Can OpenAI's TTS model convey information status using intonation like humans?

Na Hu, Jiseung Kim, Riccardo Orrico, Stella Gryllia, Amalia Arvaniti

Radboud University, the Netherlands

na.hu@ru.nl, jiseung.kim@ru.nl, riccardo.orrico@ru.nl, stella.gryllia@ru.nl, amalia.arvaniti@ru.nl

Abstract

Chatbots powered by Large Language Models (LLMs) such as OpenAI's ChatGPT have demonstrated impressive capabilities in understanding and generating text and their potential applications in humanities research have been extensively explored. Recently, OpenAI launched its first Text-To-Speech (TTS) model, which has demonstrated the ability to convert text into highly realistic speech. This opens up various potential applications for prosodic research. However, before such applications are in place, a systematic evaluation is needed to determine the extent to which the synthesized speech resembles human speech in terms of prosody. This study aims to contribute to this endeavor by comparing how information status is conveyed by intonation in British English speech synthesized using OpenAI's TTS model to the speech produced by native speakers of the same English variety. Through Functional Principal Component Analysis (FPCA) and statistical modelling, we found that OpenAI's TTS model can generate F0 contours with various shapes. However, the F0 contours generated by OpenAI's TTS model conveying information structure differ from those produced by the human speakers. This indicates that the speech generated by OpenAI's TTS model may not be ready for use in prosody research, yet.

Index Terms: synthesized speech, OpenAI, intonation, information status

1. Introduction

Large Language Models (LLMs), such as OpenAI's GPTs, Google's Bard, and Meta's LLaMa, are massive statistical models that generate contextually coherent texts that resemble human-produced texts in response to human inquiries. In essence, LLMs are machine learning algorithms with complex architectures and a vast number of parameters trained on substantial amounts of data. GPT-3 developed by OpenAI, for example, is a transformer-based generative pre-trained language model with 175 billion parameters trained on vast amounts of text sourced from the internet [1].

Amongst the various LLMs available on the market, the GPT model family has demonstrated remarkable capabilities in comprehending and generating texts as well as human-like language processing patterns across multiple language processing tasks originally designed for human participants such as syntactic ambiguity resolution [2]. These language abilities demonstrated by the GPT models have led to a surging number of applications in several research fields, including pedagogy, medical research, psychology, and linguistics (see [3] for a comprehensive summary). In the field of psychology, LLMs are used as a window to query people's mental models of themselves and their environment (social and cognitive psychology), infer individual differences in coping strategies (personality psychology), and help people reappraise stressful experiences (affective and clinical psychology); see [4] for a review.

At DevDay on November 6, 2023, OpenAI launched its first Text-To-Speech (TTS) model (https://openai.com/blog/new-models-and-developer-products-announced-at-devday), which converts text into speech. According to user reactions on the internet (e.g., https://community.openai.com/t/tts-api-service-usability/482163/7), the English speech generated by this model is highly realistic (listen to the demo here: https://platform.openai.com/docs/guides/text-to-speech).
Similar to models generating text, the performance of TTS models depends on system architecture and the amount of training data. As of the writing of this paper, OpenAI has not publicly disclosed the specific architecture or detailed information regarding the training data for their TTS model. What we do know is that a typical TTS system comprises two key components: a front-end responsible for analyzing the linguistic structure of input text, and a back-end that generates the actual speech signal. The back-end is commonly trained on vast amounts of real human speech data to establish connections between linguistic and phonetic information and corresponding acoustic features of speech. With a sophisticated model structure and large training data, it is reasonable to expect the back-end model to capture subtle nuances in human speech, producing speech with prosody that faithfully replicates that produced by an "average" human. Such a model can serve various purposes in prosodic research, including generating speech stimuli for perception experiments, creating prompts for production experiments, and simulating speech data for power analysis. Most intriguingly of all, such a TTS model could provide a window to understanding how humans convey information through prosody. Researchers can use synthesized speech directly as speech samples in production studies to explore prosodic variations across different experimental conditions. This prospect is potentially valuable for research on speech production, particularly considering the commonly shared experience that recruiting and testing human participants is both costly and time-consuming. While the potential for LLM-powered phonetic research is appealing, it is important to note that LLMs still have issues such as hallucinations, the use of private data for training, and their role in perpetuating biases, etc. (see [5] for detailed information).

Prosody plays a crucial role in conveying information about the linguistic context of utterances across various linguistic levels (see [6], [7] for extensive reviews). Intonation, in particular, can be used to signal information structure. For example, in English, new information, i.e., information that should be added to the common ground, is typically conveyed

using high pitch (H* pitch accent, in autosegmental terms) on relevant items in the utterance [8], [9]. Though this much is clear, there is a longstanding debate in English regarding the mapping between prosody and information status. One viewpoint posits that while items that represent new information are H* accented, as mentioned, items presenting contrastive or corrective information are realized with an F0 rise, denoted as L+H*. However, evidence from both production and perception does not support such a one-to-one mapping. For example, it is found in [10] that H* is compatible with both new and contrastive contexts. It is reported in [11] that L+H* is relatively infrequent and used not only with contrastive topics or foci, but also with non-contrastive items, though less frequently.

While additional studies are needed for a better understanding of how information status is conveyed through prosody, this study is not designed to contribute to this goal. Instead, our aim is to compare information status as conveyed through F0 in synthesized British English speech to that in speech produced by native Standard British English (SBE) speakers. To achieve this, speech produced by human speakers and that synthesized using OpenAI's TTS endpoint was analyzed following the procedures normally used to analyze human speech, including F0 curve modelling using Functional Principal Component Analysis (FPCA) followed by statistical analysis of the PC coefficients.

2. Methodology

2.1. Human speech data

We obtained unscripted speech from 8 native speakers of Southern British English (5 females, 3 males, age 18-54, mean age 29.25) using three tasks: a storytelling task, a map task, and an informal discussion about a set of unusual objects.

The recorded speech data were annotated for pitch accents and information status separately. Pitch accent annotations were performed by an expert annotator based solely on F0 shape, resulting in 2,450 instances of H* or L+H*. A second expert annotator independently annotated 8% of the accents. High inter-annotator agreement was observed (Unweighted Cohen's Kappa = 0.84, C.I. = 0.79 – 0.89). Information status annotations relied solely on orthographic transcripts: items were marked as *corrective* if explicitly correcting a previously mentioned item, and as *contrastive* if part of an implicit set of alternatives. The resulting pragmatic labels were then paired with the accents, with each accented item labelled as *corrective* or *contrastive* if marked as such in the orthographic transcript, and as *new* otherwise.

2.2. Synthesized speech data

The input texts used for speech synthesis were 108 question-answer pairs, where the questions were designed to elicit answers evenly distributed across the same three pragmatics conditions as present in the human speech data: new, contrastive, or corrective information (see Table 1 for examples). The answers were composed of either single words or two content words. The former had only one accent, the target of analysis (in italics in Table 1); the latter had two pitch accents, with the first word carrying a prenuclear accent and the second word carrying the target accent (in italics in Table 1). Thus, all the accents under investigation were utterance-final nuclear accents.

Table 1: *Example dialogues.*

	Single words	**Compounds**
Non-contrastive (new)	Q: What's your first name? A: *Mary.*	Q: What is the fruit on the table? A: A yellow *melon*.
Contrastive	Q: Who should I call first, Mary or Lana? A: *Mary.*	Q: Did you buy a yellow melon or a yellow pumpkin? A: A yellow *melon*.
Corrective	Q: Are you friends with Miriam? A: *Mary.*	Q: Is that a yellow pumpkin? A: A yellow *melon*.

The question-answer dialogues shown in Table 1 (excluding the "Q" and "A" identifiers before the sentences) were converted to audio using OpenAI's TTS API in Python, following the tutorial provided at https://platform.openai.com/docs/quickstart?context=python and https://platform.openai.com/docs/guides/text-to-speech. Specifically, we used the `tts-1-hd` model, which is optimized for quality. Another option is `tts-1`, which is optimized for speed and more suitable for real-time synthesis tasks. The `tts-1-hd` model offers six preset voices, including three male and two female voices with an American English accent and one male voice ("`Fable`") with a British English accent. We used the voice of `Fable`. The input texts contained no marks that could affect the resulting prosodic markup, such as capitalization, italics, or exclamation marks. The output audio files were saved to the local disk in .mp3 format. Despite its lossy nature, the .mp3 format has no impact on F0 [12], the measurement that the present study focused on.

2.3. F0 curve modelling

The F0 contours of the target accents in both the human speech data and the synthesized speech data were modeled in the following steps. First, F0 values (in Hz) were extracted at a time-step of 5 ms using PRAAT [13]. Second, the raw F0 curves were processed to remove F0 doubling and halving and were interpolated to fill F0 gaps. Third, the resulting curves were normalized by speaker using z-scores to eliminate individual speaker characteristics. Forth, the normalized F0 curves were modelled using FPCA in one analysis, following standard procedures outlined in [14]. In essence, FPCA models input curves using principal component curves, turning each input curve into a mathematical function of time:

$$f(t) \approx \mu(t) + s1 \times PC1(t) + s2 \times PC2(t) + \ \ldots \quad (1)$$

In eq. (1), $f(t)$ represents the modelled F0 curve, $\mu(t)$ is the mean curve of all input curves, $PC1(t)$, $PC2(t)$, etc. are the principal component curves, each representing a dominant mode of variation across curves, and $s1$, $s2$, etc. are the coefficients associated with the respective PCs, reflecting the contribution of the respective PCs to the resulting curve's shape. Each input curve has its unique set of PC coefficients, thereby characterizing the raw curves. The PC coefficients were used as dependent variables in statistical analysis.

2.4. Statistical analysis

To investigate potential differences in F0 contour shapes across different information statuses between human and synthesized speech, we fitted Bayesian mixed-effect models in R [15] using

the `Brms` package [16], a wrapper package for the probabilistic programming language `Stan` [17]. The model included PC coefficients (`PC1`, `PC2`, etc.) as dependent variables. The constant effects were `pragmatics` with three levels: non-contrastive ("`NC`") as the reference level, contrastive ("`CT`"), and corrective ("`CR`"), `source` with two levels: human ("`H`") and synthesized speech (referred to as machine, "`M`"), and the interaction between `pragmatics` and `source`. The model's random structure included `item` with fixed slopes and `speaker` with varying-slope for `pragmatics`. The models were fitted using 4 chains, 10,000 iterations each, including 4,000 warm-ups, and uninformative priors with a normal distribution with a mean of 0 and a standard deviation of 5.

For the effects of interest, we report the mean and the lower and upper bounds of 95% credible interval (95% CrI) of the posterior probability distribution. If the 95% CrI excludes zero, indicating that zero is an implausible value for the model parameter given the data, it indicates that the probability that the effect of interest is present is 95%. Conversely, if the interval includes zero, the effect of interest is likely to be absent.

3. Results

The FPCA of the F0 curves obtained from the human speakers and the TTS model shows that the first three PCs explained 96.4% of the variance among the curves (PC1: 75.1%; PC2: 15.3%; PC3: 6.0%). Our analysis focuses on those. The effect of each PC on the mean curve is visualized in Figure 1 by applying different scores of the PC to the mean curve (solid black line). The scores range from +1 to -1 standard deviation of the PC's coefficient in increments of 0.25. As shown in the figure, PC1 controls scaling, with larger weights resulting in curves with higher scaling. PC2 influences the slope of the curve, with larger weights reducing the steepness of the F0 fall. PC3 modifies the curve's curvature, with smaller weights resulting in a convex shape with a F0 peak.

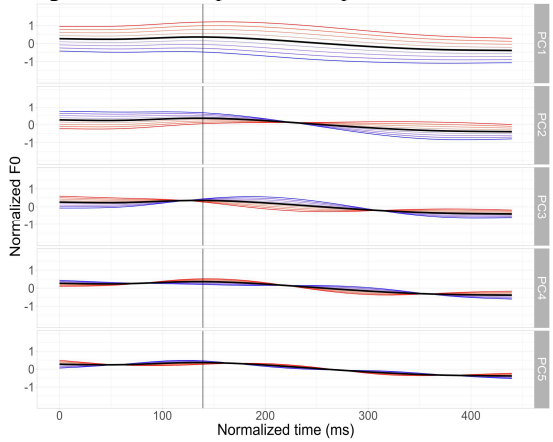

Figure 1: *Color-coded curves illustrate the effect of each PC on the mean curve (solid black line). The vertical line indicates the onset of the accented vowel.*

Descriptive statistics on the sub-dataset containing only data for the machine speech suggest variations in the PC coefficients among the input curves (PC1: range [-21.5, 41.8], median = 0.94, SD = 11.7; PC2: range [-29.7, 9.2], median = -7.5, SD = 6.7; PC3: range [-22.4, 8.7], median = -4.2, SD = 4.9). This suggests that the TTS model is capable of generating F0 contours with different shapes. Figure 2 presents 20 randomly selected F0 contours from the synthesized speech subset. As shown in the figure, some curves, e.g., 70, 100, and 158, feature a falling shape (represented as H*), whereas others, e.g., 144, 286, and 301, feature a rise-fall shape (L+H*).

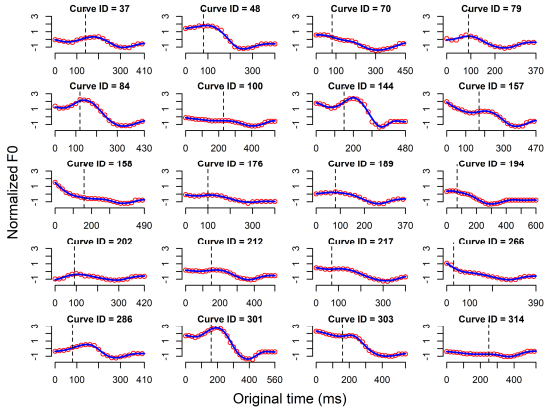

Figure 2: *20 randomly selected F0 contours from the data set containing only the machine speech. Red circle: raw F0 points at times; blue solid line: smoothed curve; vertical dash line: stressed vowel onset.*

Pairwise comparisons shown in Figure 3 reveal that, for the speech produced by human speakers, there was a 95% probability that F0 curves in the corrective condition had lower PC1 and PC3 coefficients, suggesting lower scaling and a more convex shape, compared to those in the non-contrastive condition. Moreover, F0 curves in the contrastive condition had lower PC3 coefficients, hence a more convex shape, compared to the curves in the non-contrastive condition. Furthermore, F0 curves in the contrastive and corrective conditions mainly differed in terms of PC1, with the curves in the contrastive condition having higher scaling than those in the corrective condition. As for the synthesized speech, there is no discerning evidence suggesting a difference in any of the PCs between the three pragmatic conditions, as the 95% CrIs include 0 as a plausible value.

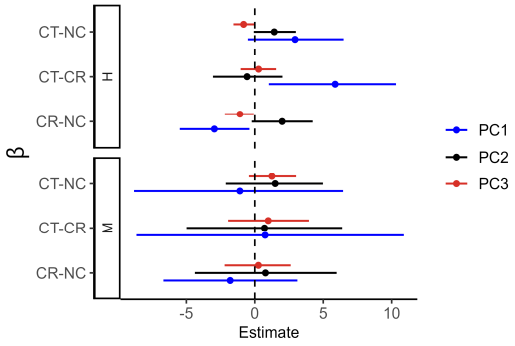

Figure 3: *95% CrIs of the posterior probability distributions of the slope parameter β for the effect of pragmatics on the three PCs.*

Figure 4 presents the estimated difference between human and machine in each PC in each pragmatic condition. It shows that the curves obtained from the two sources in the non-contrastive condition differ in all three PCs. The curve generated by the machine has higher scaling (PC1), steeper slope (PC2), and a more convex shape (PC3), compare to that produced by human speakers. In the contrastive and corrective conditions, the

curves generated by the machine differ from those produced by human speakers mainly in terms of slope, having a steeper slope compared to those produced by human speakers. The above-mentioned shape differences can be best appreciated by referring to the F0 curves reconstructed using Equation 1 involving the PC coefficients derived from the Bayesian mixed-effect model (Figure 5).

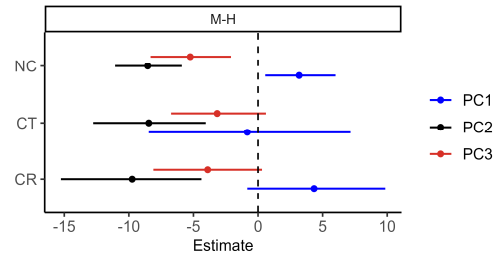

Figure 4: *95% CrI of the estimated difference between machine and human (M-H) in each PC in each pragmatic condition.*

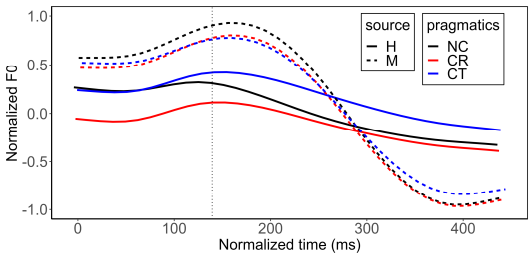

Figure 5: *F0 contours in the three pragmatic conditions for human and machine reconstructed using eq. (1) involving PCs 1, 2, and 3 and the estimates of the Bayesian mixed-effect model. The vertical dotted line indicates the onset of the accented vowel.*

4. Discussion

In this study, we compared how information status is conveyed by intonation in British English utterances synthesized using OpenAI's TTS model and utterances produced by native speakers of this variety. Our analysis revealed that while the TTS model can generate F0 contours with diverse shapes, these contours differ from those produced by humans across all three pragmatic conditions. Although the text used for speech synthesis differed from what human speakers produced, this discrepancy is not a concern, as FPCA captures only the dominant variance among F0 curves. Considering the human data (2,450 accents produced by 8 speakers) is unscripted and sizeable by the standards of phonetic research, we are inclined to say that the F0 patterns found in human speech more truthfully reflect how information status is conveyed in unscripted British English than what was revealed from the synthesized speech. Future research should explore the relevance of the observed differences between human and AI-generated speech for communication. However, it is clear that the differences between our expectations and the AI-generated F0 contours were audible to the authors of the present study, and in some instances, the differences were meaningful. Further discussion on this topic follows below.

The reason why the F0 contour shapes generated by the TTS model deviate from those produced by humans in the pragmatic conditions is not entirely clear due to the lack of transparency in the workings of complex machine learning models. There are two plausible explanations. One possibility is that the labeling of information status in the training data of the TTS model is not consistent, considering that the concept has different definitions and categories across linguistic theories, potentially leading to confusion. Another plausible explanation is that the accent in the synthesized two-word utterances is not always on the expected item. Visual and auditory inspection of the TTS model output indicates that in roughly 5% of the utterances, the focus is placed on the first word instead of the second, which is deaccented. This phenomenon applies particularly in the non-contrastive condition, though this accentuation pattern is not attested in British English: accenting only *yellow* in phrases like *a yellow melon* automatically leads to a narrow interpretation, viz. *a yellow, not a green, melon*. Considering that the sentences only contained two words, one can imagine that for sentences with more complex structures the variability in focus location may well increase.

Although the speech synthesized by OpenAI's current TTS model may not fully replicate human speech in terms of conveying information status, it would be premature to dismiss the usefulness of synthesized speech in prosodic research. It remains necessary to examine the system's prosodic performance in other linguistic aspects. Additionally, other LLM-based TTS systems such as those introduced by SpeechLab (https://www.speechlab.ai/) and Microsoft (https://learn.microsoft.com/en-us/azure/ai-services/speech-service/text-to-speech) all demonstrate a sufficient level of competence in generating natural-sounding speech. Evaluating the prosodic performance of these systems would provide valuable insights into the current state of TTS development.

This study may also contribute to approaches for evaluating TTS systems. Traditionally, TTS system performance is evaluated using mainly two approaches: subjective assessments, e.g., ratings based on factors such as naturalness, clarity, and quality (e.g., Mean Opinion Score (MOS)); objective assessments, such as Mel Cepstral Distortion (MCD), which measures how well the synthesized speech matches the natural speech in terms of spectral characteristics (See [18] for a detailed review). While these approaches help engineers understand overall system performance, they do not offer insights into specific prosodic nuances. In this study, we adopted the method normally used in the study of human speech, involving controlled testing materials and F0 modelling techniques. Using this approach to evaluate TTS speech could provide an understanding of the system's performance in specific prosodic aspects.

5. Conclusions

Using a well-tested approach in prosodic research, we found that OpenAI's TTS model (`tts-1-hd`) is capable of producing a variety of natural sounding F0 contour shapes in British English utterances. However, the contours do not match those produced by human speakers of the same English variety when it comes to encoding information structure differences.

6. Acknowledgements

The work presented in this paper has received financial support from the European Research Council (grant no. ERC-ADG-835263) awarded to Amalia Arvaniti.

7. References

[1] T. B. Brown *et al.*, "Language models are few-shot learners," 2020, [Online]. Available: arXiv:2005.14165

[2] Z. G. Cai, X. Duan, D. A. Haslett, S. Wang, and M. J. Pickering, "Do large language models resemble humans in language use?," 2024, [Online]. Available: arXiv:2303.08014

[3] Y. Liu *et al.*, "Summary of ChatGPT-Related research and perspective towards the future of large language models," *Meta-Radiology*, vol. 1, no. 2, 2023, doi: 10.1016/j.metrad.2023.100017.

[4] D. Demszky, D. Yang, D. S. Yeager, and et al., "Using large language models in psychology," *Nature Reviews Psychology*, vol. 2, pp. 688–701, 2023, doi: https://doi.org/10.1038/s44159-023-00241-5.

[5] OpenAI, "GPT-4 Technical Report," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.08774

[6] J. Cole, "Prosody in context: a review," *Lang Cogn Neurosci*, vol. 30, no. 1–2, pp. 1–31, 2015, doi: 10.1080/23273798.2014.963130.

[7] M. Wagner and D. G. Watson, "Experimental and theoretical advances in prosody: A review," *Lang Cogn Process*, vol. 25, no. 7–9, pp. 905–945, 2010, doi: 10.1080/01690961003589492.

[8] J. Pierrehumbert, "The phonology and phonetics of English intonation," PhD dissertation; Massachusetts Institute of Technology, 1980.

[9] D. R. Ladd, *Intonational Phonology*. 2008. doi: 10.1017/cbo9780511808814.

[10] D. G. Watson, M. K. Tanenhaus, and C. A. Gunlogson, "Interpreting pitch accents in online comprehension: H* vs. L+H*," *Cogn Sci*, vol. 32, no. 7, pp. 1232–1244, 2008, doi: 10.1080/03640210802138755.

[11] N. Hedberg and J. M. Sosa, "The Prosody of Topic and Focus in Spontaneous English Dialogue," in *Topic and Focus*, 2006. doi: 10.1007/978-1-4020-4796-1_6.

[12] R. Fuchs and O. Maxwell, "The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range," in *Proceedings of the International Conference on Speech Prosody*, 2016. doi: 10.21437/speechprosody.2016-107.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]. Version 6.0.43," retrieved 8 September 2018.

[14] M. Gubian, F. Torreira, and L. Boves, "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts," *J Phon*, vol. 49, pp. 16–40, 2015, doi: 10.1016/j.wocn.2014.10.001.

[15] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.r-project.org/

[16] P. C. Bürkner, "brms: An R package for Bayesian multilevel models using Stan," *J Stat Softw*, vol. 80, 2017, doi: 10.18637/jss.v080.i01.

[17] B. Carpenter *et al.*, "Stan: A probabilistic programming language," *J Stat Softw*, vol. 76, no. 1, pp. 1–32, 2017, doi: 10.18637/jss.v076.i01.

[18] P. Wagner *et al.*, "Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program," 2019. doi: 10.21437/ssw.2019-19.