Voice and choice: Investigating the role of prosodic variation in request compliance and perceived politeness using conversational TTS

Éva Székely[1], Jeff Higginbotham[2], Francesco Possemato[2,3]

[1]Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden
[2]Department of Communicative Disorders and Sciences, University at Buffalo, NY, USA
[3]Centre for Language and Cognition, Rijksuniversiteit Groningen, The Netherlands
szekely@kth.se, cdsjeff@buffalo.edu, f.possemato@rug.nl

Abstract

As conversational Text-to-Speech (TTS) technologies become increasingly realistic and expressive, understanding the impact of prosodic variation on speech perception and social dynamics is crucial for enhancing conversational systems. This study explores the influence of prosodic features on listener responses to indirect requests using a specifically designed conversational TTS engine capable of controlling prosody, and generating speech across three different speaker profiles: female, male, and gender-ambiguous. We conducted two experiments to analyse how naturalistic variations in speech rate and vocal effort impact the likelihood of request compliance and perceived politeness. In the first experiment, we examined how prosodic modifications affect the perception of politeness in permission- and action requests. In the second experiment participants compared pairs of spoken requests, each rendered with different prosodic features, and chose which they were more likely to grant. Results indicate that both faster speech rate and higher vocal effort increased the willingness to comply, though the extent of this influence varied by speaker gender. Higher vocal effort in action requests increases the chance of being granted more than in permission requests. Politeness has a demonstrated positive impact on the likelihood of requests being granted, this effect is stronger for the male voice compared to female and gender-ambiguous voices.

1 Introduction

The importance of pragmatics in the development of conversational technologies is becoming increasingly critical (Levinson, 2024). As Text-to-Speech (TTS) systems achieve greater realism in speech generation, a significant gap persists in understanding the pragmatic effects these technologies have within interactions. By modeling prosodic features based on empirical research, conversational TTS can be made more engaging and effective in a variety of interactive contexts. The necessity for human-oriented pragmatics in these systems is particularly evident in scenarios requiring compliance to requests. Moreover, understanding the subtleties of how politeness is conveyed in language and speech is crucial for comprehending the factors influencing request compliance, especially in conversational systems that utilise TTS.

People use a variety of spoken strategies to manage the potential threats that communication can pose to their own and others' self-esteem and autonomy (Brown and Levinson, 1987). Indirect requests are employed to mitigate face-threatening acts, demonstrate respect for the listener's autonomy, and maintain the fundamentally cooperative and prosocial nature of human communicative behaviour and respect between interlocutors (Rossi et al., 2023). They are a fundamental aspect of polite discourse, reflecting the speaker's sensitivity to social dynamics and the listener's ability to interpret and respond to nuanced communicative cues (Drew and Couper-Kuhlen, 2014).

Initiating actions, such as requests, can be seen as a basic form of social coercion (Enfield et al., 2019). Requests have a bearing on the sequential organisation of the ensuing talk, while also restricting the agency of the requestee, and even threatening their autonomy (Soubki and Rambow, 2024). The linguistic structure of requests influences politeness and compliance. For example, the choice between using an imperative form, which might seem direct and blunt, and opting for a more conditional or interrogative form can alter the level of imposition perceived by the interlocutor. Chalfoun et al. (2024) emphasise the strategic use of politeness markers like 'please' in everyday requests, demonstrating how these markers are employed to manage face-threats in ill-fitted interactional contexts, particularly when requests could be seen as intrusive or when they encounter resistance from the requestee. Research on modal constructions

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 466–476
September 18–20, 2024. ©2024 Association for Computational Linguistics

in requests reveals further nuances in how these requests are framed and understood in different contexts (Steensig and Heinemann, 2014). Modal verbs like ‘could’ or ‘might’ introduce a level of uncertainty or optionality into the request, thereby softening it and enhancing its politeness.

Enfield (2014) highlights the importance of the “infrastructure” that underpins requests, which includes the social and interactional contexts influencing how requests are made and received. Understanding the interplay between prosodic features and sociolinguistic norms is essential for designing effective conversational agents.

In our study, we develop and employ a prosody-controllable gender-ambiguous TTS system as a research tool to conduct controlled experiments assessing the role of prosodic variation in request compliance. This approach allows us to isolate the impact of vocal traits from gender biases without relying on the ability of voice actors to consistently reproduce prosodic variations.

The key contributions of this research are the following: We pioneer the use of a gender-ambiguous neural TTS built on spontaneous speech in perceptual studies, which allows for an unprecedented exploration of how gender perception influences listener responses to prosodic variations. Moreover, this study provides empirical evidence on how natural variations in speech rate and vocal energy influence the listener’s perception of politeness, and their responsiveness to indirect requests. Our findings illustrate that the impact of prosodic variations can differ based on the speaker’s gender profile, contributing to a more tailored approach in the design of TTS systems to accommodate diverse user interactions.

2 Background

2.1 Prosody in social signaling

Prosody contributes significantly to signaling speaker attitudes and interpersonal stances (Ward, 2019). Various aspects of stance can be predicted from prosodic features with significant accuracy beyond mere chance (Ward et al., 2017). Politeness strategies and their impact on compliance are not only influenced by the linguistic content but are also significantly modulated by prosodic features such as intonation, pitch, and speech rate. Research indicates that variations in these prosodic features can critically affect listeners’ perceptions and their subsequent responses to requests (Kendrick and Drew, 2014). Trott et al. (2023) explore how prosodic features help in disambiguating English indirect requests, highlighting the complex interplay between acoustic signals and intended meanings in speech. They find that prosodic cues such as duration, pitch, and pitch slope significantly correlate with a speaker’s intent, influencing how listeners interpret pragmatically ambiguous utterances.

Vergis and Pell (2020) explore the effects of linguistic structure, imposition, and prosody on the perception of politeness in requests. Their findings show that prosody significantly affects politeness ratings, with prosodic features e.g. appropriate intonation and pitch enhancing perceived politeness. The study highlights that not only the content but the manner of speech delivery plays a critical role in social interactions. Similarly, Caballero et al. (2018) examine the acoustic cues of politeness, demonstrating that prosodic variations such as changes in pitch, intonation, and speech rate are essential for conveying politeness. Specifically, they found that higher pitch, increased pitch range, and a melodic intonation contour are perceived as more polite, whereas rude request displayed slower speech rate, lower pitch and tended to fall in pitch. Their analysis of verbal requests shows that while a specific prosody of politeness may not exist, these features significantly influence how politeness is perceived, with certain prosodic patterns leading to higher politeness ratings. Gryllia et al. (2018) investigated the role of pragmatics and politeness in prosodic variability in Greek wh-questions. Their study showed that context and social factors, such as the power and solidarity between interlocutors, influence prosodic patterns. The findings suggest that prosodic modifications are not merely stylistic but are pragmatically motivated to achieve desired social outcomes, such as politeness or authority.

2.2 Spontaneous TTS as a research tool

Voice talents can effectively use prosodic cues to convey subtle pragmatic nuances across various speech acts (Hellbernd and Sammler, 2016). At the same time, the reliance on actors to generate experimental stimuli introduces variability, as personal interpretations of how specific utterances should be delivered may differ. While analysing speech patterns in corpora of ecologically valid, spontaneous speech data avoids this bias, this approach often lacks the necessary control over linguistic content and prosodic realisations needed for conducting rigorously controlled experiments.

An emerging alternative methodology is to use state-of-the-art TTS built on spontaneous speech data to create experimental stimuli. This method combines advantages from both traditional approaches, relying on the authenticity of natural speech for modelling, and providing the controllability required for experimental rigor. Several previous works employed prosody-controllable neural TTS as a research tool in controlled listening experiments, with the aim to discover new knowledge about various aspects of speech perception. Székely et al. (2017) investigated the interaction of vocal effort and hesitation disfluencies in synthesised speech, focusing on how these factors influence the perception of uncertainty. Székely et al. (2019) discovered using spontaneous TTS that filled pauses improved the perception of speaker authenticity and engagement. Elmers et al. (2023) looked into the perceptual impact of tongue clicks using neural TTS, revealing that their inclusion can alter perceived speaker confidence. O'Mahony et al. (2024) extends this methodology with a corpus-based approach to investigate the prosody and pragmatic functions of the discourse marker "well".

As the capabilities and controllability of spontaneous TTS systems continue to evolve, this methodology is gaining increased attention for its potential to uncover new insights into how subtle signals in speech are interpreted by listeners. These detailed insights increase our understanding about speech perception in general, and they are particularly applicable in dialogue systems and Augmentative Communication Technologies (ACT), since these applications directly employ TTS.

In the current study, we use spontaneous conversational TTS as a research tool, and we advance this methodology by training a prosody-controllable multi-speaker TTS system that can generate male, female, and perceptually gender-ambiguous (Sutton, 2020) TTS. Using this we investigate the impact of prosodic features on listener's willingness to comply with an indirect request.

3 Overview of the method

In this study, we develop and utilise a multi-speaker TTS model built upon two corpora of spontaneous speech, to investigate the impact of prosodic variations on listener responses to indirect requests.

TTS model development: A spontaneous conversational TTS system is engineered to include gender-ambiguous voice capabilities, using a modified Tacotron 2 architecture (Shen et al., 2018; Székely et al., 2023b) that allows for the dynamic control of prosodic features at the utterance level. This setup enables exploring how different prosodic renderings affect listener perceptions.

Stimuli design and synthesis: Stimuli for the experiments were designed to directly address our research questions regarding the social dynamics of request-making in conversation. Using an interactive interface, we synthesised these stimuli, ensuring each varied systematically in prosody according to predefined settings. This approach allowed precise control over the acoustic and prosodic variables of interest.

Verification of stimuli: Before deployment in experiments, all stimuli are tested for naturalness, gender ambiguity, and the presence of significant acoustic-prosodic differences across conditions, using objective measures. These verifications support the reliability and validity of the stimuli used in the subsequent online listening tests.

Experimental setup: The experiments are conducted as online listening tests on a crowd-sourcing platform, where participants are presented with synthesised speech samples. This method facilitates the collection of data on how listeners perceive and react to variations in speech delivery within an imagined conversational context.

4 Text-to-Speech Synthesis

4.1 Corpora

Two corpora of spontaneous conversational speech were used to build the TTS model. The first is a multimodal multi-party dataset called AptSpeech, described in Kontogiorgos et al. (2018). This dataset comprises 15 multi-party interactions involving a single moderator, a male speaker of General American English, and two distinct participants per session, engaged in a collaborative task. The speech data from the moderator was used to create a TTS corpus, along with additional recordings of reading newspaper articles and the Arctic sentences (Kominek and Black, 2004). The complete corpus has a duration of approximately 8 hours: 2h 26min of reading and 5h 40min of spontaneous speech. The second corpus was created from 14h 43min of conversational podcast recordings of a female speaker of General American English, who consented to make the recordings available for TTS research purposes. The speaker supplemented the material with 1h 52 min of reading non-fiction.

Both corpora were segmented into breath groups (stretches of speech delineated by two breath events) using the method proposed by Székely et al. (2020). Automatic Speech Recognition (ASR) was used to transcribe the utterances. The transcriptions were annotated for spontaneous speech events, such as filled pauses, breathing and repetitions, as well as turn-internal pauses and turn endings, in order to be able to produce these behaviours at synthesis. To balance the corpora, the number of breath groups from each style per speaker was set to the minimum of the two speakers, 480 breath groups of read speech and 2788 breath groups of spontaneous speech.

4.2 Prosody-controllable gender-ambiguous conversational TTS

We developed a prosody-controllable multi-speaker TTS model specifically for the purpose of the experiment. Our method follows (Székely et al., 2023a) closely, with the main difference being the use of spontaneous conversational speech corpora. We use a modified Tacotron 2 (Shen et al., 2018) TTS architecture which allows for features appended to the encoder output which can be controlled on a gradient at synthesis time (Székely et al., 2023a). The publicly available pre-trained gender-ambiguous model trained on 20 hours of speech data[1] is used as base, from which the existing speaker embedding is dropped and training is reinitialised as a warm start with the new spontaneous corpora. The prosodic features speech rate and energy were added to the encoder, using values normalised across the corpora. In this architecture, the appended features can be controlled at synthesis time on utterance- or phrase level. Speech rate is calculated as syllables per second, including silences, which is different from articulation rate, which excludes silences. As a result, slower speech rate values at synthesis time result in insertion of longer pauses as well as a slowed down articulation rate. Energy is calculated as the Root Mean Square of power of the signal, using a window of 20ms and step size of 5ms (Suni et al., 2017). Because energy is an acoustic feature that correlates with other prosodic features in spontaneous conversational speech, increasing the energy feature of the TTS results in a natural increase of pitch and articulatory effort, as well as more pronounced emphasis patterns. This results in a prosodic rendering that perceptually translates to a quality which can be described as "speaking up". This combination of prosodic features is sometimes referred to as upgraded, salient, or marked prosody (Selting, 1996) in the fields of Conversation Analysis (Sidnell and Stivers, 2013) and Interactional Linguistics (Couper-Kuhlen and Selting, 2018). In order to be descriptive without implying a direct emotional connotation, we will be referring to the prosodic realisation that this method of feature control in spontaneous neural TTS produces through the modification of the energy input feature as a level of *vocal effort*.

The model was trained for 45k iterations on 4GPUs (batch size 28). The speech signal is decoded from the output using the neural vocoder HiFi-GAN (Kong et al., 2020). The model published model by the authors is fine-tuned on the combined corpora for 180k iterations on 4GPUs (batch size 28).

5 Experiments

5.1 Hypotheses

The present study aims to systematically investigate the interactions between prosodic features, perceived politeness, request compliance and gender by positing several hypotheses:

Prosodic variation hypothesis (H1): Changes in prosodic features speech rate and vocal effort influence the perceived politeness and compliance rates of requests. Faster speech rate and higher vocal effort are predicted to enhance perceived politeness and likelihood of compliance.

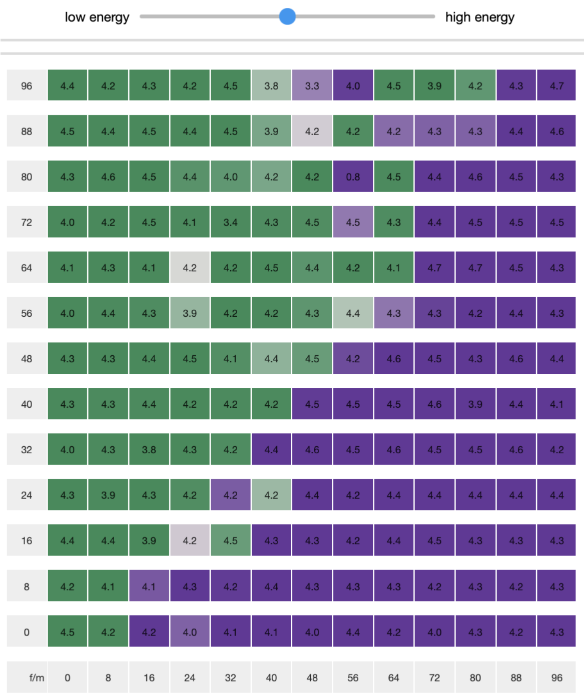
Gender perception hypothesis (H2): The gender perception of a TTS voice (male, female, or gender-ambiguous) mediates the effect of prosodic variations on politeness and compliance.

Request type hypothesis (H3): Listener responses are influenced by the interaction between prosodic features and the type of request: permission versus action requests.

5.2 Stimuli

We designed 8 indirect requests that are formulated to be considerate of the listener's capacity to grant them, possibly at a minor inconvenience, yet also allowing room for a polite refusal. All sentences are similar in length and they all contain politeness markers that express a positive face and attempt to mitigate the controlling the threat to autonomy expressed by the utterance (e.g.: *would you mind*, *can I please*). Whilst slight differences

[1] https://github.com/evaszekely/ambiguous



96	4.4	4.2	4.3	4.2	4.5	3.8	3.3	4.0	4.5	3.9	4.2	4.3	4.7
88	4.5	4.4	4.5	4.4	4.5	3.9	4.2	4.2	4.2	4.3	4.3	4.4	4.6
80	4.3	4.6	4.5	4.4	4.0	4.2	4.2	0.8	4.5	4.4	4.6	4.5	4.3
72	4.0	4.2	4.5	4.1	3.4	4.3	4.5	4.5	4.3	4.4	4.5	4.5	4.5
64	4.1	4.3	4.1	4.2	4.2	4.5	4.4	4.2	4.1	4.7	4.7	4.5	4.3
56	4.0	4.4	4.3	3.9	4.2	4.2	4.3	4.4	4.3	4.3	4.2	4.4	4.3
48	4.3	4.3	4.4	4.5	4.1	4.4	4.5	4.2	4.6	4.5	4.3	4.6	4.4
40	4.3	4.3	4.4	4.2	4.2	4.2	4.5	4.5	4.5	4.6	3.9	4.4	4.1
32	4.0	4.3	3.8	4.3	4.2	4.4	4.6	4.5	4.6	4.5	4.5	4.6	4.2
24	4.3	3.9	4.3	4.2	4.2	4.2	4.4	4.2	4.4	4.4	4.4	4.4	4.4
16	4.4	4.4	3.9	4.2	4.5	4.3	4.3	4.2	4.4	4.5	4.3	4.3	4.3
8	4.2	4.1	4.1	4.3	4.2	4.4	4.3	4.3	4.3	4.2	4.3	4.2	4.3
0	4.5	4.2	4.2	4.0	4.1	4.1	4.0	4.4	4.2	4.0	4.3	4.2	4.3
f/m	0	8	16	24	32	40	48	56	64	72	80	88	96

Figure 1: An example of the So-to-Speak interface, adapted to display gender-ambiguity ratings on a color gradient (green = female, purple = male) and automatic evaluation of naturalness MOS on a scale of 1-5. On the axes x and y, the percentage of respectively the male and female speaker embedding input is displayed. Samples play upon clicking on a cell. Moving the bar on top updates the grid to display a new set of samples corresponding to the setting. Step size and feature range are adjustable. In this example, speech rate is set to fast, speaker embedding and energy input vary.

in the verbal formulations might impact the willingness of the listener to comply, the decision to vary the types of politeness formulas is driven by our goal to avoid experimental monotony and to realistically reflect the variation in everyday speech. We included two types of requests, 4 sentences in each: *Permission Requests*: requests that seek authorisation or consent to perform an action. *Action Requests*: requests that involve asking for a particular service to be provided. The sentences included in the experiment are listed in Table 1.

Stimuli were created using an adaptation of the So-to-Speak interface (Székely et al., 2023b), which is an open source exploratory platform designed to help researchers interact with multi-dimensionally controllable TTS systems[2]. The interface enables the synthesis and playback of hundreds of samples simultaneously, displayed on an interactive grid, varying both low level prosodic features and high level style controls. Automatic estimates of naturalness Mean Opinion Scores (MOS) (Huang et al., 2022) are presented for each sample. For this work, we created an adaptation of the interface where the output of an automatic gender classifier (Rizhinashvili et al., 2022) is displayed on a color gradient. Figure 1 shows an example of the interface.

Table 1: Sentences synthesised for the experiment

Permission Requests
Is it ok if I switch off these lights? *Would you mind if I opened the windows?* *Do you mind if I adjust the thermostat?* *Can I please take down these posters?*
Action Requests
Would you please turn down the music? *Would you mind changing the channel?* *Could you please turn off the air conditioning?* *Do you mind closing the curtains?*

For the speech rate feature, two settings were chosen: *fast* and *normal*, using normalised speech rate values from the corpora, where *fast* is defined as 2 std higher than the mean. To create stimuli displaying different levels of vocal effort, three settings, *low*, *medium* and *high* were synthesised, using normalised values of the energy feature, as described in Section 4. Note that because of the corpus-driven approach to prosody control, these input features impact other characteristics in the speech samples as well, such as the increased presence of reduced articulation in fast speech rate, and higher f0 and more pronounced emphasis patterns in high energy settings. For the three gender types, three different speaker embeddings were used as input to the multi-speaker model: *male*, *female*, and *gender-ambiguous*, amounting to a total of 144 speech samples.

To verify the stimuli, we used three objective measures: Gender-ambiguity was automatically evaluated using a gender-classifier first described by Rizhinashvili et al. (2022) and adapted by Székely et al. (2023a) trained on the LibriTTS dataset (Zen et al., 2019). Additionally, all samples were evaluated through an automatic MOS rating introduced by Huang et al. (2022) which has been shown to correlate highly with perceptual ratings of naturalness. This test ensured that all stimuli had a minimum of 4.5 MOS rating. A third test was

[2] https://github.com/evaszekely/So_To_Speak

carried out to ensure that speech rate and energy features are indeed significantly different across conditions. This test was deemed necessary because of the way the prosody-control in the TTS architecture is designed, the features are not explicitly modified, rather they are an input to the TTS model. For this reason, we validate with acoustic measurements on the samples, that the output of the TTS reflected the change in input values. The value ranges measured on speech rate and energy when these are varied in the inputs for the different settings are significantly different. This was further confirmed by a series of one-sided paired t-tests over stimuli between each combination of settings (all p<0.01). The measurements of speech rate, energy and f0 are found in Appendix A. Note that these values are specific to the individual speakers' own register, as represented in the training data. As such, they should not be used as independent references. The audio samples are available online[3].

5.3 Experiment 1: politeness ratings

In order to not prime participants into specific behavior patterns regarding politeness and compliance, we conducted separate experiments concerning these two aspects. The first experiment was specifically designed to investigate how variations in prosody influence perceptions of politeness. This experiment sought to isolate politeness as variable, assessing its effects through a structured rating system. Stimuli were presented one per trial, listeners were asked to rate *"How polite does this request sound to you?"* on a scale of 1-5 (where 1 = very impolite, 2 = impolite, 3 = neutral, 4 = polite, 5 = very polite). To avoid any bias that might arise from participants recognising the experiment's focus on potential gender differences, each gender type (male, female, and gender-ambiguous) was rated by a separate group.

5.4 Experiment 2: request compliance

The goal of this experiment was to determine how different levels of speech rate and vocal effort (increased pitch and energy) contribute to conveying a tone of voice which makes it more likely to result in an indirect request being granted. The situational context presented to the participants was the following: *"Imagine that each request causes you a minor inconvenience—for example, if asked to open a window, consider that you are already feeling a bit cold and would prefer it closed. However, depending on how the request is conveyed, you might be more inclined to accommodate the speaker if it seems particularly important to them."* Participants were presented with pairs of stimuli where each pair consisted of the same request rendered with different prosodic features. The task required participants to listen to each pair and decide based on the tone, which version they were more likely to grant: ((a) = **A** much more likely, (b) = **A** more likely, (c) = both equally likely, (d) = **B** more likely, (e) = **B** much more likely). The pairwise design was chosen for 2 main reasons: firstly, to mitigate the effect of the differences of lexical content and formulations and topics among the individual sentences.

To gain further insight into what aspects of the speech samples people considered important while listening, at the end of the experiment, participants were asked the question: *"Could you tell us what helped you make your decisions?"*. The same between-subjects design was used as in Experiment 1: a different group of participants was recruited for each gender type.

6 Results

6.1 Experiment 1: politeness ratings

This listening test was completed by 90 people, 30 in each gender condition. The experiment took on average 10 minutes to complete and participants were paid £12 per hour. Everyone was asked to confirm that they were using headphones or earphones while listening to the stimuli. Participants' age ranged between 23 and 69, 45 identified as female and 45 as male.

A linear regression analysis was performed to evaluate the influence of speech rate, vocal effort, type of request, and gender on the average politeness rating of the various speech stimuli. For this analysis, the request type was coded 0 for permission requests and 1 for action requests. Gender also received an ordinal coding as this reflects the way

Table 2: Regression analysis of factors affecting politeness ratings

Variable	Coefficient	Std. Error	t-value	P-value
Constant	3.2705	0.050	65.729	<0.001
VoiceGender	0.1212	0.035	3.444	0.001
RequestType	-0.3950	0.057	-6.875	<0.001
SpeechRate	0.3184	0.057	5.542	<0.001
VocalEffort	0.4390	0.035	12.479	<0.001

[3] https://www.speech.kth.se/tts-demos/sigdial2024-request

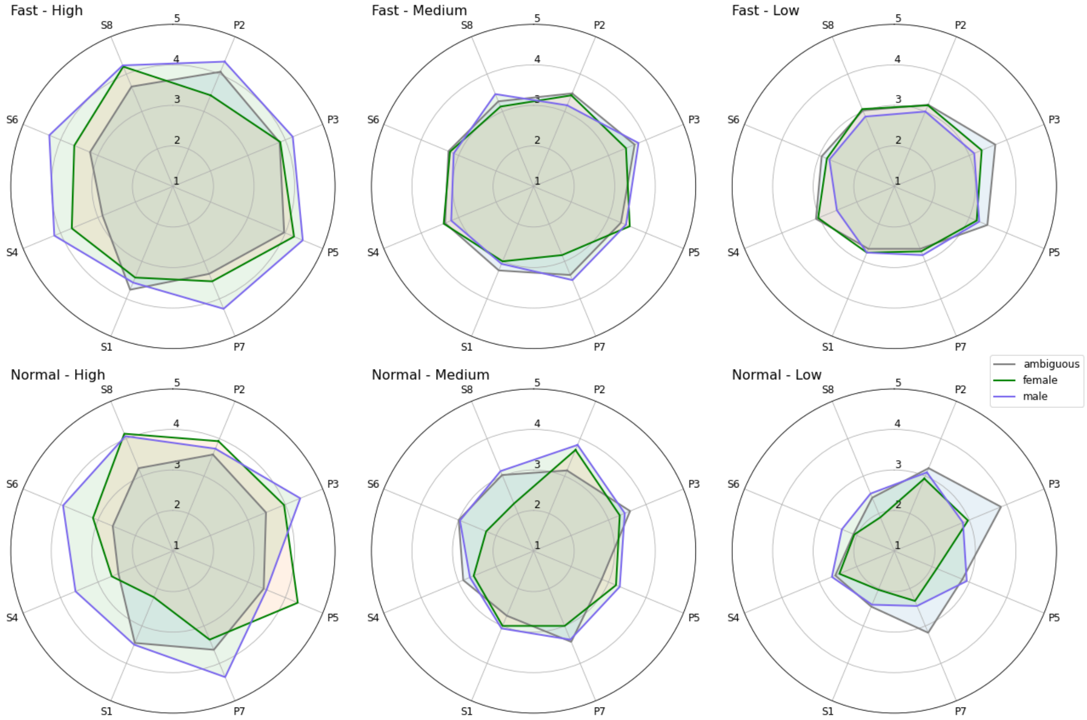


Figure 2: Results of the politeness rating per stimulus. Stimulus names on the left starting with **P**, indicate *permission requests*, and stimulus names on the right, starting with **S** indicate *action requests*. Ratings range from the center of the circle (1 = very impolite) to the edge of the circle (5 = very polite).

the TTS voices were generated: -1 for female, 0 for ambiguous and 1 for male. The results, included in Table 2 show that all four variables are significant in explaining the variance in average politeness rating. The model, based on R-squared explains 63.9% of the variance in average politeness ratings between the stimuli.

Results of this rating experiment are illustrated in Figure 2. Speech rate, vocal effort and gender have an increasing impact on perceived politeness. Action requests are rated significantly less polite than permission requests. The result that lower pitch and slower speeech rate is considered less polite confirms the findings of Caballero et al. (2018).

6.2 Experiment 2: request compliance

90 native speakers of English, recruited through the Prolific[4] platform completed the study. Participants' age ranged between 23 and 75, 42 identified as female and 48 as male. Recruitment of participants followed the same setup as in Experiment 1. The experiment took on average 18 minutes to complete. Results of the test are in Figure 3 as the proportion of the pairs in which the stimulus for that condition was preferred over another, excluding no-preference cases. Confidence intervals are calculated based on the standard error of the proportion of preferences and then applying the normal distribution's critical value to get 95% confidence intervals. We evaluated the effect of differences in speech rate and vocal effort between the two samples for each voice on the individual preference results, also controlling for participant gender. Linear regression models were applied separately for each voice. Results in Table 3 show that increases in both speech rate and vocal effort had a significant and consistently positive impact on the preference rating. The gender of the participants did not significantly influence preference for any of the voices.

Answers to the follow-up question about what helped listeners make their decision revealed that participants' decisions were influenced by their perceptions of politeness in the spoken requests. Several participants indicated that a friendlier or gentler tone made them more inclined to grant the requests, whereas harsh or demanding tones tended to deter compliance. This feedback highlights that, alongside the prosodic features conveying request importance to the speaker, the perceived politeness - or lack thereof - conveyed through prosody is a secondary factor in participants' decision-making.

[4] app.prolific.com

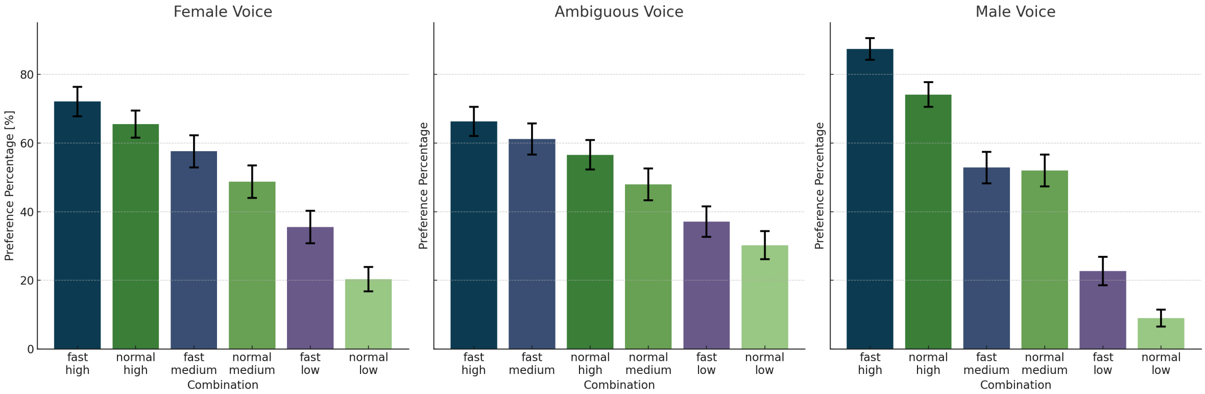


Figure 3: Percentage of evaluations where a stimulus with given speech rate: *fast* (blues), *normal* (greens)) and vocal effort: *high, medium, low* (darker to lighter shades) was preferred over other combinations, excluding ties.

Table 3: Regression Analysis Results for Compliance

Variable	**Coeff.**	**P-Value**	**95% CI**
	Female		
Intercept	-0.0325	0.522	[-0.132, 0.067]
Δ Speech Rate	0.2596	<0.001	[0.143, 0.376]
Δ Vocal Effort	0.3991	<0.001	[0.354, 0.444]
Participant	0.0569	0.322	[-0.056, 0.170]
	Ambiguous		
Intercept	-0.1206	0.046	[-0.239, -0.002]
Δ Speech Rate	0.3815	<0.001	[0.251, 0.512]
Δ Vocal Effort	0.3216	<0.001	[0.270, 0.373]
Participant	-0.0451	0.473	[-0.168, 0.078]
	Male		
Intercept	-0.0076	0.867	[-0.097, 0.082]
Δ Speech Rate	0.2480	<0.001	[0.148, 0.348]
Δ Vocal Effort	0.6213	<0.001	[0.583, 0.660]
Participant	-0.0028	0.954	[-0.098, 0.092]

6.3 Influence of perceived politeness on request compliance

Combining findings from both experiments we can examine the impact of perceived politeness differences between paired stimuli on request compliance. The average perceived politeness for each sample from Experiment 1 is introduced as an explanatory variable in the analysis of the results of Experiment 2.

The significant results of the ordinal logistic regression explaining compliance, considering main and interaction effects are presented in Table 4. Controlling for the difference in perceived politeness between the stimuli, increasing speech rate or vocal effort still have a significant positive impact on the likelihood that a request is granted. On their own, the request type and gender of the voice do not show a significant effect. For action requests, increases in vocal effort are more effective, while the effect of politeness is more limited.

Table 4: Significant Effects in the Combined Analysis

Variable	**Coeff.**	**P-value**	**95% CI**
Δ Speech Rate	0.2764	<0.001	[0.135, 0.418]
Δ Vocal Effort	0.3224	<0.001	[0.205, 0.440]
Δ Politeness	0.8803	<0.001	[0.671, 1.090]
Δ Proj.* Request	0.1738	0.008	[0.045, 0.303]
Δ Pol. * Request	-0.3904	<0.001	[-0.605, -0.176]

7 Discussion

One of the limitations in our study is that our experiments utilised only one voice per gender. To enhance the generalisability of our findings, future work will explore the use of voice conversion technologies to create a wider variety of stimuli across different gender profiles. Additionally, while this study primarily focused on prosodic features such as pitch, energy and speech rate, there are numerous other features in request articulation that warrant exploration. These include voice quality, placement of emphasis, the strategic use of pauses, and utterance-final intonation which can influence the perception of requests. Moreover, as Levinson (2024) points out, utterances are unlikely to be action-determinate by virtue of their form alone. Experiment 2 addresses this to an extent by presenting listeners with an imaginary scenario, but it is important to acknowledge the inherent limitations of controlled listening experiments in simulating the complex dynamics of wider social contexts. Consequently, the findings from this experiment should be further evaluated in more realistic, interactive scenarios where deeper contextual embeddings can be implemented.

Reflecting the rate of technological advancements, we expect to see an increasing demand for personalised, conversational TTS to represent and display the identity of individuals in real and virtual environments. One group, individuals with disabilities (e.g., cerebral palsy, autism, adult-onset disorders) who rely on computer-based Augmentative Communication Technologies, already use TTS to engage in real-time spoken conversations. The lack of pragmatically appropriate and effective TTS to accomplish various conversational tasks, including indirect requests, is a common critique of commercial ACTs. Our findings specifically show that adjustments to prosodic features such as speech rate and vocal effort significantly impact the perceived politeness of requests, and also affect compliance rates. This is particularly important for the design of TTS systems in ACTs, where effectively conveying requests in a polite manner is essential for users with communication challenges. Our hope is that incorporating these insights, developers can better equip conversational systems to meet the varied communication demands of individuals, ensuring more respectful and successful interactions across both real and virtual settings.

8 Conclusions

This study has demonstrated that the perception of politeness significantly enhances the likelihood of requests being granted. The effectiveness of changing politeness through prosody is stronger for the male voice compared to female and gender-ambiguous voices. Additionally, higher vocal effort in action requests significantly increases the chances of compliance, more so than in permission requests. This highlights the significant role that prosodic manipulation of TTS can play in enhancing the effectiveness of communicative acts within spoken dialogue systems to accommodate diverse user interactions more effectively.

Acknowledgments

This work was supported by This research was supported by the Swedish Research Council project Perception of speaker stance (VR-2020-02396), the Riksbankens Jubileumsfond project CAPTivating (P20-0298), the Engelke Family Foundation and National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR grant number 90DPCP0007).

References

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Jonathan A. Caballero, Nikos Vergis, Xiaoming Jiang, and Marc D. Pell. 2018. The sound of im/politeness. *Speech Communication*, 101:14–27.

Andrew Chalfoun, Giovanni Rossi, and Tanya Stivers. 2024. The magic word? face-work and the functions of please in everyday requests. *Social Psychology Quarterly*, page 01902725241245141.

Elizabeth Couper-Kuhlen and Margret Selting. 2018. *Interactional linguistics: Studying language in social interaction*. Cambridge University Press.

Paul Drew and Elizabeth Couper-Kuhlen. 2014. Requesting – from speech act to recruitment. In Paul Drew and Elizabeth Couper-Kuhlen, editors, *Requesting in social interaction*, pages 1–34. John Benjamins Publishing Company.

Mikey Elmers, Johannah O'Mahony, and Éva Székely. 2023. Synthesis after a couple PINTs: Investigating the role of pause-internal phonetic particles in speech synthesis and perception. In *Proc. Interspeech*, pages 4843–4847.

N. J. Enfield. 2014. Human agency and the infrastructure for requests. In Paul Drew and Elizabeth Couper-Kuhlen, editors, *Requesting in social interaction*, pages 35–54. John Benjamins Publishing Company.

Nicholas J Enfield, Tanya Stivers, Penelope Brown, Christina Englert, Katariina Harjunpää, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Tiina Keisanen, Mirka Rauniomaa, et al. 2019. Polar answers. *Journal of Linguistics*, 55(2):277–304.

Stella Gryllia, Mary Baltazani, and Amalia Arvaniti. 2018. The role of pragmatics and politeness in explaining prosodic variability. In *Proc. Speech Prosody*, pages 158–162. Speech Prosody Special Interest Group.

Nele Hellbernd and Daniela Sammler. 2016. Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, 88:70–86.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2022. The VoiceMOS Challenge 2022. In *Proc. Interspeech*, pages 4536–4540.

Kobin H. Kendrick and Paul Drew. 2014. The putative preference for offers over requests. In Paul Drew and Elizabeth Couper-Kuhlen, editors, *Requesting in social interaction*, pages 87–114. John Benjamins Publishing Company.

John Kominek and Alan W Black. 2004. The CMU arctic speech databases. In *Proc. SSW*, pages 223–224.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

Dimosthenis Kontogiorgos, Vanya Avramova, Simon Alexanderson, Patrik Jonell, Catharine Oertel, Jonas Beskow, Gabriel Skantze, and Joakim Gustafson. 2018. A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *Proc. LREC*, pages 119–127.

Stephen C Levinson. 2024. The dark matter of pragmatics: Known unknowns. *Elements in Pragmatics*.

Victoria S McKenna and Cara E Stepp. 2018. The relationship between acoustical and perceptual measures of vocal effort. *The Journal of the Acoustical Society of America*, 144(3):1643–1658.

Johannah O'Mahony, Catherine Lai, and Éva Székely. 2024. "Well", what can you do with messy data? Exploring the prosody and pragmatic function of the discourse marker "well" with found data and speech synthesis. In *Proc. Interspeech*.

Davit Rizhinashvili, Abdallah Hussein Sham, and Gholamreza Anbarjafari. 2022. Gender neutralisation for unbiased speech synthesising. *Electronics*, 11(10):1594.

G. Rossi, M. Dingemanse, S. Floyd, et al. 2023. Shared cross-cultural principles underlie human prosocial behavior at the smallest scale. *Scientific Reports*, 13:6057.

Margret Selting. 1996. Prosody as an activity-type distinctive cue in conversation: The case of so-called 'astonished'. *Prosody in conversation: Interactional studies*, 12:231.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *Proc. ICASSP*, pages 4779–4783.

Jack Sidnell and Tanya Stivers. 2013. *The handbook of conversation analysis*. John Wiley & Sons.

Adil Soubki and Owen Rambow. 2024. Intention and face in dialog. In *Proc. LREC-COLING*, pages 9143–9153.

Jakob Steensig and Trine Heinemann. 2014. The social and moral work of modal constructions in granting remote requests. In Paul Drew and Elizabeth Couper-Kuhlen, editors, *Requesting in social interaction*, pages 145–170. John Benjamins Publishing Company.

Antti Suni, Juraj Šimko, Daniel Aalto, and Martti Vainio. 2017. Hierarchical representation and estimation of prosody using continuous wavelet transform. *Computer Speech & Language*, 45:123–136.

Selina Jeanne Sutton. 2020. Gender ambiguous, not genderless: Designing gender in voice user interfaces (VUIs) with sensitivity. In *Proc. CUI*, pages 1–8.

Éva Székely, Joakim Gustafson, and Ilaria Torre. 2023a. Prosody-controllable gender-ambiguous speech synthesis: a tool for investigating implicit bias in speech perception. In *Proc. Interspeech*, pages 1234–1238.

Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. 2019. Spontaneous conversational speech synthesis from found data. In *Proc. Interspeech*, pages 4435–4439.

Éva Székely, Gustav Eje Henter, Jonas Beskow, and Joakim Gustafson. 2020. Breathing and speech planning in spontaneous speech synthesis. In *Proc. ICASSP*, pages 7649–7653.

Éva Székely, Joseph Mendelson, and Joakim Gustafson. 2017. Synthesising uncertainty: The interplay of vocal effort and hesitation disfluencies. In *Proc. Interspeech*, pages 804–808.

Éva Székely, Siyang Wang, and Joakim Gustafson. 2023b. So-to-Speak: an exploratory platform for investigating the interplay between style and prosody in tts. In *Proc. Interspeech*, pages 2016–2017.

Sean Trott, Stefanie Reed, Dan Kaliblotzky, Victor Ferreira, and Benjamin Bergen. 2023. The role of prosody in disambiguating English indirect requests. *Language and Speech*, 66(1):118–142.

Nikos Vergis and Marc D Pell. 2020. Factors in the perception of speaker politeness: The effect of linguistic structure, imposition and prosody. *Journal of Politeness Research*, 16(1):45–84.

Nigel G Ward. 2019. *Prosodic patterns in English conversation*. Cambridge University Press.

Nigel G Ward, Jason C Carlson, Olac Fuentes, Diego Castan, Elizabeth Shriberg, and Andreas Tsiartas. 2017. Inferring stance from prosody. In *Proc. Interspeech*, pages 1447–1451.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A corpus derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech*.

A Acoustic measurements on the experimental stimuli

In the tables the output ranges are recorded of the measured speech rate, energy and f0 levels for the different input settings for speech rate and vocal effort respectively. Energy and f0 are selected as acoustic measurements as these have been demonstrated as significant predictors of listeners' perception of vocal effort (McKenna and Stepp, 2018). ANOVA tests were performed for each voice to validate the statistical difference of the output measurements for the different levels of input setting.

Table 5: Speech Rate Range (syl/s) by Voice and Input Speech Rate Level with ANOVA p-values

voice	female	ambiguous	male
fast	4.99 - 6.77	5.30 - 7.07	5.38 - 7.29
normal	3.83 - 5.15	4.15 - 5.25	4.25 - 6.00
p-value	*0.0*	*0.0*	*0.0*

Table 6: f0 Range (Hz) by Voice and Input Energy Level with ANOVA p-values

voice	female	ambiguous	male
high	212 - 269	177 - 261	112 - 159
medium	162 - 207	139 - 178	95 - 112
low	123 - 153	104 - 137	82 - 90
p-value	*0.0*	*0.0*	*0.0*

Table 7: Energy Range (RMS power) by Voice and Input Energy Level with ANOVA p-values

voice	female	ambiguous	male
high	0.071 - 0.124	0.070 - 0.119	0.058 - 0.091
medium	0.070 - 0.127	0.067 - 0.103	0.050 - 0.080
low	0.059 - 0.094	0.057 - 0.084	0.045 - 0.070
p-value	*0.00217*	*0.00034*	*0.00001*