“The Student Speaks”: The Relationship Between Prosodic Speech Features, Subjective Confidence and Response Accuracy in Language Learning

Thijs Peters

S5141540

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group number: 19

Supervisor: Dr. Tassos Sarampalis

Second evaluator: Dr. Olaf Dimigen

In collaboration with: Y. Daskalova, E. Fernández-Amurrio, A. Jarczynska, A. Lukassen, N. Velthuizen

June 27, 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.

Acknowledgements

First, I would like to thank Tassos for his dedication in supervising our group and for being such an approachable and supportive teacher (and for teaching me what a “bijzin” is). I truly appreciated being part of your thesis group and your approach to education. Second, I would like to thank Thomas for all his help with setting up and conducting the experiment, answering all our very specific questions and helping us figure out all the data. Third, I would like to thank the whole group for everybody’s enthusiasm and dedication while working on this project. And last but not least, I want to thank Thessa for supporting me in every step along the way. From getting me coffee and lunch, to sitting next me while I was writing and ensuring me that everything will be fine. I truly could have not done this project without you.

I acknowledge the use of ChatGPT (chatgpt.com), Claude (claude.ai) and Google Gemini (gemini.google.com) to brainstorm about my research and to assist me with improving my writing in the drafting of this assessment.

Abstract

In the last decade, advances in memory research have led to the development of adaptive learning systems (ALS). ALS use reaction time and accuracy of learners to optimize item scheduling for learning facts, like studying vocabulary in a foreign language. Previous research has shown that this approach and its benefits can be translated to speech-based ALS, by using prosodic speech features (PSF) of learner responses, which seem to be indicative of the subjective confidence and response accuracy of the learner. This study investigates whether earlier findings for single words can be replicated for multiple-word items. Additionally, it aims to compare how PSF are indicative of subjective confidence and response accuracy when studied in the native language and a new foreign language. This was studied using a 2x2 within-subject design, with a sample of Dutch speaking university students (*M* = 20.1 years old, *SD* = 1.8; 34 females and 14 males), who studied Dutch and Italian vocabulary items in a speech-based language task. Results show that PSF are also associated with subjective confidence and response accuracy for multiple-word items, but the strength of these associations may depend on language direction. These findings contribute to further development of speech-based language learning systems.

Keywords: prosodic speech features (PSF), adaptive learning systems (ALS), subjective confidence, response accuracy, foreign language learning

“The Student Speaks”: The Relationship Between Prosodic Speech Features, Subjective Confidence and Response Accuracy in Language Learning

Understanding how to accurately predict learners’ memory performance represents an important goal in translating memory research into useful applications for education. This is especially beneficial for aiding learners with memorizing big sets of facts, like needed for acquiring vocabulary in a foreign language. Recent developments in memory research have led to the creation of adaptive learning systems (ALS; Lindsey et al., 2014; Papoušek et al., 2014; Settles & Meerder, 2016), which can help with the difficult process of memorizing vocabulary of a foreign language (Wahyuningsih et al., 2023). ALS can be applied to different fact learning contexts, and in general aim to improve the memorization of items by using learning behaviour measures (reaction time and response accuracy), often based on the typing or clicking of users, which are used to estimate in real-time the most optimal presentation of items to fit the specific learning process of the user (Lindsey et al., 2014; Papoušek et al., 2014; Settles & Meerder, 2016; van Rijn et al., 2009). Using these ALS seems to result in improved learning efficiency when compared to non-adaptive alternatives (Lindsey et al., 2014; Papoušek et al., 2014; van der Velde et al., 2021; van Rijn et al., 2009).

The efficacy of these ALS depends on the accuracy of their estimations regarding the successful memorization of items by the learner (Wilschut et al., 2022). A promising type of ALS that has shown to improve learning efficiency in both experimental and applied contexts is the MemoryLab (ML) typing-based ALS (Sense et al., 2018). ML uses the ACT-R declarative model of memory equations (Anderson et al., 1998; Pavlik & Anderson, 2005, 2008), which model how items are stored in memory and decay from it at a certain rate over time. The system tracks the reaction times and response accuracy for each item being studied by the user, and uses them to estimate a rate-of-forgetting parameter (RoF) parameter (Wilschut et al., 2021). This RoF parameter predicts when an item will become irretrievable from memory, and is continuously updated while the items are being studied. Based on the RoF parameters for all these items, the ML system predicts the most optimal item scheduling to optimize the learning efficiency for a specific learner (van Rijn et al., 2009; Sense et al., 2016).

An extensive body of research (Lindsey et al., 2014; Papoušek et al., 2014; van der Velde et al., 2021; van Rijn et al., 2009) indicates that ALS have helped to improve the learning efficiency of systems that rely on typing-based input. However, the use of ALS to optimize learning with other types of input, like speech-based learning, is still in development (Wilschut et al., 2021). This highlights an interesting opportunity for language learning, considering that, in addition to studying vocabulary items, developing speech skills is an important component of becoming fluent when learning a new language. Speech-based learning is already integrated by popular language learning systems (e.g., Duolingo, Babbel), but they do not yet use information from speech to optimize the learning process in real-time. Interestingly, studies testing a speech-based variant of the ML system found that its benefits for learning efficiency seem to transfer to the speech-based variant and that this resulted in comparable learning outcomes (Wilschut et al., 2021, 2024b).

Recent research proposes that, in addition to the reaction time and response accuracy that can be extracted from spoken responses in these speech-based ALS, the speech itself also contains information about the memory performance of a learner (Wilschut et al., 2023). Specifically, there is information in prosodic speech features (PSF), which are features of speech that extend beyond the phonemic level and communicate more nuanced information about the emotions, tone, and emphasis of the speaker (Wennerstrom, 2001; Xu, 2011). PSF can be categorized in several ways; one way is to divide them into *intonation, rhythm* and *stress*. Intonation is the harmonic structure of the speech, which is defined by the patterns of pitch that are produced. Rhythm is the change in timing or speaking speed throughout the speech. Lastly, stress describes the relative intensity (i.e., loudness) that the speaker gives to different syllables that make up their speech (Xu, 2011). These PSF contain information about the confidence of a speaker in their speech (Jiang & Pell, 2017; Goupil & Autocouturier, 2021; Goupil et al., 2021) and are related to the objective accuracy of the speaker (Goupil & Aucouturier, 2021). Specifically, objective accuracy seems to be associated with the loudness of the responses, while the subjective confidence is reflected by speaking speed and pitch.

In the context of language learning using speech-based systems, PSF (in particular speaking speed, intensity and pitch) seem to be useful indicators for predicting future memorization (Wilschut et al., 2023). Recent research investigated which PSF specifically relate to subjective confidence and response accuracy in this language learning context (Wilschut et al., 2025). They used a non-adaptive language learning task, in which participants were asked to learn the English translation of Lithuanian words. They found that subjective confidence was a significant predictor for response accuracy, and found distinct patterns of PSF that reflect both accuracy and subjective confidence, in line with earlier findings (Goupil & Aucouturier, 2021). However, the findings from Wilschut et al. (2025) are limited to single-word items.

The current gap therefore concerns whether the relationship between PSF and subjective confidence and response accuracy translates to multiple-word items, given that longer items allow for more explicit display and variance in PSF (Reed, 2010). Secondly, these studies (Wilschut et al., 2023, 2025) used non-native English speakers learning the English translation of foreign language items. Therefore, this does not consider to what extent the relationship between PSF and subjective confidence and response accuracy is similar when items are studied in the native language of the learner compared to in the foreign language that is being studied. This is important to consider, given that PSF may differ in native speech compared to non-native speech, differing in for example how stress is used (Landblom & Ionin, 2022). This could possibly affect how PSF can be indicative of subjective confidence and response accuracy in this language learning context.

The current study contributes to the advancement of speech-based ALS by investigating the potential of including a broader range of items, longer and more complex ones, expanding their usefulness for language learning. Furthermore, if indeed the relationship between PSF and subjective confidence and response accuracy applies to these multiple-word items, this will help to further improve real-time estimations of subjective confidence and response accuracy without requiring explicit input from learners, helping to improve the prediction of memory performance in ALS. Lastly, investigating the relationship of PSF with subjective confidence and response accuracy for both the native and a new foreign language of learners will expand our understanding of how PSF are displayed in both languages in a language learning task. This furthers our knowledge of how confidence and memory performance relate differently to PSF in different languages in a language learning context.

The current study aims to investigate the relationship between PSF (speaking speed, intensity, average pitch and pitch change), the subjective confidence of participants, and participants' response accuracy in the context of a language learning task in which they study a set of Dutch-Italian subject-verb phrases (e.g., "The boy runs"). Additionally, it aims to investigate how the relationships between PSF and subjective confidence and response accuracy differ when participants have to respond in their native language (Dutch) compared to in a foreign language (Italian). We hypothesize that: **H1**: PSF will be more indicative of both the response accuracy as the subjective confidence for multiple-word items, compared to single-word items, and that **H2**: the relationships of the PSF with subjective confidence and response accuracy to be language dependent.

Methods

Participants

This study included 48 Dutch individuals aged between 18 and 25 years (*M* = 20.06, *SD* = 1.83). The sample consisted of 14 (29.2%) males and 34 (70.8%) females. All participants indicated having received formal language training in English, and almost all had training in German (*n* = 44) and French (*n* = 43). Other frequent languages included Spanish (*n* = 14), Latin (*n* = 12), and Greek (*n* = 11). A few participants received training in Chinese (*n* = 2), Frisian (*n* = 1), Norwegian (*n* = 1) or Danish (*n* =1). The desired sample size was estimated based on effect sizes from prior comparable research that found small significant effects (Wilschut et al., 2025), though precise calculations were difficult due to the limited existing work in this emerging area of study.

The sample was a convenience sample of mostly first-year psychology students at the University of Groningen, who received course credits for their participation. A few participants were other students recruited from our own social networks. Participants could only participate in this study if they were native Dutch speakers with no previous formal language training in Italian and no speech or hearing impairments. Data from two participants were excluded due to technical issues during testing. The study was approved by the ethics committee of the Department of Psychology of the University of Groningen (approval code: PSY-2223-S-0073).

Design and Procedure

This study employed a within-subjects design in which participants studied sets of Italian-Dutch subject-verb phrases (e.g., “The boy runs”) by completing two experimental conditions: Dutch-to-Italian and Italian-to-Dutch. The order in which they completed these conditions was randomly assigned to ensure a counterbalanced design. There were two sets of 15 phrases; which set was learned in which direction was also counterbalanced across participants. The experiment consisted of one practice block in which five example sentences were studied. This was followed by two learning conditions, each consisting of three blocks with a break in between (see Figure 1A). Each block consisted of five phrases that were first presented in a study trial and four times in a test trial (see Figure 1B). The 30 written and spoken stimuli used for the trials were also randomly allocated to both conditions for each participant.

This study used the language condition—either Dutch-to-Italian or Italian-to-Dutch—as the independent variable and measured behavioural measures (reaction time, response accuracy and subjective confidence) and acoustics (speaking speed, intensity, average pitch and pitch change) based on the participant's utterance for each trial as the dependent variables.

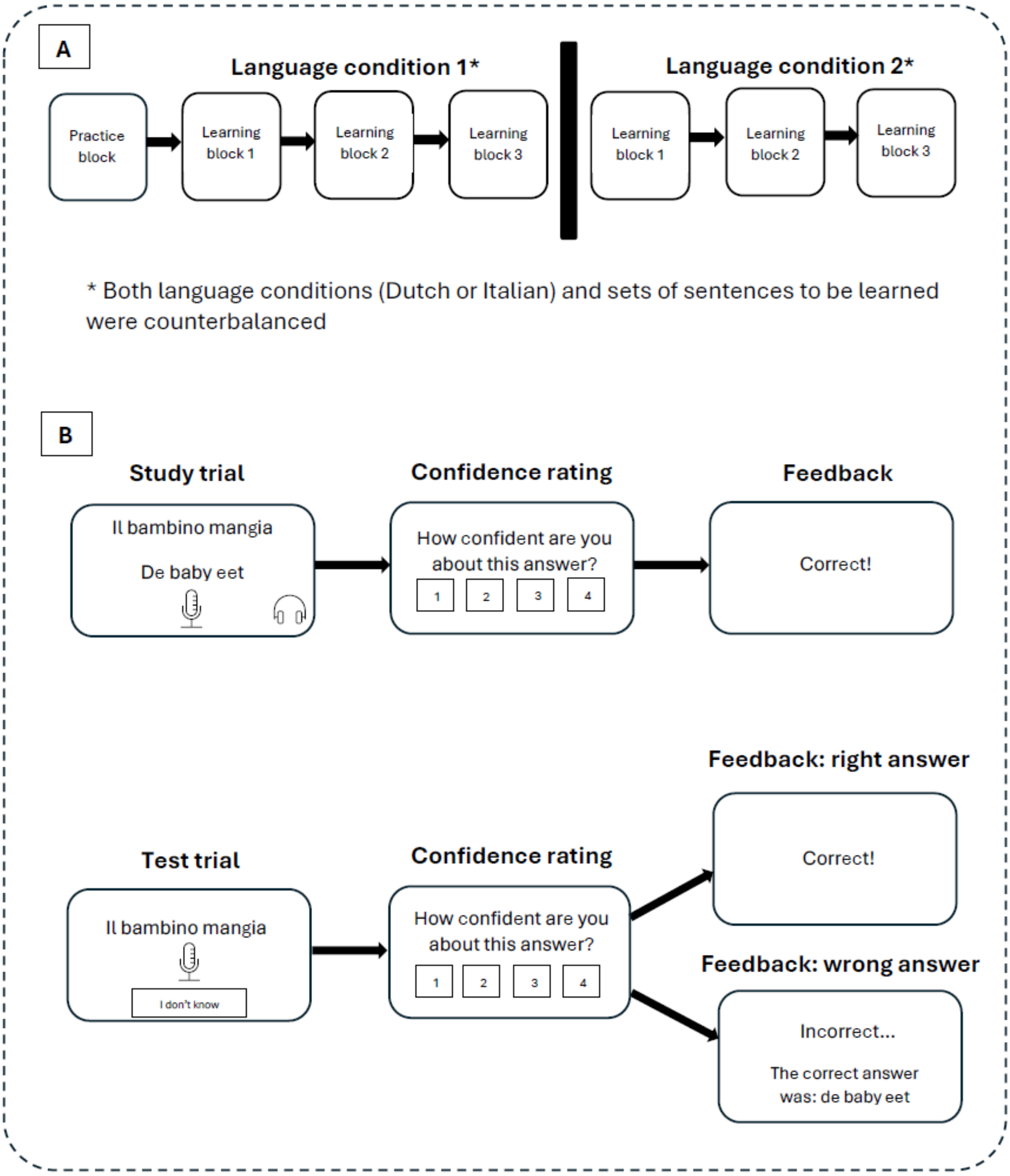
Procedure

Participants were instructed to study a set of Italian-Dutch subject-verb phrases. Before starting the learning task, participants were first asked to complete a pen-and-paper background questionnaire, read the information letter of the study, and provide written informed consent for their participation. Next, the participants were seated in a quiet testing booth, where a computer screen was used to display the learning application and over-ear headphones were used to play audio stimuli and record utterances of the participants. All participants completed two conditions: learning phrases from Dutch to Italian and from Italian to Dutch. The order of the conditions and which set of phrases was learned in which condition was counterbalanced. All possible combinations of conditions were put in a random order and each new participant was assigned based on this sequence.

The experiment started with the application displaying basic instructions about the workings of the experiment on the screen, followed by a practice block. During the practice block, the participants were able to familiarize themselves with the learning application by

Figure 1

Visualization of the Experimental Design



Note. A: Overview of the practice and learning blocks in both language conditions (Dutch and Italian). B: Visualization of the interface that participants saw when completing study and testing trials.

practicing five phrases, each one being presented once in a study trial and twice in two separate test trials, in the language of their first assigned condition. Following the practice block, participants completed three learning blocks in their first condition, each block presenting five unique phrases from their assigned fifteen stimuli set via a fixed random item scheduling. This means that sentences were randomly assigned to one of the three blocks and randomly ordered within trials, with all participants following the same predetermined randomized sequence. Each phrase was first presented to the participant once in a study trial, where they were shown the written translation of the phrases in both languages as well as hearing the spoken translation of the language of their first condition. The test trials then followed, presenting each phrase four additional times according to the same fixed schedule used for the study trials, prompting them to provide the correct spoken translation. For each trial, participants also had the option to click an “I don’t know the answer” button before providing their response. However, participants were instructed to only use this when they had no idea of any of the words or sounds that made up the phrase. See figure 1B for a visualization of the study and test trials.

Immediately after their answer was recorded, the participants rated their subjective confidence in their answer. After indicating their confidence, the recorded utterance of the participant was transcribed via the Google speech-to-text API (Google, n.d.) that recorded their answers, and compared with the correct answer. To prevent correct answers being labelled as incorrect due to small transcription errors, a Levensthein’s edit distance of 2 or smaller was used to determine whether an answer was correct (Wilschut et al., 2024b). This meant that if the transcription was two letters off, the answer was still counted as correct. If the answer was indeed correct, the screen displayed “Correct!”. If the answer was incorrect, the screen displayed “Incorrect!” together with the correct answer. Upon completing the three learning blocks of the first condition, participants took a self-paced break before proceeding to the second condition. This was identical in structure, but with a reversed learning direction of the languages and 15 new phrases from the other stimuli set. The entire experiment, including the preparation, one practice block and two learning blocks lasted approximately 40 minutes.

Materials

Stimuli

This study used a set of 40 written stimuli consisting of simple subject-verb phrases in the simple present tense in both Dutch and Italian, generated by the experimenters (see Appendix A). The stimuli were generated so that no words had a similar origin in the Dutch and Italian translation to minimize any potential linguistic similarities, to ensure low familiarity. Thirty phrases were randomly selected to be used in the learning blocks. The phrases were categorized based on the number of syllables to ensure balance in terms of difficulty, and then randomly divided into two fixed sets of 15 phrases. Out of the ten remaining phrases, four were randomly selected to be used in the practice block and the other six were discarded.

For each phrase a text-to-audio voice generator web application was used to generate the spoken stimuli in both languages (Narakeet, 2025). For all the Dutch stimuli a female voice (Famke) was used to generate the audio files. For 24 Italian stimuli a female voice (Isabella) was used and for the other 16 a male voice (Vittorio) was used.

Instruments

The computer screen that was used to display the learning application was the Iiyama ProLite G2773H (27-inch, 1920×1080 pixels, 60 Hz refresh rate; 59.8 cm W × 33.6 cm H × 68.6 cm D). To record the utterances of the participants and to present the audio stimuli the Nedis Xywayon GHST100BK USB over-ear headphones with a built-in microphone were used.

The spoken stimuli and their corresponding written stimuli were integrated into a learning application developed for this study. The application was constructed using JavaScript and HTML5, based on the *jsPsych* library for behavioural experiments. For each trial the utterance of the participant was recorded as a single audio file and transcribed via the Google Speech-to-Text API (Google, n.d.).

Variables

Demographics. Participants completed a short background questionnaire that was specifically designed for this study. In this questionnaire, they reported their age and gender, and confirmed their status as native Dutch speakers and the absence of speech and hearing impairments. Participants also indicated their second language knowledge by listing any languages in which they had received formal training.

Behavioural measures. To measure memory performance, reaction time and response accuracy were measured. Reaction time was measured in milliseconds (ms) as the time between the stimuli being shown and the participant starting to speak their answer. Response accuracy was calculated as the rate of the correct answers divided by the total amount of correct and incorrect answers. To measure subjective confidence, this study used a 4-point Likert scale ranging from *not confident* (1) to *confident* (4) for the participant to indicate their confidence in their answer. This was in contrast with the continuous confidence slider that was used in the study by Wilschut et al. (2025).

Acoustics. For each utterance the following PSF (acoustics) were measured: speaking speed, intensity, average pitch and change in pitch. Speaking speed was measured by dividing the amount of syllables spoken in the response by its total duration (syl/sec). The intensity was measured in decibel (dB). Change in pitch (Hz) was calculated based on the difference in Hz of the last five segments of the utterance compared to the first five segments. This was done by dividing each utterance into 20 equal time segments, and calculating the average pitch (Hz) for all utterances per segment. The average of the first five segments and the last five segments was calculated, and subtracted from each other to calculate the change in pitch (Hz). Average pitch was measured in Hertz (Hz) as the average over all 20 segments.

API Confidence. For each trial, the Google Speech-to-Text API calculated a confidence score between 0.0 and 1.0, indicating the confidence of the system in correctly transcribing the utterance of the participant.

Data Preprocessing

All the audio files containing the utterances were afterwards analysed with Praat 6.2.07 (Boersma, 2007). Using the Praat software, the relevant PSF for this study (speaking speed, intensity, average pitch and pitch change) were calculated and extracted from the utterances, as well as the duration of the response. For each language-trial type combination, averages were calculated per participant for all variables. These averages only included correct responses, except for the average response accuracy. For the correlation matrices, the PSF variables were standardized to *z*-scores across trials per participant.

Statistical Analyses

All calculations were conducted and visualizations were created using R (R Core Team, 2024).

Descriptive Statistics and RM-ANOVAs: Behavioural Measures and Acoustics

To calculate the descriptive statistics and generate violin plots for all the dependent variables, we included the measurements of the correct and incorrect responses from the study trial and the correct responses for the test trials. In order to compare these responses for their behavioral measures and acoustics between language conditions and trial types, we performed two-way repeated measures ANOVAs. We generated Q-Q plots for all dependent variables to check for the normality assumption, which was met if the data was approximately on the diagonal Q-Q line.

API Transcription Confidence Across Language Conditions

In order to check for any inconsistencies of the API in its accuracy to transcribe the utterances of participants between language conditions, we used the study trials to check the working of this system. As in the study trials participants only had to replicate what they heard, this served as a check to compare the performance of the system between the two languages. We did not exclude any participants based on their API confidence scores of their trials.

Associations Between Behavioural Measures and Acoustics

Correlation matrices were created in order to describe the associations between the behavioural measures and standardized acoustics for both language conditions. We used the correct and incorrect responses of all the testing trials and their behavioral and acoustic measures. R was used to conduct the Pearson's pairwise correlations (of all complete pairs) and to create a correlation matrix that uses a colour gradient to indicate the direction of the correlation and the size of the spheres are used to indicate the size of the correlation. The strengths of the correlations were interpreted using the following criteria: weak ($r < .30$), moderate ($r < .50$) and strong ($r > .50$; Cohen, 1988).

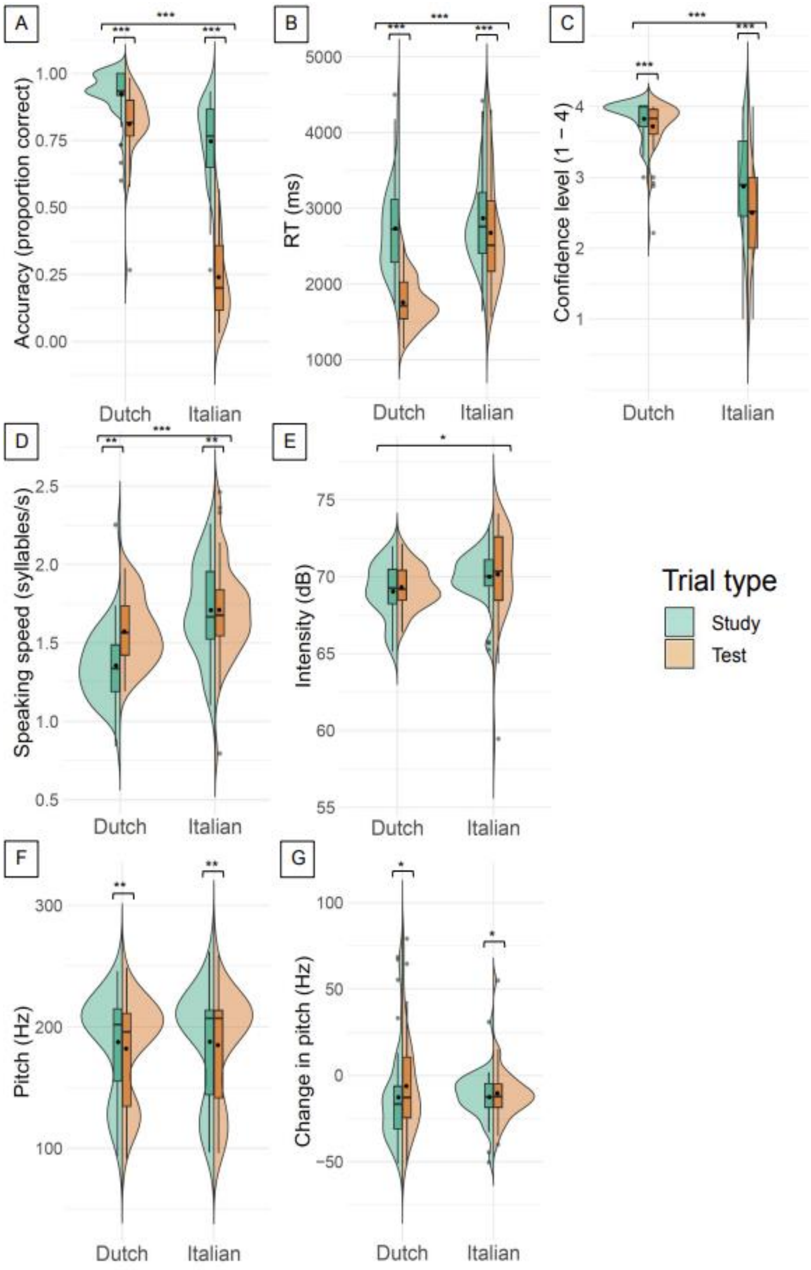
Results

Descriptive Statistics and RM-ANOVAs: Behavioural Measures and Acoustics

Figure 2 displays violin plots for the descriptive statistics of the behavioural measures (response accuracy, reaction time and subjective confidence) and acoustics (speaking speed, intensity, pitch average and pitch change) for both language conditions, across study and test trials. The Q-Q plots for the 2-way repeated measures ANOVAs indicated that the normality assumption was approximately met for all analyses.

Figure 2

Violin Plots of Descriptives Behavioural Measures and Acoustics by Language Condition and Trial Type



Note. Accuracy was calculated using correct and incorrect trials. All other variables were calculated using only correct test trials.

A: Accuracy (proportion correct), B: reaction times (ms), C: subjective confidence D: speaking speed (syl/s) E: intensity (dB), F: average pitch (Hz), G: change in pitch (Hz).

* *p* < .05. ** *p* < .01. *** *p* < .001.

Response Accuracy

For the behavioural measures, the response accuracy in the Dutch trials was high in both study trials (*M* = .92, *SD* = .09) and test trials (*M* = .81, *SD* = .13), compared to the Italian trials. These showed lower and more variable accuracy scores than the Dutch condition (see Figure 2A), and showed a bigger decrease in response accuracy between the study trials (*M* = .75, *SD* = .15) and test trials (*M* = .24, *SD* = .15), with outliers having low accuracy for Italian study and Dutch test trials. These differences seemed to depend on both the language condition (*F*(1,47) = 595.3, *p* < .001) and trial type (*F*(1,47) = 315.4, *p* < .001), with the trial type difference depending on the language condition (*F*(1,47) = 235.6, *p* < .001).

Reaction Time

The reaction time distributions (see Figure 2B) indicated relatively consistent performance for both the study trials (*M* = 2869.3, *SD* = 616.2) and test trials (*M* = 2676.3, *SD* = 722.2) for the Italian condition. In the Dutch condition the responses were visibly slower in the study trials (*M* = 2734.0, *SD* = 588.7) compared to the test trials (*M* = 1758.1, *SD* = 323.3), with one outlier having a high reaction time in both the Dutch study and test trials. The difference in how fast participants responded differed between both language condition (*F*(1,47) = 59.6, *p* < .001) and trial type (*F*(1,47) = 52.4, *p* < .001), with the trial type difference depending on the language condition (*F*(1,47) = 37.9, *p* < .001).

Subjective Confidence

Figure 2C illustrates a noticeable difference between both language conditions, with overall confidence being higher and less dispersed in the Dutch condition for both study trials (*M* = 3.83, *SD* = 0.26) and test trials (*M* = 3.72, *SD* = 0.36), with a few outliers present in both study and test trials. The Italian condition showed broad dispersion across both trial types, indicating that participants felt less confident in both the study trials (*M* = 2.87, *SD* =

0.73) and test trials (*M* = 2.5, *SD* = 0.73). The difference in how confident the participants reported to be differed between both language condition (*F*(1,47) = 204.7, *p* < .001) and trial type (*F*(1,47) = 18.1, *p* < .001), with the effect of trial type depending on the language condition (*F*(1,47) = 6.2, *p* = .020).

Speaking Speed

For the acoustics, the plots for speaking speed (see Figure 2D) indicate that the distribution for the speaking speed was nearly equal for the study trials (*M* = 1.71, *SD* = 0.30) and test trials (*M* = 1.61, *SD* = 0.31) in the Italian condition. In contrast, the average speed in both the study trials (*M* = 1.36, *SD* = 0.24) and test trials (*M* = 1.57, *SD* = 0.21) were lower in the Dutch condition compared to the Italian responses, with participants speaking faster in the test trials. There were a few outliers with high speaking speed in Dutch study trials and one with low speaking speed for the Italian test trials. The difference in the speaking speed of the participants seemed to significantly differ between both language condition (*F*(1,47) = 64.1, *p* < .001) and trial type (*F*(1,47) = 11.6, *p* = .001), with the effect of trial type depending on the language condition (*F*(1,47) = 16.8, *p* < .001).

Intensity

The intensity of the responses presented in Figure 2E show that intensity levels were similar between language conditions, with the Italian condition showing a slightly higher average intensity in both the study trials (*M* = 70.0, *SD* = 1.7) and test trials (*M* = 70.2, *SD* = 3.0) compared to the Dutch study (*M* = 69.0, *SD* = 1.83) and test trials (*M* = 69.3, *SD* = 1.4). While the dispersion was mostly similar for both conditions and trial types, the Italian condition displays a broader dispersion for the test trials. There were a few low intensity outliers in the Italian study and testing trials. The difference in the intensity levels of the responses seemed to only depend on the difference between the language conditions (*F*(1,38)

= 5.1, *p* = .030) but not on trial type (*F*(1,38) = 0.4, *p* = .557) nor an interaction effect (*F*(1,38) = 0.0, *p* = .933).

Average Pitch

The average pitch differenced minimally between trial types (see Figure 2F), with the Dutch condition showing a bit more pairwise difference between the study (*M* = 187.5, *SD* = 40.9) and test trials (*M* = 182.11, *SD* = 40.9) than for the Italian study trials (*M* = 187.8, *SD* = 45.2) and test trials (*M* = 185.0, *SD* = 44.8). The difference in the average pitch of the responses only depended on the trial type (*F*(1,38) = 8.0, *p* = .007), but not on language condition (*F*(1,38) = 0.0, *p* = .961), nor an interaction between the two (*F*(1,38) = 0.9, *p* = .341).

Change in Pitch

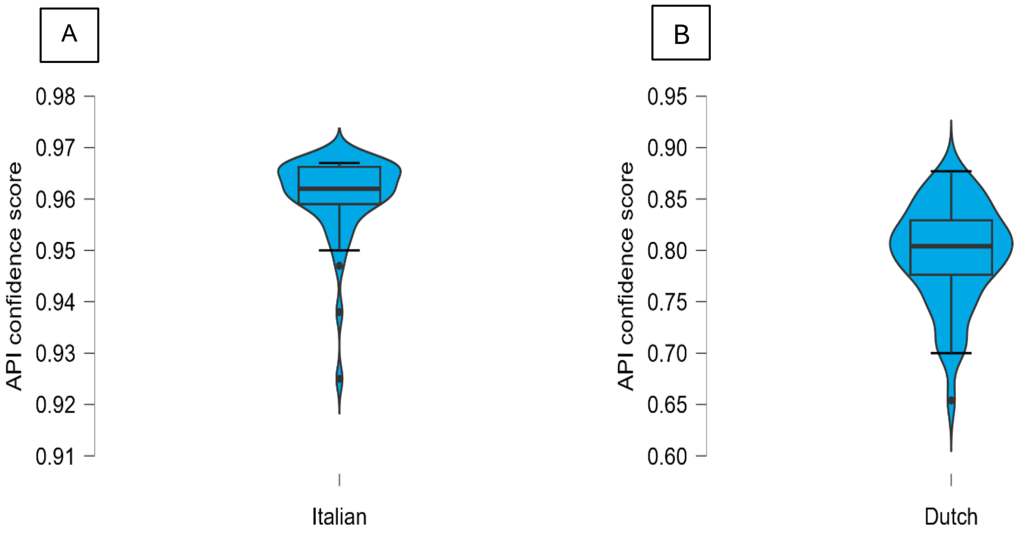
Lastly, the visualizations of the change in pitch (see Figure 2G) indicate a bigger difference and more dispersion in the Dutch condition between the study trials (*M* = -12.71, *SD* = 28.36) and test trials (*M* = -6.22, *SD* = 28.4) compared to the Italian study (*M* = -12.55, *SD* = 14.08) and test trials (*M* = -10.31, *SD* = 16.15). Outliers included multiple large changes in pitch for Dutch study and test trials and a few high and low outliers for both Italian study and test trials. The change in pitch seemed to depend on the type of trial (*F*(1,38) = 4.2, *p* = .048), but not on language condition (*F*(1,38) = 0.5, *p* = .540) or an interaction between trial type and language (*F*(1,38) = 1.4, *p* = .242).

API Transcription Confidence Across Language Conditions

To check for the consistency of the API in transcribing the responses, the average API confidence was calculated for each language condition, using the correct and incorrect responses of the study trials. Figure 3 reveals that the API was more confident and consistent in transcribing the Italian responses (*M* = .961, *SD* = .008) compared to the Dutch trials (*M* = .796, *SD* = .049). Furthermore, the confidence in the Dutch condition is far more dispersed

Figure 3

Boxplots Displaying Mean API Confidence per Language Condition



Note. Calculated using the correct and incorrect study trials.

and shows a clear outlier with an atypical value compared to the rest of the distribution, indicating a participant with a systematic low API confidence for their responses. In the Italian condition there were also a few outliers, but these were much less atypical in their deviance from the rest of the findings. This shows a systematic difference in the API's confidence between both conditions.

Associations Between Behavioural Measures and Acoustics

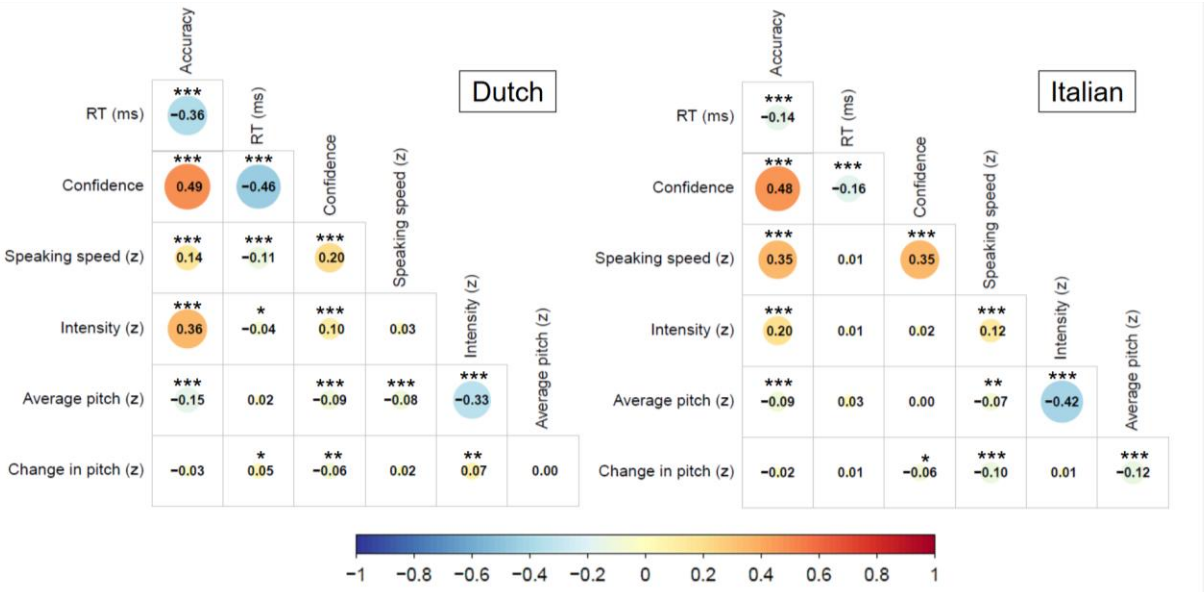
Figure 4 displays the correlation matrices of the behavioural measures (reaction time, response accuracy and subjective confidence) and the standardized acoustics (speaking speed, intensity, average pitch, and pitch change) for both language conditions.

Dutch Trials

The matrix in Figure 4 (left) displays the correlations for the dependent variables in the Dutch trials. In terms of the relationship between all the behavioural measures, reaction times were moderately negatively correlated to response accuracy and subjective confidence,

Figure 4

Correlation Matrices of Dependent Variables Separated by Language Conditions



Note. Correlations were calculated using the correct and incorrect responses from testing trials.

Acoustics (speaking speed, intensity, average pitch and change in pitch) were standardized to *z*-scores.

* *p* < .05. ** *p* < .01. *** *p* < .001.

while response accuracy and subjective confidence were moderately positively correlated. In other words, when participants were more accurate, they responded faster and were more confident about their response.

Speaking speed had a weak negative correlation with reaction time, and a weak positive correlation with response accuracy and subjective confidence. In terms of acoustics, speaking speed was weakly negatively correlated with average pitch, but not with intensity and pitch change. This means that speaking faster was associated with increased accuracy and confidence, and with higher pitch.

Intensity of the response was weakly negatively correlated with reaction times, and positively with response accuracy (moderately) and subjective confidence (weakly). Furthermore, intensity had a moderate negative correlation with the average pitch and a weak positive correlation with change in pitch, but was not correlated with speaking speed. Therefore, higher intensity was associated with higher accuracy and confidence, a higher average pitch and an increase in pitch.

The average pitch of the responses was weakly negatively correlated with response accuracy and subjective confidence, but was not correlated with reaction time. Examining the acoustics showed that average pitch was weakly negatively correlated with speaking speed and moderately negatively with intensity. It was not correlated with pitch change. A lower pitch thus indicated lower accuracy and confidence, and was related to a slower and softer response.

Lastly, change in pitch was weakly positively correlated with reaction time and weakly negatively correlated with subjective confidence, but was not correlated with response accuracy. In terms of acoustics, change in pitch was only correlated with the intensity of the response (a weak positive correlation), but not with speaking speed and average pitch. This means that a decrease in the pitch was associated with higher confidence, a quicker response and lower intensity compared to an increase in pitch.

Italian Trials

The left correlation matrix in Figure 4 displays the correlations for all the Italian trials. For the behavioural measures, the reaction times were weakly negatively correlated with the response accuracy and subjective confidence, and the response accuracy was moderately positively correlated with the subjective confidence. This means that when participants were more accurate, they responded faster and were more confident about their response.

Speaking speed was moderately positively correlated with response accuracy and subjective confidence, and weakly negatively with average pitch and pitch change. It was weakly positively correlated with intensity. Faster speaking speed was thus indicative of higher accuracy and confidence and associated with lower, less variable pitch in the response.

Intensity of the response was weakly positively correlated with response accuracy, but not with reaction time and subjective confidence, and it was moderately negatively correlated with pitch average, but not pitch change or speaking speed. This means that a louder response was associated with higher accuracy and a lower pitch.

The pitch average had a weak, negative correlation with response accuracy, but was not correlated to reaction time and subjective confidence. In terms of acoustics, pitch average was negatively associated with speaking speed (weakly), intensity (moderately) and pitch change (weakly). Responses with a higher pitch were thus indicative of lower accuracy, and were spoken slower, less loud and with less variation in pitch.

Lastly, pitch change was weakly negatively correlated with subjective confidence, speaking speed, and average pitch. This means that while a higher pitch of a response was associated with less accuracy, it was not associated with the reaction time or confidence levels.

Discussion

The first aim of this study was to investigate the relationship between prosodic speech features (PSF), the subjective confidence of participants and their response accuracy in the context of a language learning task. Specifically, we aimed to investigate this relationship for multiple-word items, given that these allow for more variation in PSF and therefore might contain more information about the subjective confidence and response accuracy of the learner. Furthermore, we investigated how the relationship between PSF, subjective confidence and response accuracy are affected by studying these items in a native language

(Dutch) or in the acquired foreign language (Italian), given that the display of subjective confidence and response accuracy in PSF might differ in native speech compared to the speech in a new language that is learned.

Based on Wilschut et al. (2025) we expected that the PSF indicative of the subjective confidence and response accuracy to be stronger related in this study. For intensity we found larger correlations for response accuracy compared to Wilschut et al. (2025), indicating that intensity of a response is also indicative of objective accuracy for multiple-word items, which aligns with the expectations based on Goupil and Aucouturier (2021) findings which suggest that intensity is an indicator of objective accuracy. Our findings for speaking speed deviated from this expectation, despite them being stronger correlated for both the subjective confidence, they showed a positive relationship. This is not in line with the findings of Wilschut et al. (2025) which showed a negative relationship with speaking speed. This might be explained by our use of subject-verb phrases instead of single words, as our findings for this relationship do align with other previous study that found that same relationship for confidence and speaking speed (Jiang & Pell, 2017). In terms of pitch change, we found a similar but smaller correlation for response accuracy, but we did not find a significant correlation for subjective confidence compared to previous findings (Goupil & Aucouturier, 2021; Wilschut et al., 2025). These differences could be due to the linguistic nature of subject-verb phrases compared to single words. Pitch segment visualizations by Wilschut et al. (2025) show that single words have a relatively stable pattern, followed by an increase or decrease depending on the participants confidence. It could be that for longer items, like subject-verb phrases, there is more opportunity for the rise and fall of pitch in the pitch signal throughout the phrase. However, with our calculation of pitch change we only included the first and last part of the pitch signal, therefore possibly not accurately measuring change in pitch in a way that may fit more with the pattern of subject-verb phrases. This could have led to these correlations to be less accurate.

In terms of the difference in the relationship of PSF, subjective confidence and response accuracy between both language conditions, we expected that these relationships would differ, in the sense that some PSF might be more predictive of subjective confidence and response accuracy in one of the languages. For both language conditions the speaking speed seemed to be a significant indicator of accuracy and subjective confidence, but this association was stronger for Italian than Dutch. Intensity was also correlated with response accuracy in both languages, but stronger for Dutch than Italian. In Dutch it was also weakly correlated with subjective confidence. Change in pitch had the same weak, negative correlation with subjective confidence in both languages. For the Italian condition, the speed at which the participants spoke their answer seemed to be more strongly related to whether their answer was correct and how confident they felt about their answer compared to Dutch. This means that the faster they spoke the more likely it was that their response was accurate and felt more confident about it. This is interesting, as this would mean that speaking speed is a better predictor for subjective confidence and response accuracy in the non-native language. However, intensity was a better predictor for accuracy and subjective confidence in the native language compared to the foreign language, while change in pitch was significantly correlated with subjective confidence in both languages. These findings show that there are differences between the language conditions, but these differences are mostly in the strengths of the correlations.

The current study has several limitations. First, we used a fixed-items schedule to present the items to the participants instead of using an ALS, as this helped to simplify the data collection and analysis. Given that the order of sentences was fixed for each practice block, this might have allowed participants to memorize the sequence of items. If this was easier in Dutch than in Italian (which may have been the case, as accuracy and confidence were much higher in the Dutch condition), this could mean that the correlations between PSF and response accuracy and subjective confidence might have in part reflected the process of memorizing a sequence of easy sentences in their native language, rather than learning to translate the sentences from a foreign language. All in all, this could have led to a ceiling effect for the Dutch accuracy and confidence scores, which can be seen in the lack of variability in the distributions of these variables. For accurate estimations of Pearson's correlation coefficient, variability in both variables is needed, as a lack of variability in a variable can lead to underestimations of the correlations (Goodwin & Leech, 2006). Thus, the correlations between PSF and accuracy and confidence for the Dutch trials may underestimate their actual relationship. Furthermore, this could make the comparison of correlations between language conditions inaccurate. To avoid this ceiling effect, future studies should ensure variability in accuracy and confidence by making the task more difficult by not presenting items in a fixed order, for example by using an ALS, which will also repeat difficult items more often, helping to increase variability.

Secondly, for our analysis we only looked at Pearson's correlations, which do not control for correlations of other variables, meaning we cannot say anything about the unique predictive value of any one PSF on subjective confidence and response accuracy. Future research could include partial correlations or linear regression analysis to control for the other variables and find the unique predictive value of each PSF.

Third, our method for calculating pitch change may have limited our ability to detect meaningful correlations between subjective confidence and response accuracy for multiple-word items. We divided the pitch signals of trials into 20 equal segments, for which we arbitrarily averaged the first five and last five segments, and calculated the difference between these points expecting to provide a useful measure of overall pitch change.

However, by using this method we did not take into account any potential fluctuations in the middle segments of the pitch signal, which might have affected the correlations we found. Future research should consider the statistical approach to calculating the change in pitch of previous research, to see if using a least squares linear regression to analyse the pitch signal might be a better way to capture the change in pitch for multiple-word items compared to our approach (Wilschut et al., 2025).

Despite the mentioned limitations and suggestions for improvements, this study has several important implications. First, these findings help deepen our understanding of how PSF can be used as additional predictors to standard learning behaviour measures (reaction time and response accuracy) for memorization of items in language learning tasks. This is interesting for the development of speech-based ALS, given that real-time extracting of PSF during learning sessions can help to optimize their prediction of memorization of items, meaning that this would improve the learning efficiency by providing learners with better item presentations for memorization. Furthermore, these PSF could be added as behavioral indicators to ALS of confidence, to replace learners having to continuously indicate their level of confidence for item adaption, enabling the incorporation of confidence as a memory performance indicator without disrupting the learning process too much. Further improving and testing these speech-based ALS is important, as it helps optimize these systems to the needs and proficiency of individual learners (in different educational contexts) and helps to make these systems available for users with certain learning impairments that would benefit from using a speech-based system over a text-based system (e.g. in case of dyslexia; Wilschut et al., 2024a). By expanding the findings from single-word (Wilschut et al., 2025) to multiple-word items, this is a first step in exploring the potential of using PSF in speech-based ALS to include longer and possibly more complex items as part of foreign language learning. Finally, these findings contribute to a better understanding of how PSF reflect subjective confidence and response accuracy in native language and foreign language in the context of language learning tasks.

Conclusion

In summary, this study found that the distinct correlations between PSF and subjective confidence and response accuracy in a language learning task for single words can mostly be translated to multiple-word items. Additionally, we found that these associations seemed to be different for studying items in a native or foreign language. Future research could design a task that is better adjusted for difficulty between language conditions and should employ more sophisticated statistical analyses to further explore the individual associations of PSF with subjective confidence and response accuracy. The findings of the current study contribute to the advancement of speech-based adaptive language learning systems, a step in making language learning more efficient and engaging using science-based technology.

References

Boersma, P. (2007). Praat: Doing phonetics by computer. http://www.praat.org/.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Psychology Press.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of r. *The Journal of Experimental Education*, *74*(3), 249–266. https://doi.org/10.3200/JEXE.74.3.249-266

Google. (n.d.). *Cloud Speech-to-Text* [Computer software]. Google Cloud. https://cloud.google.com/speech-to-text

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

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., & Aucouturier, J.-J. (2021). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature Communications*, *12*(1), 861. https://doi.org/10.1038/s41467-020-20649-4

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

Landblom, S. A., & Ionin, T. (2022). Nuclear accent placement in broad focus intransitives in native and non-native English: An investigation of syntactic and pragmatic factors. *Glossa: A Journal of General Linguistics*, *7*(1), Article 1. https://doi.org/10.16995/glossa.5810

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

Narakeet (2025). Narakeet Text to Voice Over (Version 2.43.81) [Web application]. https://www.narakeet.com/docs/text-to-voice-over/

Papoušek, J., Pelánek, R., & Stanislav, V. (2014). Adaptive practice of facts in domains with varied prior knowledge. *Proc. of Educational Data Mining. 6-13.*

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*(4), 559–586. https://doi.org/10.1207/s15516709cog0000_14

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

Reed, B. S. (2010). Analysing conversation: An introduction to prosody. Macmillan International Higher Education.

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

Sense, F., Behrens, F., Meijer, R. R., & Van Rijn, H. (2016). An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, *8*(1), 305–321. https://doi.org/10.1111/tops.12183

Sense, F., Meijer, R. R., & Van Rijn, H. (2018). Exploration of the rate of forgetting as a domain-specific individual differences measure. *Frontiers in Education*, *3*, 112. https://doi.org/10.3389/feduc.2018.00112

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

Van Rijn, D., van Maanen, L., & van Woudenberg, M. (2009). Passing the test: Improving learning gains by balancing spacing and testing effects. In A. Hoses, D. Peebles, & R.

Cooper (Eds.), *Proceedings of the 9th International Conference on Cognitive Modeling* (pp. 108-114). Article 187.

van der Velde, M., Sense, F., Borst, J., and Van Rijn, H. (2021). Alleviating the cold start problem in adaptive learning using data-driven difficulty estimates. *Comput. Brain Behav. 4*, 231–249. https://doi.org/10.1007/s42113-021-00101-6

Wahyuningsih, L., Huwaidah, H. K., Maryam, F. F. D., & Arochman, T. (2023). Learners strategies used by non-English department students in learning English: Students' perspective. *KABASTRA: Kajian Bahasa Dan Sastra*, *3*(1), 12–22. https://doi.org/10.31002/kabastra.v3i1.1083

Wennerstrom, A. (2001). The music of everyday speech: Prosody and discourse analysis. Oxford University Press. https://doi.org/10.1093/oso/9780195143218.001.0001

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*, 780131. https://doi.org/10.3389/frai.2021.780131

Wilschut, T., Sense, F., Scharenborg, O., & van Rijn, H. (2022). Beyond responding fast or slow: Improving cognitive models of memory retrieval using prosodic speech features. *International Conference on Cognitive Modeling 2022.*

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 - 24th International Conference, AIED 2023, Proceedings* (pp. 255-266). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 13916 LNAI). Springer. https://doi.org/10.1007/978-3-031-36272-9_21

Wilschut, T., Sense, F., & Van Rijn, H. (2024a). Modality matters: Evidence for the benefits of speech-based adaptive retrieval practice in learners with dyslexia. *Topics in Cognitive Science*, *17*(1), 57–72. https://doi.org/10.1111/tops.12769

Wilschut, T., Sense, F., & Van Rijn, H. (2024b). Speaking to remember: Model-based adaptive vocabulary learning using automatic speech recognition. *Computer Speech & Language*, *84*, 101578. https://doi.org/10.1016/j.csl.2023.101578

Wilschut, T., Sense, F. & van Rijn, H. (2025). 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), 85–115. https://doi.org/10.20396/joss.v1i1.15014

Appendix A

Table 1

List of the 40 Generated Simple Subject-Verb Phrases in Dutch and Italian

ID	Dutch phrase	# of Syllables	Translation (English)	Italian phrase	# of Syllables
1	De nicht glimlacht	4	The niece smiles	La nipote sorride	7
2	De baby eet	4	The baby eats	Il bambino mangia	6
3	De kikker springt	4	The frog jumps	La rana salta	5
4	Het meisje leest	4	The girl reads	La ragazza legge*	6
5	De ober kookt	4	The waiter cooks	Il cameriere cucina*	8
6	De broer rijdt	3	The brother drives	Il fratello guida*	6
7	De trein wacht	3	The train waits	Il treno aspetta*	6
8	De zus loopt	3	The sister walks	La sorella cammina*	7
9	De hond speelt	3	The dog plays	Il cane gioca*	5
10	De schilder tekent	5	The painter draws	Il pittore disegna*	7
11	De nacht begint	4	The night starts	La notte inizia*	6
12	De brandweerman spreekt	5	The firefighter speaks	Il pompiere parla*	6
13	De jongen rent	4	The boy runs	Il ragazzo corre*	6
14	De man liegt	3	The man lies	L’uomo mente*	4
15	Het dier reist	3	The animal travels	L’animale viaggia*	6
16	De vrouw komt	3	The woman comes	La donna viene*	5

17	De oma bakt	4	The grandma bakes	La nonna inforna*	6
18	De prijs verandert	5	The price changes	Il prezzo cambia	5
19	De deur kraakt	3	The door creaks	La porta scricchiola	6
20	Het bot breekt	3	The bone breaks	L’osso si rompe	5
21	De klok tikt	3	The clock ticks	L'orologio ticchetta	7
22	De boom bloeit	3	The tree blooms	L’albero fiorisce	6
23	De dochter schreeuwt	4	The daughter shouts	La figlia grida	5
24	De vriend lacht	3	The friend laughs	L’amico ride	5
25	De vogel zingt	4	The bird sings	L’uccello canta	5
26	De groep wint	3	The group wins	Il gruppo vince	5
27	De haai zwemt	3	The shark swims	Lo squalo nuota	5
28	De zon schijnt	3	The sun shines	Il sole splende	5
29	De familie betaalt	6	The family pays	La famiglia paga	6
30	De leeuw slaapt	3	The lion sleeps	Il leone dorme	6
31	De maan bestaat	4	The moon exists	La luna esiste	6
32	De stoel valt	3	The chair falls	La sedia cade	5
33	De muis verdwijnt	4	The mouse disappears	Il topo scompare	6
34	De kat snurkt	3	The cat snores	Il gatto russa	5

35	De buurman begrijpt	5	The neighbour understands	Il vicino capisce	7
36	Het hout brandt	3	The wood burns	Il legno brucia	5
37	De plant groeit	3	The plant grows	La pianta cresce	5
38	De docent vermenigvuldigd	8	The teacher multiplies	L'insegnante moltiplica	8
39	Het vat explodeert	5	The barrel explodes	Il barile esplode*	7
40	Het schip zinkt	3	The ship sinks	La nave affonda*	6

Note. * indicates phrases for which a male voice was used.