Creakiness, Breathiness, and Nasality Contribute to the Perceived Suitability of Synthesized Speech in a Pragmatically-Rich Domain

Harm Lameris[1,2], *Nigel G. Ward*[1]

[1]University of Texas at El Paso, USA
[2]Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
lameris@kth.se, nigelward@acm.org

Abstract

We aim to improve the suitability of speech synthesis output for applications that are situated, embodied, and/or involve rich user interaction. For such purposes, better control of prosody is a priority. Basic research on prosody has found that voice quality features, notably creakiness and breathiness, and also probably nasality, play central roles in conveying various pragmatic functions. This paper investigates the extent to which proper control of these three feature can improve the perceived suitability of synthesized speech. Participants used the voice conversion tool VoiceQualityVC to make fine-grained adjustments to parameters affecting perceived voice quality and nasality. Working with utterances taken from a corpus of collaborative gameplay, they were able to modify synthesized speech to better match how they thought it should sound. A subsequent perception experiment showed that these adjusted utterances were rated as more suitable than the baseline. These findings demonstrate both the potential value and the feasibility of exploiting more prosody-related parameters in speech synthesis. Samples can be found at www.cs.utep.edu/nigel/lameris.
Index Terms: voice quality, pragmatics, voice conversion, human in the loop, text-to-speech, speech editing

1. Introduction

Increasingly, AI systems and robots are being viewed as potential partners or teammates to interact with, rather than mere appliances to control. While speech that is merely intelligible and natural is adequate for many current applications, collaborative and embodied interactions require more [1, 2]. Future systems will increasingly need to signal complex intents, guide users in real time, show awareness of the environment and situation, mark how the information in an utterance connects to other information and to the action plan, coordinate joint actions and turn-taking, express assessments, attitudes, and stances, and so on. Until they can do such things, AI systems will never be fully trusted or acceptable [3, 4, 5, 6].

Prosody is known to have an important role in conveying many such pragmatic functions [7], and there is a growing body of work relevant to the creation of text-to-speech systems with controllable or modifiable prosody. We here briefly survey related research on two themes.

The first theme is that of methods for prosodic control in TTS systems [8, 9, 10]. Two of the main approaches are "style"-based modifications and feature-based modification. Style-based control, as in [11, 12] involves conditioning TTS on style embeddings, often referred to as style tokens, either learned from a large multi-speaker corpus or from an acted corpus. These style tokens can then be used at inference to control the synthesis output either globally as in [11] or locally as in [12]. However, style tokens are intransparent and lack a clear link between style and prosodic realization. In contrast, feature-based control targets perceptually meaningful features, generally pitch, speech rate, and energy [10, 13, 14]. Recently it has been common to use such features in conditioning the signal, either at the phoneme level [10] or the utterance level [13, 14].

While such work has shown that it is possible to control prosody at a low level, in practice attention has been limited to only to pitch and duration, and occasionally also energy. Other prosodic features remain underexplored. These include voice quality features. Voice quality (in the narrow sense, i.e. voice properties arising from laryngeal activity [15]), is known to play an important part in conveying emotions, attitudes, and social cues [16]. More recent work is discovering how voice quality can also be used to signal pragmatic information, such as stance, emphasis and certainty, which are key elements of the pragmatic repertoire required for teammate-style tasks [7].

A second major theme in related research concerns the role of voice properties in speech synthesis. Prior to the advent of neural speech synthesis, several studies focussed on the synthesis of voice quality [17, 18]. More recent work focused on synthesizing different voice qualities exists, but has several limitations [19, 20, 21]. In [19], only the communicative functions of creaky voice were analyzed, which was perceived as less positive, less certain, and more turn-final than modal voice. In [20], breathy and creaky voice were investigated, with breathy voice being rated as more intimate and more invested, while creaky voice was rated as less intimate and less positive. These voice qualities were investigated in isolation, however, despite often co-occurring [22]. In [21], speech is synthesized with different degrees of roughness, breathiness, resonance etc., but the evaluation was limited to expert ratings of the accuracy of conveying the desired voice quality.

Thus there is an unmet need to directly examine how these speech properties affect perceptions. Specifically, we ask: can control of voice quality properties and nasality improve the quality of speech synthesizer output? In line with our interest in conveying more pragmatic functions, we address this question in the context of a collaborative video game player's utterances.

We address this question through three stages: first, development of a method for controlling these properties; second, a human-in-the-loop style experiment where five participants ("adjusters") modified the presence of creakiness, breathiness, and nasality; and third, a perception experiment evaluating the quality of these modified utterances. Later sections discuss each in turn. The contributions reported in this paper are: 1) a toolchain and workflow tool that enables human modification to the degrees of voice quality and nasality across utterances, and 2) a demonstration that such modifications can improve the suitability of synthesized speech.

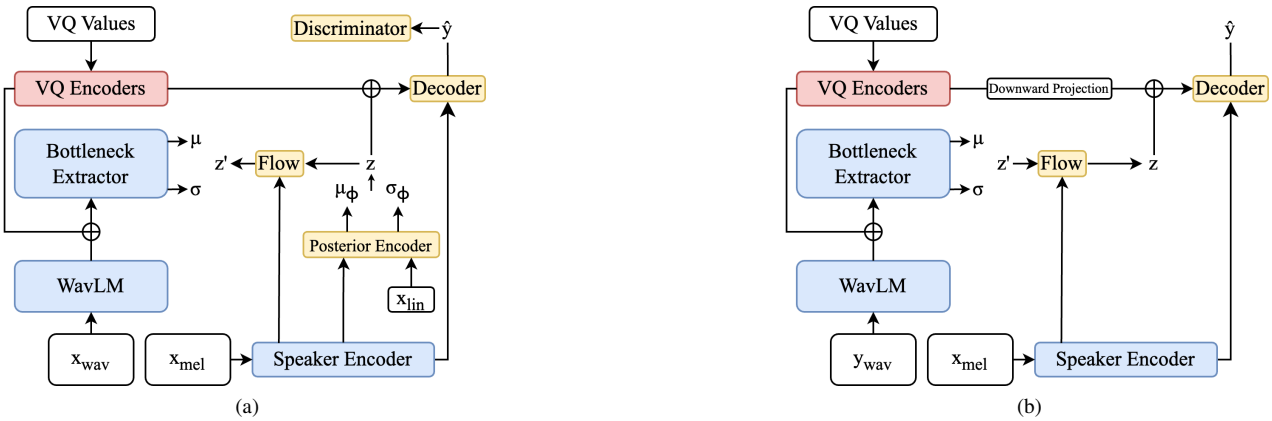

Figure 1: *The architecture of the system during finetuning (a) and inference (b)*

2. Models, Data, and Training

In order to answer our question, we needed to first develop a workflow to support the creation of samples with and without appropriate use of creakiness, breathiness, and nasality. The solution we developed is far from ideal, but allowed us to leverage existing systems and resources.

In short, we first synthesized basic versions of the utterances in US English using XTTS, exploiting its voice cloning feature to make them close to its target voice. We then did voice conversion, using the newly developed VoiceQualityVC tool, to slightly change these utterances to make them even closer: the result was our baseline utterances. Finally, participants used VoiceQualityVC again to adjust the properties of interest.

2.1. VoiceQualityVC and its Extension

While VoiceQualityVC is described elsewhere [20], it is new enough to merit a brief description here. Its architecture is based on a modification of FreeVC [23], a voice conversion system that is capable of modifying voice quality features. The main part of the architecture consists of a Conditional Variational Autoencoder (CVAE) with adversarial training that is conditioned on WavLM embeddings, and an RNN-based speaker encoder. It uses a prior encoder, in which the waveform is embedded using WavLM features, representations from an audio language model. These embeddings are passed through an information bottleneck, which is in turn passed through a normalizing flow to learn a more complex distribution. During training, it additionally uses a posterior encoder to extract speaker information from RNN-based speaker encodings and a linear spectrogram. These speaker representations are used to condition the normalizing flow. During inference, however, this information is inferred from the inverse normalizing flow, as detailed in [23].

The architecture of VoiceQualityVC during finetuning and at inference can be seen in Figures 1a and 1b, respectively.

To support the present experiments, we added separate encoders to this architecture for six features: two perceptually-meaningful features, namely pitch and pitch variation (st.d.), and four lower-level features, namely creakiness, Cepstral Peak Prominence Smoothed (CPPS), H1–H2, H1–A3. These six features are each handled by a voice quality encoder consisting of an affine layer in Figure 1 that projects the frame-level features into a 1024-dimensional space for the conditioning of the WavLM features, and a 192-dimensional space for conditioning the decoder inputs. The four lower-level features were not exposed to the adjusters directly. Rather, these properties were determined by the values they chose for creak, breathiness, and nasality, according to the mappings shown in Table 1, which we arrived at after some trial-and-error exploration. An informal evaluation of the distinctiveness of these voice types was performed, and while creaky voice was always accurately identified, perceptions of the outputs intended as breathy and nasal voice aligned with the intentions only about 80% of the time, as one would expect given the perceptual ambiguity between these properties for certain phonetic content [24].

Table 1: *The feature settings for each property.*

	Creak	CPPS	H1-H2	H1-A3
Creakiness	3	-1	-2	-2
Breathiness	-2	-1	3	3
Nasal	0	1	-3	3

2.2. Data Selection and Use

To generate the baseline samples and to create the manipulation functionality required the use of three corpora.

Given our interest in synthesis to support diverse pragmatic functions, we chose to use a corpus of Fireboy and Watergirl gameplay [25]. In this game, participants navigate a 2D landscape with various obstacles, widgets, and puzzles, some of which require cooperation to solve. Communication in this game is typically diverse, including self-talk while puzzle solving, various explanations, suggestions or instructions regarding what actions to take next, joint planning, narration of action, side comments, words to manage the interpersonal relation, and smalltalk. In this corpus, one person, EC, played with a dozen novice partners across a dozen 10-minute games. EC was chosen for his skill in making the game a fun experience; he used his voice very expressively. Our long-term aim is to build an AI player (NPC) that exhibits the same level of skill. For this experiment, we selected 17 of EC's utterances as he guided one novice through the first level of the game. These we term the "original" utterances. They are seen in the appendix.

In our experiment, the task for the adjusters was to modify baseline utterances to be more appropriate, taking inspiration from the original utterances. To make this easier for them, we

wanted the baseline utterances to be similar to what EC might have produced had he spoken the same words in a perfectly neutral style. This turned out to be difficult, as XTTS could not acceptably reproduce EC's speech patterns, perhaps because of his distinctive voice (very low and generally creaky and nasal), or because of his dialect, which exhibited some Spanish influences. We therefore used another American English speaker for the voice cloning. This was the mediator in the AptSpeech corpus [26]. Four utterances of between 8-12 seconds from this speaker were chosen for zero-shot voice cloning with XTTS [27]. We selected these from a segment where the mediator gave instructions to the participants, as this was a reasonable style match for at least some of EC's utterances.

Training the voice conversion model, VoiceQualityVC, requires a corpus annotated with voice quality features. For this, we automatically annotated LibriSpeech-R [28], a restored version of LibriSpeech for creakiness, Cepstral Peak Prominence Smoothed (CPPS), H1–H2, and H1–A3, in addition to pitch, following the procedure in [20]. To this feature set, we added the feature of pitch variation. Pitch was extracted using the Wavelet Prosody Toolkit [29], which employs continuous wavelet transforms for pitch detection. Pitch variation was measured by taking the per-utterance standard deviation of F_0 in Hertz. The pitch and pitch variation features were z-standardized, and this was done over the entire corpus to allow for more variation at inference. The voice quality features were standardized per speaker. We used this annotated LibriTTS-R data to finetune the weights of the pre-trained FreeVC model. The pre-existing train-test split was used. We zero-initialized the voice quality encoders and finetuned the model for 118k iterations on 2 24 GB NVIDIA GeForce RTX 3090 GPUs using a batch size of 8. We did not use the spectrogram distortion-based data augmentation. The model has a total of 39,354,304 parameters, an increase of 14,592 parameters compared to FreeVC [23].

We used the resulting version of VoiceQualityVC to create the baseline utterances. As noted above, this took the XTTS outputs and modified them to better match the average voice and style of EC. Specifically, conversion was done to attain the mean values for the voice quality for this speaker, which in practice were close to modal, and to attain -1.5 std. dev. from the mean for the average pitch and mean pitch variation, to roughly match EC's range. These modifications brought the baselines closer to the voice of the exemplary speaker, enabling the adjusters to focus on the task of interest, namely customizing the levels of creakiness, breathiness, and nasality.

3. Modification Process

In overview, five participants, the adjusters, modified the breathiness, creakiness, and nasality of synthesized speech using VoiceQualityVC.

3.1. Approach

In line with suggestions from [30], we investigate the appropriateness of the speech in the context of an actual use case: serving as the voice for a game player agent.

Our method is rather indirect, reflecting both the limited research goals of this paper (which do not include explicitly modelling the pragmatics-prosody mapping) and the inspiration we take from previous work. Specifically, we here rely upon human-in-the-loop modification of synthesizer output. Other work taking this approach includes [31], in which a grid display is used to enable the user to systematically listen to modified utterances within a specified modification range in order to select utterances that match the intended purpose. Our most direct inspiration is from a human-in-the-loop prosody-editing study for a cross-text prosody transfer task [32]. Findings from that work included: a majority of evaluators preferred the edited samples over the baseline, the samples suffered from decreased naturalness, especially as more edits were made, and participants did not modify each feature to an equal extent, with F_0 being modified to a greater degree than energy and duration.

3.2. Procedure

In the modification experiment, five adjusters, all with at least basic training in identifying prosodic properties, were tasked with modifying the 17 baseline utterances from the Fireboy and Watergirl corpus that were synthesized using XTTS [27]. These synthesized utterances were first converted to match the speaker identity of the experienced player in the Fireboy and Watergirl corpus. The adjusters were asked to modify the utterances to improve the in-context suitability. All were aware of our ultimate goal of building a highly supportive and trustable AI game-playing partner, and all had additionally watched and discussed recordings of the original human-human gameplay. While their goal was deliberately left somewhat vague, they were given access to the original human audio as a reference and used that as inspiration as they worked to improve the suitability for the context.

The adjusters were given the ability to add creak, breathiness and nasality to arbitrary regions of the utterance, and to do so to arbitrary degrees, between 0 and 1 or, rarely, higher. As the baseline utterances were always in modal voice, they never needed to reduce the values for these properties. In addition they were able to change the average pitch and pitch range, although this was not a focus of this study and they did this only rarely. The adjusters used a simple custom tool enabling them to specify regions of the audio and for each specify the intended voice quality or nasality degree. To make region selection easier, they were shown a timeline of the word and word boundaries derived from WhisperX [33] transcriptions. The tool was built in Jupyter Notebook, and adjusters were encouraged to check how their manipulated output sounded, and to make adjustments until it sounded acceptable. Multiple regions could overlap, thus, for example, adjusters could create a breathy nasal region. The adjusters performed the modifications unsupervised and at their own pace, with most deciding to split the work across multiple sessions. Adjusters were also asked about the ease of use and their confidence as to whether the modified utterances were improvements over the TTS baseline.

3.3. Observations

The five adjusters spent an average of 10 minutes on each utterance, with earlier utterances taking approximately 15 minutes and later utterances taking approximately 5 minutes. There was a similar learning curve for the perceived difficulty of the task, with most participants mentioning that they initially found the modifications challenging but became more confident and efficient as they progressed. Participants often combined several voice qualities in the same utterance. Out of the 85 total audio samples, 83 had added creak, 75 had added breathiness, and 69 had added nasality. Several tactics were used by the participants, with some participants annotating the use of voice quality in the original audio before implementing these voice

qualities with VoiceQualityVC, while others utilized an iterative approach, adjusting small segments, then listening to the result before making more adjustments. The adjusters expressed varied confidence in having improved the pragmatic suitability of the speech, but all were confident that they improved some, although not all, of the utterances.

We wondered which of the properties were easiest to modify and which were most often modified. There seemed to be a hierarchy, for which the quantitative evidence is seen in Table 2. Participants reported finding creak the easiest to identify in the original utterances and being most confident in their modifications for creak. Modifications to creak were also the most common, and tended to have higher intensity than for the other properties. Breathiness was towards the middle on all these dimensions, and nasality was the most difficult and least used. We also found that the modified regions for nasality tended to be longer than for the other properties. We think that this latter finding reflects a tendency in the originals for nasality to more often span entire utterances or large fractions of them, while the other voice quality properties may have been more often word- or phrase-bound.

We had several concerns going in, but in the event these did not seem to be problematic. We feared that some settings for the properties would lead to impairments in naturalness, and this was sometimes the case, but causing glitches that didn't sound like speech at all seemed to occur only rarely—for example, when setting both creak and nasality to high values—and the adjusters could easily fix this by backing off to less extreme values. We feared that locally modifying only these three properties would lead to unintended pragmatic functions, since often these properties probably work together with, or at least correlate with, other prosodic properties. However, this did not seem to be the case, likely in part because our voice conversion model tends to "drag along" other properties as it finds something that approximately meets the specification but remains within the distribution of seen data. Conversely, we feared that the adjusters would feel that the output was not really matching up with their specification, but in fact they seemed to be satisfied with whatever they got, even if it remained quite different from the original. We also feared that the adjusters would treat the properties as binary, but in fact they used a wide range of values for each property. There was little agreement among the adjusters on the specific modifications. To assess this, we examined inter-annotator agreement on the per-frame presence of modification for each voice quality feature, regardless of the magnitude of the change. As shown in Table 3, there was at best marginal agreement for creakiness, and no agreement for breathiness and nasality.

Table 2: *The average number of modifications, the mean intensity of the modifications, and the mean duration of the modifications for each modification type.*

	# of mods.	mean intensity	mean duration
Creaky	1.71	0.80	0.51
Breathy	1.62	0.69	0.46
Nasal	1.20	0.69	0.57

Table 3: *Fleiss' Kappa for each modification type.*

	Creaky	Breathy	Nasal
Fleiss' Kappa	0.06	0.00	-0.05

4. Evaluation

We hypothesized that these modifications — for creakiness, breathiness and nasality — would be beneficial: specifically, that the customized utterances would be rated higher than the baseline utterances.

Our measure was a Comparative Mean Opinion Score (CMOS) evaluation comparing the customized utterances to the corresponding baseline utterances. Crowdworkers were asked: "rate the suitability of the voice for an in-game supportive co-player for playing a collaborative video game." The ratings were performed on a scale from "1 – Much less suitable than the reference" to "7 – Much more suitable than the reference" with the additional anchor of "4 – Equally suitable as the reference". At the end of the experiment, the participants had the opportunity to comment on the experiment and they factors affecting their judgments.

In all we had 100 participants perform these judgments. Each evaluated the 17 creations of one of the adjusters; thus there were a total of 1700 judgments. These participants were recruited through Prolific[1] and were required to be native speakers of English residing in the United States of America. They made their judgments online, and were each paid £2 for their participation, with the average completion time being 7 minutes and 45 seconds.

Table 4: *Average preference for the modified samples. * indicates significance ($p < 0.005$).*

Adjuster	Average CMOS
#1	+0.35*
#2	+0.32*
#3	+0.26*
#4	+0.61*
#5	+0.19*
all	+0.35*

The hypothesis was supported across all the utterances, and indeed, also for each of the adjusters. That is, each adjuster's modified utterances were generally rated better, as seen in Table 4, and this was significant, using a single-sample *t*-test. The average improvement varied by adjuster, ranging from 0.19 to 0.61. Further, for each of the 17 pairs, the modified versions were on average rated more suitable than the baseline versions. We probed further with some post hoc and qualitative work. First, in an attempt to examine the individual contribution of each of the prosodic properties, we tested the hypothesis that utterances for which property i had been modified would have relatively higher ratings than those for which property i was unchanged. Using a Generalized Linear Mixed Model regression grouped by adjuster with the voice types and pitch and pitch variation as fixed effects, with random slopes for the voice types, we found positive trends for creakiness and breathiness, but these were not significant. There were no significant trends for pitch, which had a slight negative effect on the CMOS ratings, or pitch variation, which had a slight positive effect. Second, we examined the participants' comments on what helped them make their decisions. They variously noted improved prosody and sounding less robotic. Additionally, some participants mentioned preferring the emotion of the modified samples or imagining the suitability specifically for a video game that they frequently play.

[1] http://app.prolific.com/

5. Discussion

Every set of utterances modified by any of the adjusters was significantly better than the baseline audio according to the subjective CMOS evaluation. This suggests the importance of voice quality and nasality in conveying pragmatic functions.

However, so far, a true quantitative understanding is lacking. Not only are the individual contributions of each voice property unclear, we also do not know the magnitude of the contribution of these three properties relative to other prosodic features. Although most participants had positive comments about the modified utterances, and, to a lesser extent, about the baseline synthesis quality, some expressed the opinion that even the best were far from truly suitable. Seeking further insight, we compared the modified utterances of the highest-rated adjuster to the original utterances, to judge which of the functions listed in the Appendix she had been successful in conveying. Overall our impression was that maybe a quarter were managed adequately, although never unambiguously, and maybe for a quarter more the counterindications present in the baselines had been alleviated. Thinking about why this should be true, one possible confounding factor could be that the transcripts fed to the synthesizer were sometimes distant from the actual original speech, which included spontaneous speech behaviours such as repetition and pauses which were not replicated by XTTS. However the likely major factor is that, as some participants noted, the controls provided were inadequate to produce fully satisfactory utterances: the adjusters variously expressed the desires to have more precise control of speaking rate, pitch contours, and level of reduction, and the ability to more easily modify the intensity of voice quality per syllable. In general, quantifying the relative importance of the various prosodic features for various purposes is an important unsolved problem [34].

The low agreement scores for the specific modifications were somewhat surprising. We see various possible reasons. While the adjusters worked after listening to the original recordings, they did not necessarily attempt to mimic them. Their task was to modify the baseline to increase the in-context pragmatic suitability. They may have chosen different modifications that still serve the same pragmatic functions, or idiolectally differed in their ideas of the prosody-pragmatic mappings, or they may have had different ideas of which pragmatic functions were present in the original or were most important to include. They probably also perceived the presence of the prosodic properties differently, both in the original speech and the post-modification speech. In particular, it is likely that the original speaker selectively deployed different types of creakiness [35], that the adjusters perceived these differently, and/or that they differently perceived the specific kind of creakiness added in the modifications. Another possible cause for the lack of agreement for breathiness and nasality is the fact that the speech of the original speaker could generally be characterized as breathy and slightly nasal, therefore leading to the lower number of breathy and nasal modifications and less agreement for those ratings. Higher agreement existed for sentences where the voice quality of the original speaker was very salient. For example, in the utterance *right, next level*, the original speaker uttered *right* with pronounced creak, which was copied to some extent by all of the adjusters.

While prompt-based TTS systems powered by large language models are burgeoning (e.g., [36, 37]), they offer limited insight into the underlying mechanisms of prosody. Voice quality is a prime example: despite its importance, our understanding of how it contributes to pragmatic meaning remains limited. Although certain associations, such as creakiness signalling turn-finality are relatively well-documented [38], other relationships are speculative or entirely unknown. As noted above, the low agreement scores may suggest that there is not a single correct way to achieve a given pragmatic effect. Instead, different combinations of voice quality features might serve the same function, allowing for multiple valid modifications that enhance pragmatic suitability in different ways. Another important question is whether certain pragmatic functions can be reliably conveyed through voice quality modifications alone, or whether they require co-occurring prosodic and other cues, as has been seen for breathy voice [22].

These findings have potential downstream applications: a clearer understanding of how voice quality contributes to pragmatic meaning could inform the design of loss functions by incorporating measures of pragmatic appropriateness or listener preference into model training, rather than relying solely on acoustic similarity or naturalness scores. For instance, if certain combinations of voice quality reliably signal specific pragmatic functions (e.g., creakiness for turn-finality), models could be trained to produce these patterns in appropriate contexts by minimizing a loss that penalizes mismatches between target pragmatic functions and prosodic realizations. Additionally, it could guide the selection of training data, as current training datasets rarely possess the pragmatic diversity needed to convey pragmatic meaning appropriately. Our results can also shape annotation practices for future TTS systems. Rather than relying on high-level labels like "emotion: happy" or "style: formal", annotations could include fine-grained prosodic and pragmatic cues—such as voice quality, syllable-level emphasis, or discourse function (e.g., "acknowledgment", "contrastive statement"). This kind of labelling would make it easier to train models to generalize from voice quality to intended communicative function.

Future work could also develop ways for automating voice quality modifications, leading to a more scalable approach for implementing and evaluating voice quality in speech, perhaps avoiding the need for such post-processing by including control of such features in the speech synthesis model itself.

6. Implications

We may have taken a first step towards creating a new generation of synthesizers that can control these properties. In this section we briefly consider three likely types of benefits and how we may proceed.

First, including control of voicing properties and nasality may broaden the range of styles that can be effectively achieved. In particular, this may help lift synthetic voices beyond the competent-but-cold style that is so common today. While pitch-only prosody control may be acceptable, and even desirable, for intelligibility-prioritizing applications, we may need these additional properties to achieve styles that are perceived as more individual, sincere or trustworthy. To the extent that this is an important component of natural speech in many genres, we would advocate for selecting suitable data and adding terms to the prosody loss function to favour fidelity for these properties.

Second, extending synthesizers to handle these properties may enable them to learn a more comprehensive context-to-prosody mapping from training data. To some extent these properties seem to reflect aspects of the interlocutor's speech and local context, for example whether information is new or already likely part of common ground. The mappings between such context properties and the uses of these properties could

likely be learned by end-to-end methods.

Third, extending synthesizers to convey more aspects of agent intention and feeling — expressing confidence or uncertainty, a strong suggestion versus a mild one, dominance or submissiveness, engagement or lack of involvement, and so on — will likely require explicit representation of how inputs with such pragmatic functions map to appropriate prosody, including the features that we have focused on.

Building support for the last use case will likely require us to elucidate exactly how these properties serve each such function. For this, basic empirical research will be required. To date, systematic study of the functions of nasality and voicing properties has been hampered by the lack of a way to test hypotheses by creating stimuli with these properties controlled. Now, however, this is possible using the techniques described here.

7. Summary

In this paper, we examined the effect of voice quality and nasality on the perceived suitability of utterances for the role of a supportive co-player for a collaborative video game. In a modification experiment, five participants modified 17 baseline synthesized utterances using VoiceQualityVC. The participants had precise control over the creakiness, breathiness, and nasality, as well as some control over the average pitch and pitch range of the utterances. In line with our hypothesis, the modifications were rated as more suitable for the envisaged AI co-player.

Acknowledgement: This research was supported in part by the Air Force Office of Scientific Research under award number FA9550-24-1-0281 and the National Science Foundation through award 2348085.

Appendix

This appendix provides a transcript of the original utterances presented to the adjusters as they appeared in the original gameplay, with annotations to illustrate the rich variety of pragmatic functions in this data. The original audio is available at [redacted for anonymous review].

A few notes: These annotations had no part in the experiments. These functions were apparent to us from the words said, from the actions seen in the video, from the context, and from the prosody. For almost all the annotated functions it seems that prosody is important in conveying that intent or feeling. We do not know which of these functions the adjusters were sensitive to, or whether there were others they noticed, as we never asked them to explicitly identify the pragmatic functions.

Generalizing over what we see here we note that pragmatics here is much more varied than one of a few dozen dialog acts commonly seen in commercial dialogues [39]. We also see that, although much recent work in synthesis aims to improve emotional expressivity or to improve style diversity, those two factors account relatively little of what is actually happening here, in an effective, cooperative, real-time situated dialog.

EC: [some self-talk as he synchronizes the recording, etc.]

EC1: **alright, have you played this game before?** *(A new topic, shifting from self-directed to other-directed, announcing that he's ready to begin the game, an implicit invitation to get ready to start, friendly, positive tone, grounding.)*

Novice: nope

EC2: **alright, so you're going to be the lava boy, down there.** *(Acknowledges response, slowing down his speaking rate to speak more clearly now that he knows his partner will need some proper explanation, gives instructions, establishes common ground regarding an on-screen referent.)*

EC3: **you're going to move with the arrow keys.** *(Continuation of instructions, grounding, request for confirmation.)*

Novice: okay

EC4: **basically you can't touch what's opposite of you.** *(Continuation of instructions, giving a warning.)*

EC5: So, if, you're fire, so you can go into the fire, but you can't go into the water. *(continuing after a slight pause, a small false start but recovering, paraphrasing and elaborating, asking for confirmation.)*

Novice: ah, okay

EC6: **right there, you can walk down.** *(A digression from explaining the rules of the game to comment on what's happening now, suggesting an action, politely assuming the other sees what to do.)*

EC7: **and then this green mud, neither of us can touch it, because if either of us touch it, we both die.** *(Resumption of previous topic, calling attention to something salient, explaining a non-obvious rule, marking the word "die" as metaphorical.)*

Novice: okay

EC8: **so jump.** *(Direct instruction, urgent but polite, as the novice has already started to move.)*

EC9: **there you go.** *(Praise, authoritatively judging the novice's performance.)*

EC10: **yeah, these are the introductory levels.** *(A non-important aside, presumably intended to fill an awkward silence, in effect, apologizing for the simplicity of this game, making a contrast with the implied challenge in the later levels, implying that it will get more interesting.)*

EC11: **alright, I need you to push the button so the ledge goes out, yeah.** *(Recognizes hesitation, explains non-obvious action, requests cooperation, cues immediate action, offers praise upon completion.)*

Novice: ah, okay, got it (steps on the button but then moves forward)

EC12: **stay on it, and then let it go.** *(Flags incorrect action, gives a more explicit paraphrase, and then clearly instructs the next step, indicates no urgency.)*

EC13: **then I'll touch it, so you can get up.** *(Indicates successful preparation for the key move, announces specific intent, instructs the novice to wait, highlights reciprocity, cause-effect relationship.)*

EC14: **come on down.** *(Cues next subgoal, implies it's self-evident, delivered in a jokey, narrative tone, perhaps to back off from the face-threatening overly directive nature of 12.)*

EC15: **there you go; you've got the hang of it.** *(General praise, highlighting obvious progress rather than authoritative assessment.)*

EC: [laughs], no, it's not Mario

Novice: ah, okay, I did like go up *(apologetically)*

EC16: **yeah, I know, like in Mario you normally go up.** *(Accepts apology, minimizes error, empathizes, contrasts with another game's logic.)*

EC17: **right, next level.** *(Marks completion, gives praise, foreshadows greater complexity, indicates the intent to continue playing the game in full expectation that the novice also wants to continue.)*

8. References

[1] J. Cambre and C. Kulkarni, "One voice fits all? social implications and research challenges of designing voices for smart devices," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, pp. 1–19, 2019.

[2] M. Marge, C. Espy-Wilson, N. G. Ward, A. Alwan, Y. Artzi, M. Bansal, G. Blankenship, J. Chai, H. Daumé III, D. Dey *et al.*, "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, vol. 71, p. 101255, 2022.

[3] K. Lieberman and N. Sarter, "A comparison of auditory and visual representations of system confidence to support trust specificity, attention management, and joint performance in human-machine teams," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 65, no. 1, 2021, pp. 67–71.

[4] J. Wester, S. de Jong, H. Pohl, and N. van Berkel, "Exploring people's perceptions of LLM-generated advice," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 2, p. 100072, 2024.

[5] E. L.-C. Law, A. Følstad, J. Grudin, and B. Schuller, "Conversational agent as trustworthy autonomous system (trust-ca), dagstuhl seminar 21381," *Dagstuhl Reports*, vol. 11, no. 8, pp. 76–114, 2022.

[6] M. J. Barnes, N. Wang, D. V. Pynadath, and J. Y. Chen, "Human-agent bidirectional transparency," in *Trust in human-robot interaction*, C. S. Nam and J. B. Lyons, Eds. Elsevier, 2021, pp. 209–232.

[7] N. G. Ward, *Prosodic patterns in English conversation*. Cambridge University Press, 2019.

[8] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. ICASSP*, 2020, pp. 6189–6193.

[9] S. Shechtman and A. Sorin, "Sequence to sequence neural speech synthesis with prosody modification capabilities," in *Proc. SSW*, 2019, pp. 275–280.

[10] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *Proc. ICASSP*, 2021, pp. 6588–6592.

[11] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018, pp. 5180–5189.

[12] M. Lenglet, O. Perrotin, and G. Bailly, "Local style tokens: Fine-grained prosodic representations for TTS expressive control," in *Proc. SSW*, 2023, pp. 120–126.

[13] T. Raitio, J. Li, and S. Seshadri, "Hierarchical prosody modeling and control in non-autoregressive parallel neural TTS," in *Proc. ICASSP*, 2022, pp. 7587–7591.

[14] H. Lameris, S. Mehta, G. E. Henter, J. Gustafson, and É. Székely, "Prosody-controllable spontaneous TTS with neural HMMs," in *Proc. ICASSP*, 2023.

[15] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London*, vol. 31, pp. 1–186, 1980.

[16] R. J. Podesva, "Gender and the social meaning of non-modal phonation types," in *Annual meeting of the Berkeley linguistics society*, 2011, pp. 427–448.

[17] T. Drugman, J. Kane, and C. Gobl, "Modeling the creaky excitation for parametric speech synthesis." in *Proc. Interspeech*, 2012, pp. 1424–1427.

[18] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[19] H. Lameris, É. Székely, and J. Gustafson, "The role of creaky voice in turn taking and the perception of speaker stance: Experiments using controllable TTS," in *Proc. LREC-COLING*, 2024, pp. 16 058–16 065.

[20] H. Lameris, J. Gustafson, and É. Székely, "VoiceQualityVC: A voice conversion system for studying the perceptual effects of voice quality in speech," in *Proc. Interspeech*, 2025.

[21] F. Rautenberg, M. Kuhlmann, F. Seebauer, J. Wiechmann, P. Wagner, and R. Haeb-Umbach, "Speech synthesis along perceptual voice quality dimensions," in *Proc. ICASSP*, 2025.

[22] N. Ward, A. Kirkland, M. Włodarczak, and É. Székely, "Two pragmatic functions of breathy voice in American English conversation," in *Proc. Speech Prosody*, 2022, pp. 82–86.

[23] J. Li, W. Tu, and L. Xiao, "FreeVC: Towards high-quality text-free one-shot voice conversion," in *Proc. ICASSP*, 2023, pp. 1–5.

[24] S. Imatomi, "Effects of breathy voice source on ratings of hypernasality," *The Cleft Palate-Craniofacial Journal*, vol. 42, no. 6, pp. 641–648, 2005.

[25] N. G. Ward and S. Abu, "Action-coordinating prosody," in *Proc. Speech Prosody*, 2016, pp. 629–633.

[26] D. Kontogiorgos, V. Avramova, S. Alexanderson, P. Jonell, C. Oertel, J. Beskow, G. Skantze, and J. Gustafson, "A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction," in *Proc. LREC*, 2018.

[27] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea, L. Hart, A. Aljafari, J. Meyer, R. Morais, S. Olayemi *et al.*, "XTTS: a massively multilingual zero-shot text-to-speech model," *arXiv preprint arXiv:2406.04904*, 2024.

[28] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "LibriTTS-R: A restored multi-speaker text-to-speech corpus," in *Proc. Interspeech*, 2023, pp. 5496–5500.

[29] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Comput. Speech Lang.*, vol. 45, pp. 123–136, 2017.

[30] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. E. Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander *et al.*, "Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program," in *Proc. SSW*, 2019, pp. 105–110.

[31] É. Székely, S. Wang, and J. Gustafson, "So-to-speak: an exploratory platform for investigating the interplay between style and prosody in TTS," in *Proc. Interspeech*, 2023, pp. 2016–2017.

[32] H. Maurya and A. Sigurgeirsson, "A human-in-the-loop approach to improving cross-text prosody transfer," in *Proc. Interspeech*, 2024, pp. 2295–2299.

[33] M. Bain, J. Huh, T. Han, and A. Zisserman, "WhisperX: Time-Accurate Speech Transcription of Long-Form Audio," in *Proc. Interspeech*, 2023, pp. 4489–4493.

[34] N. G. Ward, D. Marco, and O. Fuentes, "Which prosodic features matter most for pragmatics?" in *Proc. ICASSP*, 2025.

[35] P. A. Keating, M. Garellek, and J. Kreiman, "Acoustic properties of different kinds of creaky voice." in *ICPhS*, vol. 1, 2015, pp. 2–7.

[36] D. Yang, S. Liu, R. Huang, C. Weng, and H. Meng, "InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt," *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2024.

[37] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "PromptTTS: Controllable text-to-speech with text descriptions," in *Proc. ICASSP*, 2023, pp. 1–5.

[38] M. Włodarczak and M. Heldner, "Contribution of voice quality to prediction of turn-taking events," in *Proc. Speech Prosody*, 2022, pp. 485–489.

[39] H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. C. Fang, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria *et al.*, "Towards an ISO standard for dialogue act annotation," in *Proc. LREC*, 2010.