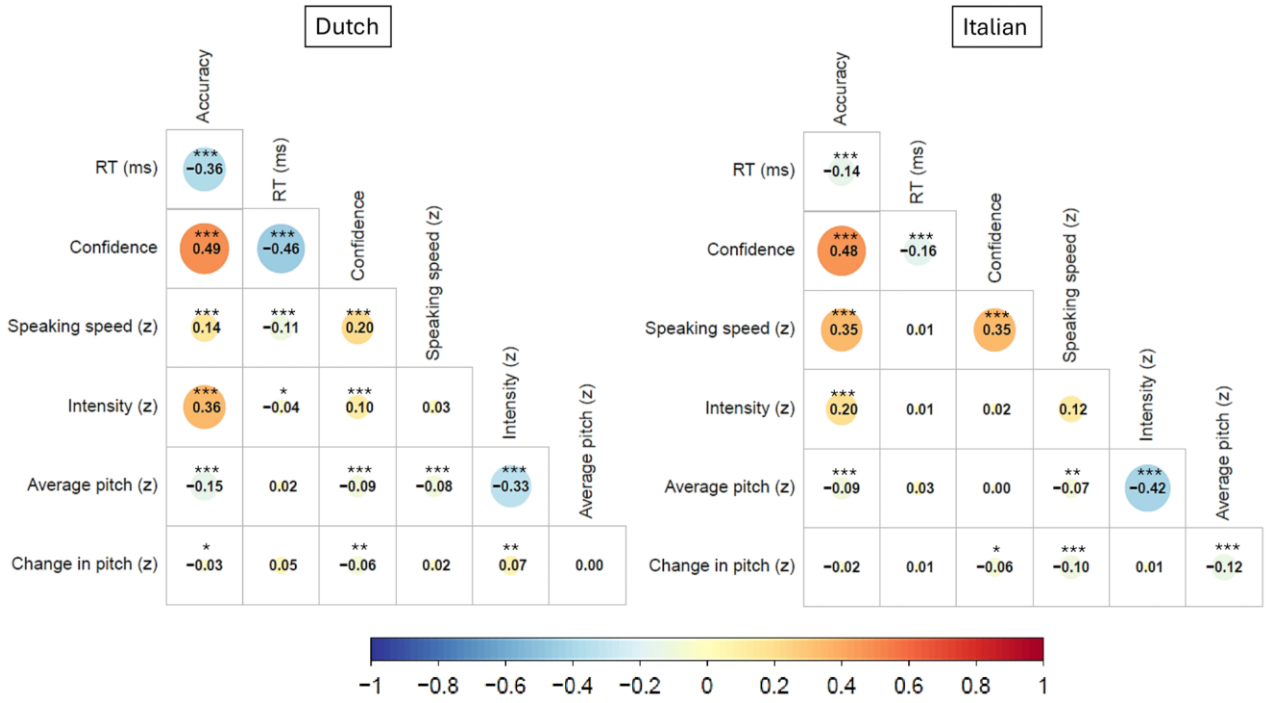
*Distributions of responses for all 7 dependent variables*




*Note: Distributions for each variable are split by language response condition and trial type.*

**Figure 5**

*Standardised Correlations Matrix for the Both Language Conditions*




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

## Discussion

The main aim of the current study was to explore the relationship between intensity, pitch and speaking speed as prosodic speech features, subjective confidence and memory trace (i.e. measured by accuracy and reaction time) in a language learning task, consisting of simple subject-verb sentences.We wanted to compare whether prosodic information differs between single-word cues, as examined by Wilschut et al. (2025) and Goupil and Aucouturier (2021), and simple subject-verb sentences. In addition we aimed at finding whether the language direction, i.e. whether a response is spoken in the native or the foreign language, has influence on the prosodic information.

Upon analyzing the prosodic data, the results revealed a nuanced pattern of relationships across the Dutch and Italian response conditions. Intensity and speaking speed generally showed significant moderate correlations to accuracy and subjective confidence, similar to Wilschut et al. 's 2025 study, suggesting that higher intensity in voice and faster speaking speed are generally associated with higher accuracy and subjective confidence rates. Yet more specifically, and as we compared our findings to Wilschut et al.'s (2025), speaking speed is more associated with confidence, while intensity correlates more with accuracy. Pitch values (i.e. pitch change and average pitch) showed small yet significant associations with accuracy and confidence, as compared to the findings in Wilschut's 2025 study. Notably, the pitch variation between correct and incorrect responses in the Italian response condition did not differ significantly (see Figure 5). The way we estimated pitch was by dividing each signal to segments, 10 in the beginning and 10 in the end of the utterance (Section 3, Figure 2). By doing this we might have potentially lost some information, which can be accounted for, if pitch is calculated in several ways, which we implemented regardless, as we calculated both pitch change and pitch slope. Yet there is a chance of missed information in our approach. Future studies can attempt to calculate pitch in different ways, to account for such loss of information. On another note, Jiang & Pell (2017) showed that generally higher variation in pitch was associated with confident responses. We observed overall higher variation for correct responses, especially in the Italian response condition, which 1) may be a potential solution for the limitation of variation in pitch calculation and 2) confirm the high relation between accuracy and subjective confidence we found.

It is important to mention here that participants' confidence ratings were congruent with their accuracy, suggesting that subjective confidence evaluation was a good predictor of memory trace strength (judging by accuracy and the corresponding reaction times). As previously

suggested, a pattern of confident or accurate response, as in our study subjective confidence was a good predictor of accuracy, may follow a rising-falling pitch pattern (Goupil & Aucouturier, 2021), which was evident in the Dutch response condition but not in the Italian response condition, which may simply be due to prosodic information being less reliable in speech while speaking a foreign language (van Maastricht et al., 2016). Yet that finding can be a potential future research direction, as to see whether prosodic information might have been influenced by other factors in the foreign language condition, such as the fact that we used an AI speaker. Alternatively, the voice of a human native speaker may influence the prosody differently in the foreign response condition.

As we aimed at exploring whether simple sentences contain more prosodic information than the single word cues, we explored the correlations between accuracy and each prosodic feature. As compared to Wilschut's 2025 study, the correlations we obtained were overall higher. For instance, Wilschut et al. reported a correlation between accuracy and intensity of $r = 0.12$ ($p < 0.001$), while we reported a correlation of $r = 0.36$ ($p < 0.001$) in the Dutch response condition. The same pattern was evident for speaking speed, yet not for pitch. Yet, this may suggest that simple sentences can convey more prosodic information than single word cues. An improved learning system, and potential object for future research in this field, could be a hybrid design, where single words and simple sentences are counterbalanced, as the learner starts with learning single words and shifts towards sentences, with the help of an adaptive scheduling model. This would benefit as it will be a form of spaced learning, which is more beneficial for the learner, especially if personalized (Khajah et al., 2014; Lindsey et al., 2014), and as the difficulty will be counterbalanced, avoiding the drawback of extensive task difficulty, pointed out by some of our participants.

As previously discussed, there may be systematic differences in prosodic information between response conditions, that is, between spoken responses produced in the native language versus those in the foreign language (Swerts et al., 2002; van Maastricht et al., 2016). Indeed, it was evident that prosodic features differed between the language response conditions, as the prosody in the Dutch response condition had stronger correlations between the memory trace, confidence, and the prosodic features, than in the Italian condition. This would suggest that indeed speech in the native language (in our case Dutch), is richer in prosodic variation than a newly acquired language. Besides being a foreign language, the non-plastic nature of Italian might have contributed for less variation in prosody, compared to the plasticity in Dutch language (Swerts et al., 2002), as it was previously discussed in the introduction of this paper. Thus native speech can be more informative for predicting the learner's performance. Another relevant point to consider is that during the debrief sessions, participants commonly reported the Italian response condition being more challenging than initially anticipated, which could indicate that our design might have been made more complicated than desired. In future research, a hybrid design combining elements from the current study and that of Wilschut et al. (2025), may be employed to better account for the increased cognitive demands, associated with longer linguistic stimuli.

Aside from our main research aims, we evaluated whether the language learning task was appropriately calibrated in terms of difficulty for the participants and whether the automatic speech recognition (API) system reliably assessed the correctness of their responses. Regarding task difficulty, our findings suggest that the task may have been more challenging than initially anticipated. Several participants reported during the debriefing session that the sentences were too complex to memorise. This subjective feedback aligns with the objective accuracy and RT

measures (see Section 3), which indicated low accuracy levels in the test trials for the Italian response condition (M = 0.24). Concerning the reliability of the API system in determining response accuracy, we examined the API's confidence scores (see Section 3), which showed exceptionally high confidence in the Italian response condition (M = 0.961) and moderate confidence in the Dutch response condition (M = 0.795), suggesting that the system's estimations of accuracy were generally reliable. However, upon manually inspecting some of the transcriptions, particularly within the Dutch response condition, we identified several anomalies, including implausible transcriptions that likely did not reflect the participants' actual spoken responses. That may suggest the API had difficulty in correctly identifying the accuracy of participants' responses in Dutch, potentially leading to misinterpretations of accuracy estimates in this condition, and should be considered when interpreting those results. Indeed, as shown by Zhang and colleagues (2025), there are certain low-resource languages, i.e. with which the given software has 10h or less of training, that may not have been sufficiently represented or analysed by the speech recognition system we used, or the lack of advanced vocabulary data in Dutch (Kuhn et al., 2024), thus this potential lack of representation could have contributed to the recognition errors we observed.

Our findings may further facilitate the improvement of speech adaptation within ALSs like MemoryLab, which may additionally be beneficial for dyslexia patients (Wilschut et al., 2025) and practicing speaking skills, an essential part of achieving comprehensive language knowledge.

## Conclusion

This study investigated which cognitive and metacognitive indicators of memory retrieval can be detected in the speech signal during spoken recall. Participants learned vocabulary

through verbal retrieval practice, producing responses bidirectionally, either translating from Italian to Dutch or vice versa. The findings revealed two main outcomes. First, pitch, intensity and speaking speed as prosodic features of speech can be informative of memory trace during a language learning task, as well as higher subjectively evaluated confidence being a good predictor of accuracy, together with faster reaction time. In addition, we were the first to examine the difference of prosodic information between single word and simple subject-verb sentence cues, which may contribute to improving the performance of ALS in language learning settings, especially if a hybrid model, incorporating both single words and simple sentences, with adaptive scheduling, is incorporated. Second, we used a bidirectional design, which revealed that the direction of the language of the spoken response did show to be of importance for the prosodic information. More specifically, spoken responses in the native language were more informative regarding prosodic information than were the responses in the foreign newly acquired language. In conclusion, our findings offer valuable insights for the improvement of Adaptive learning systems incorporating speech input. Specifically, the inclusion of simple sentence structures and native-language responses, appears to enhance the informativeness of prosodic features, thereby improving the model's evaluation of memory strength in language learning tasks.

## References

Boersma, P. (2006). Praat: doing phonetics by computer. http://www.praat.org/

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30* (1–2), 1–31. https://doi.org/10.1080/23273798.2014.963130

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

JASP Team (2024). JASP (Version 0.18.3)[Computer software]

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

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Khajah, M. M., Lindsey, R. V., & Mozer, M. C. (2014). Maximizing Students' Retention via Spaced Review: Practical Guidance From Computational Models of Memory. *Topics in Cognitive Science*, *6*(1), 157–169. https://doi.org/10.1111/tops.12077

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the Accuracy of Automatic Speech Recognition Solutions. *ACM Trans. Access. Comput.*, *16*(4), 25:1-25:23. https://doi.org/10.1145/3636513

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

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

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

Swerts, M., Krahmer, E., & Avesani, C. (2002). Prosodic marking of information status in Dutch and Italian: A comparative analysis. *Journal of Phonetics*, *30*(4), 629–654. https://doi.org/10.1006/jpho.2002.0178

van Maastricht, L., Krahmer, E., & Swerts, M. (2016). Prominence Patterns in a Second Language: Intonational Transfer From Dutch to Spanish and Vice Versa. *Language Learning*, *66*(1), 124–158. https://doi.org/10.1111/lang.12141

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.

Wilschut, T., Sense, F., Scharenborg, O., & van Rijn, H. (2023). Improving Adaptive Learning Models Using Prosodic Speech Features: 24th International Conference on Artificial Intelligence in Education, AIED 2023. *Artificial Intelligence in Education - 24th International Conference, AIED 2023, Proceedings*, 255–266. https://doi.org/10.1007/978-3-031-36272-9_21

Wilschut, T., Sense, F., van der Velde, M., Fountas, Z., Maaß, S. C., & van Rijn, H. (2021). Benefits of Adaptive Learning Transfer From Typing-Based Learning to Speech-Based Learning. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.780131

Wilschut, T., Sense, F. & van Rijn, H. (In review) The sound of recall: Cognitive and Metacognitive Markers of Memory Retrieval Performance in Speech Prosody.

Xu, Y. (2011). Speech prosody: A methodological review. *Journal of Speech Sciences*, *1*(1), 85.

Zhang, L., Wu, S., & Wang, Z. (2025). End-to-End Speech Recognition with Deep Fusion: Leveraging External Language Models for Low-Resource Scenarios. *Electronics*, *14*(4), Article 4. https://doi.org/10.3390/electronics14040802