Generation of speech and facial animation with controllable articulatory effort for amusing conversational characters

Joakim Gustafson
KTH Royal Institute of Technology
Stockholm, Sweden

Éva Székely
KTH Royal Institute of Technology
Stockholm, Sweden

Jonas Beskow
KTH Royal Institute of Technology
Stockholm, Sweden

ABSTRACT

Engaging embodied conversational agents need to generate expressive behavior in order to be believable in socializing interactions. We present a system that can generate spontaneous speech with supporting lip movements. The neural conversational TTS voice is trained on a multi-style speech corpus that has been prosodically tagged (pitch and speaking rate) and transcribed (including tokens for breathing, fillers and laughter). We introduce a speech animation algorithm where articulatory effort can be adjusted. The facial animation is driven by time-stamped phonemes and prominence estimates from the synthesised speech waveform to modulate the lip- and jaw movements accordingly. In objective evaluations we show that the system is able to generate speech and facial animation that vary in articulation effort. In subjective evaluations we compare our conversational TTS system's capability to deliver jokes with a commercial TTS. Both system succeeded equally good.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

ECAs, speech synthesis, facial animation, humour generation

ACM Reference Format:
Joakim Gustafson, Éva Székely, and Jonas Beskow. 2023. Generation of speech and facial animation with controllable articulatory effort for amusing conversational characters. In *ACM International Conference on Intelligent Virtual Agents (IVA '23), September 19–22, 2023, Würzburg, Germany*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3570945.3607289

1 INTRODUCTION

There is a rich history in the development of embodied conversational agents (ECAs) designed to engage in spoken interactions with users [9]. Early examples include the astronomy guide Gandalf [50], the desktop agent PPP persona [1], the virtual tutor Steve [39], the publicly available August [17] and the real estate agent REA [8]. In recent years, ECAs have been employed in various applications, such as museum guides [5, 26, 44], educators [3, 19, 60], computer game characters [16, 40], and virtual companions [6, 7, 52]. There is also a recent trend to explore humour in virtual agents and social robots [34], where some even work towards a robot theatre where robots can track the audience response to their jokes [22]. This trend stems from humor being found to promote engagement and increase motivation to follow the advice of coaching ECAs [33]. Zhang et al. investigated how humor styles influence the perception of joke-delivering robots [62]. They selected 20 jokes in five styles (Affiliative, Self-enhancing, Self-defeating, and Aggressive) from famous comedians and the Jester joke dataset, finding that robots telling self-defeating jokes received higher scores and more laughter detected from the users.

Numerous studies have examined the perception of different voice types for interactive agents [41]. Social robots and ECAs typically use the same kind of TTS as voices assistants, which are designed for simple, transactional interactions where users ask questions or issue commands, and the agent responds verbally or performs actions. As a result, these TTS voices aim to emulate a neutral, warm, and informative speaking style. However, to develop more amusing and opinionated conversational characters, it is crucial to incorporate more engaging vocal performances in their TTS voices [2]. In real-world scenarios, conversational agents must adapt their speaking styles according to the situation, such as speaking more clearly to securely convey messages or varying vocal effort to deliver dramatic or engaging content. Lindblom's H&H theory posits that human speech production is influenced by physiological economy constraints [30], with hypo-articulated speech requiring minimal articulation effort and hyper-articulated speech maximizing clarity. One study found that neural TTS underperformed in speech intelligibility in noisy environments compared to the clearer concatenative TTS [11].

In the current study, we have developed a neural conversational TTS system capable of controlling the articulatory effort of its synthesized speech. Furthermore, we have integrated an avatar with lip movements that are coherent with the generated speech, considering both phonetic content and articulation effort. We conducted objective evaluations where we generated 100 sentences in different manners of speaking and measured the results automatically. To assess its effectiveness subjectively we decided to conduct a very demanding task - the Ebert test, as detailed below. We utilized the large language model GPT-4 [35] as a joke generator. Subsequently, we generated 16 word pun jokes with a self-defeating twist that required our amusing conversational character to transition from hyper-articulation to hypo-articulation.

> **The Ebert Test:** *If the computer can successfully tell a joke, and do the timing and delivery as well as Henny Youngman, then that's the voice I want!* —Roger Ebert, 2011

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
IVA 2023, September 19–22, 2023, Würzburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-XXXX-X/18/06.
https://doi.org/10.1145/3570945.3607289

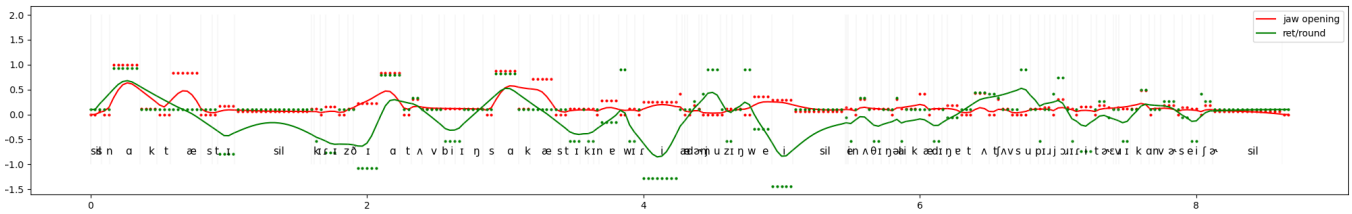

Figure 1: Animation tracks (solid lines) and targets (dots) for jaw opening and retraction-rounding for a conversational TTS utterance, showing larger movements in the first (hyper-articulated) part and smaller in the last (hypo-articulated) part.

2 RELATED WORK

In a study comparing the likability of human speech with TTS voices of the robot Sophia and IBM Watson [27], TTS voices were perceived negatively, often characterized as too smooth or lacking comprehension. Conversely, natural voices garnered positive feedback for their prosody, paralinguistic, and extralinguistic cues, such as audible breathing and smiling voice. A recent study investigated the perceived personality in a virtual agent controlled by human speech and gestures or using TTS and state-machine-like animation [49]. Extroversion was mainly communicated through motion, while speech influenced agreeableness and emotional stability.

To develop style-specific TTS, researchers often use corpora with specific speaking styles. This method was employed to create unit-selection TTS voices with distinct personalities for animated characters in a speech-enabled computer game [18]. Another approach involves using a large corpus containing varied speaking styles and automatically detecting a given number of Global Style Tokens (GST) and then categorize these by listening to them [58]. Style tokens have also been used for emotional TTS [53], training the system on voice actors who read drama scripts manually tagged for emotions (happiness, sadness, anger, or neutral). However, as argued by Marge et al. [31], it is not obvious that it is possible to extend style tokens from book reading or acted emotions to the kinds of communicative functions [59] you would need in human-machine interactions. Efforts have been made to create a TTS corpus closer to a conversational speaking style by recording an actor reading chatbot scripts [61]. However, spontaneous conversational speech has been found to be more varied in pitch and speaking rate than scripted conversational speech [28]. Some systems are trained on a large number of speakers speaking in a range of styles, where the manner of speaking is controlled using a reference audio that is given as input along with the text [10, 29]. However, it can be difficult to chose the set of reference audios to use given a specific situation. The neural conversational TTS system presented in this paper was trained on a corpus of a male speaker who either read texts in a clearly articulated manner or was the moderator in casual three-party interactions [25]. This combined corpus of read and spontaneous speech of the same speaker contains a range of verbal behaviours. We have then developed a method where we can both mix the speaking styles and control the prosodic realization.

Embodied conversational agents need to have lip movements that are synchronized and coherent with the synthesized speech. Taylor et al. trained a deep neural network on a large audio-visual dataset containing a single actor reciting 2543 phonetically diverse sentences in neutral tone [48]. Given speech and the phonetic transcription, the system generates lip movements that represent the phonetic content of the speech, but not the manner of speaking. JALI is an animator-centric workflow for the automatic creation of lip-synchronized animations [14]. They introduce the JALI viseme field with a lip and a jaw axis. This is used to capture speaking styles like mumbling, screaming and normal conversation. In recent years, there have been notable advancements in multimodal systems capable of generating speech accompanied by non-verbal behaviors such as co-speech gestures [57] and facial expressions with synchronized lip movements [20]. Additionally, high-fidelity talking systems based on neural radiance fields have emerged, exemplified by the work of Guo et al. [15]. However, one limitation of these systems is their rendering time, which currently takes approximately 12 seconds per frame on an RTX 3090 GPU. This rendering delay poses a challenge for real-time conversational applications where instantaneous generation is crucial. While real-time alternatives like NVIDIA Audio2Face [51] exist, they typically require extensive training data of around 60 minutes per actor. Fortunately, recent advancements by Pan et al. [36] present a promising lip sync system that can be trained on smaller datasets. In our paper, we introduce a flexible lip sync system capable of adapting to various facial rigs. This means that the system can drive any blendshape-based rig used in virtual agents or robots. Moreover, our method offers explicit control over articulation effort, which means that we can make the lip movements hyper or hypo clear, in order to match different speech styles and allow this to be manipulated. This flexibility enables manipulation in research settings. Notably, our approach is training-free and addresses a common use-case that is often overlooked in virtual agent and social robot applications.

3 SYSTEM DESCRIPTION

3.1 Conversational Text-To-Speech system

The conversational TTS was trained using Tacotron 2 with an added utterance-level prosody control method, similar to [38], and a speaking style control using an 8-dimensional speaker-like embedding, similar to [54]. In a three-party dialogue corpus the moderator turns were used as a TTS corpus [25]. These were segmented into breath groups (stretches of speech delineated by breath events) using a deep learning-based breath detector [46]. These were then transcribed using Whisper ASR [37] and subsequently corrected to ensure accurate transcription of all fillers and repetitions. For each breath group, we measured the mean f0 and speech rate (approximated by the peaks of the wavelet matrix) using the Wavelet

Prosody Analyzer [43]. The mean f0 and speech rates of the breath groups were normalized by aligning the 1st and the 99th percentile points of the data to -1 and 1, respectively, while allowing outliers to extend beyond that range. At inference it is possible to extrapolate on the features by going beyond this normalised range, enabling the model to generate hyper- and hypo articulated speech using both a limited set of actual training data in this range, but also relying on the full corpus for robustness. Normalized values for these two features were appended to each utterance's encoded text and passed to the attention and decoder blocks from the pre-trained model. A model is first initialised on a pre-trained read speech model, and trained for 70k iterations on the corpus with two embeddings, indicating whether the utterance is from the *read* or *spontaneous* part of the corpus. This model is then further trained including the prosodic features for an additional 100k iterations. We used a HiFi-GAN [24] vocoder fine-tuned on the same corpus for 383k iterations on the top of the published model. Data collection is further described in Sec. 4.1.

3.2 Speech Animation with Adjustable Effort

We introduce a new, pseudo-biomechanical algorithm for generating speech animation. This algorithm offers a straightforward yet effective approach to account for co-articulation with adjustable articulatory effort by minimizing energy while maximizing adherence to articulatory targets. By varying the weight between these (often conflicting) goals, it is possible to produce speech animation with varying degree of clarity, in line with Lindblom's H&H theory [30] where articulatory effort (here: the required energy) stands in proportion to informational requirements (here: adherence to targets). The algorithm requires no training data and can be applied to different facial rigs. The input to the speech animation algorithm is a time-stamped phoneme sequence. In our experiments, this is derived from a phoneme recognizer based on wav2vec2.0 [4]. We use five high-level parameters to describe visual speech targets and articulatory motion. Drawing inspiration from Öhman's model of coarticulation [32], which proposes to view articulation as the superposition of continuous vowel motion and rapid consonant articulations, we use two parameters for vowel articulation: jaw [0..1] (degree of jaw-opening) and *retraction-rounding* [−1..1] (negative values correspond to lip retraction, positive values to lip rounding), and three parameters for consonant articulations: *bilabial* [0..1] (where 1 means lip closure, regardless of state of jaw), *labiodental* [0..1] (lower lip/upper front teeth contact + raised upper lip) and *dental* [0..1] (parted lips + elevated tongue), see Fig. 2

For each phoneme, and for each parameter, a tuple t, w describes articulatory target position and a weight that dictates the importance of the target. As an example, the consonant k may be either rounded or retracted and will therefore have a weight of $w = 0$ for the *retraction-rounding* parameter. In the case of the bilabial consonant b, the *bilabial* parameter target $t = 1$ is paired with a weight of $w = \infty$ ensuring that the target will always be reached for this phoneme. To synthesize a new animation x_i, a target sequence t_i is formed by placing targets on a timeline according to a provided time-stamped transcription, along with the corresponding target weights w_i. We approximate articulatory effort by the total acceleration summed over an articulatory parameter trajectory x_i as $E_1 = \sum_i |x_{i-1} + x_{i+1} - 2x_i|$. We approximate information content loss as a weighted sum of deviation from articulatory targets: $E_2 = \sum_i |x_i - t_i| w_i$ and calculate the final parameter trajectory x_i that minimizes the sum $E = E_1 + E_2$, where the E_1 term effectively tries to *straighten* the track while E_2 tries to adhere to the defined targets as closely as possible. In order to model varying levels of prominence in the articulation, we can do two things in this model: 1) increase the weight of targets belonging to prominent syllables, thereby forcing the trajectory closer to the target; and 2) shift the target to a more extreme position; this applies to the vowel parameters (jaw opening and retraction-rounding) which may simply be scaled up or down. In practice, we use prominence estimates to modulate both the target weights and the vowel target scaling, along with a global scaling for hyper-hypo articulation, see Fig. 1. In our experiments, the generated articulation tracks are be used to drive the Furhat social robot or its digital twin simulator, but they can be easily implemented on other facial animation rigs. Videos of 16 sentences generated with our conversational TTS in high or low articulatory effort and with the standard Furhat lipsync (baseline in our experiments) and the Amazon Polly TTS voice Matthew can be found at https://www.speech.kth.se/tts-demos/iva2023/.

4 METHOD

4.1 Speech synthesis corpus

Developers of conversational systems should ensure that the TTS voices they use are trained on ecologically valid data [2]. Our long-standing goal is to build social robots capable of engaging in multi-party interactions. To achieve this, we require data to train models that generate appropriate speech, facial gestures, and gaze behaviors. Consequently, we have recorded a corpus in which the same male American speaker acted as a moderator in 15 one-hour, three-party interactions [25].

In these recordings, the moderator and two participants were assigned the task of decorating an apartment using a GUI on a large touch screen. The recordings took place in a motion capture lab, where all participants wore headset microphones, eye-tracking glasses, and gloves, and were filmed by three video cameras. All channels were synchronized and timestamped using the Farmi framework [21], ensuring that all aspects of the multi-party interaction were captured effectively, see fig. 3.

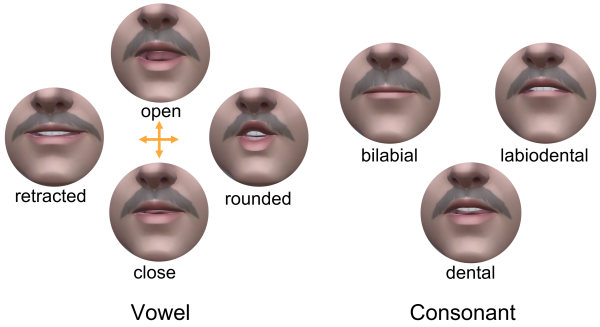

Figure 2: The 5 parameter model for the lip synchronization.

In each interaction, the moderator first engaged in small talk with the participants before introducing the task at hand. He then assumed the role of an interior decorator, offering suggestions on how to decorate the apartment and providing instructions on using the GUI for this purpose. Occasionally, he adopted a self-directed speaking style while contemplating design options or commenting on the users' progress. As the moderator switched between small talk, instructions, advice-giving, and casual commentary, the resulting corpus encompassed a wide range of spontaneous speaking styles. In total, the conversational TTS corpus contains 5 hours and 40 minutes of moderator speech. To facilitate the generation of hyper-articulated speech, the TTS corpus was supplemented with 2 hours and 30 minutes of clear speech, in which the moderator read sentences from the CMU Arctic [23] and newspaper texts. The total TTS corpus spans approximately 8 hours.

4.2 Joke delivery generation

Typically, TTS voices are evaluated using mean opinion scores (MOS), where generic sentences suitable for reading aloud are synthesized. However, this approach tends to favor neutral, warm, and informative speaking styles, which are best suited for reading news or engaging in transactional interactions with voice assistants. In this paper, our goal is to evaluate a conversational TTS voice capable of expressing different attitudes while speaking. As mentioned earlier, the ultimate test for an expressive TTS voice would involve delivering jokes with the appropriate timing and intonation. We challenged ourselves by selecting joke delivery as the speech synthesis evaluation task. Instead of using existing jokes from corpora like the Jester joke dataset, we decided to generate the jokes to be synthesized using the large language model GPT-4. The prompt used for generating the joke candidates was: *"Can you invent words that do not exist and then describe what they mean in a fun and entertaining manner?"*. We also generated self-defeating comments for each joke, using the following prompt: *"Can you give a sarcastic comment as a response to this joke?"*. Examples of the jokes are listed in Table 2.

We used the recently proposed So-to-Speak system [47]] to generate three different levels of articulation (see Tab. 1). This interface allows users to generate and interact with hundreds of synthetic speech samples using multi-dimensionally controllable TTS. The design displays prosodic feature variations on the axes of an interactive grid, where samples can be played by selecting them. The style function can be varied interactively, with a slider enabling users to scroll through grids exhibiting various levels of conversational and read speech styles. The samples displayed on the grid are playable upon clicking, and they are marked and colored according to an automatically generated naturalness MOS score using [12]. The scores range from 1-5 (with 5 being "completely natural"), and the corresponding colors range from red (1) to green (5). This provides users with an estimate of how the settings on the controllable features affect the quality of synthetic speech. The control interface and an example grid with TTS samples are illustrated in Fig. 4. Using this interactive tool, one of the authors selected specific ranges of the controllable features to create three manners of perceptually distinctly different articulation, hyper-, normal and hypo-articulation, as presented in Tab. 1. Since the TTS engine is built on Tacotron 2 [42] which is probabilistic at inference, the samples are synthesized with natural variation within each setting.

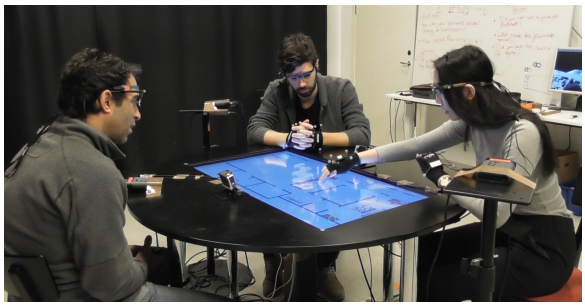

Figure 3: Picture from the data collection where the moderator and two participants are decorating an apartment.

Table 1: The style and prosody controls used for the articulation efforts. Input values are based on normalized utterance-level averages for f0 and speech rate, where -1 corresponds to the 1st percentile in the corpus and 1 to the 99th percentile.

Articulation	Read/Conversational	Pitch	Speech rate
hyper	80/20	0 to 1	-2.0 to -1.0
normal	20/80	-0.5 to 0.5	-0.5 to 0.5
hypo	0/100	-2.0 to -1.0	1.0 to 2.0

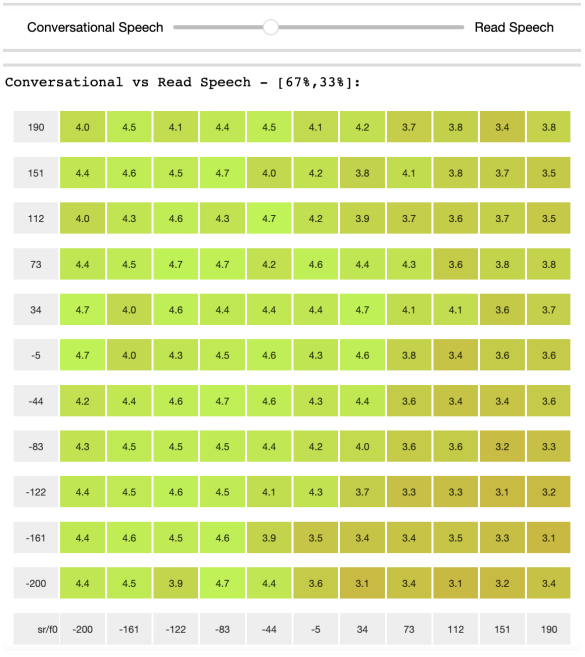

Figure 4: An example grid with a sentence synthesized with 7 style settings and 11 prosodic feature steps, totalling 847 unique speech samples. The audios play upon clicking on a cell. Style slider on top, updates the grid to the requested style. Colors correspond to estimated MOS scores.

Table 2: Examples of the GPT-4 invented words, funny descriptions and sarcastic self-mockery

Invented word (Hyper-articulation)	Description (Normal articulation)	Self-mockery (Hypo-articulation)
Fail-forward	It is the act of failing multiple times, but continuing to learn from your mistakes and moving forward in a positive direction.	Just what we need, more people failing their way to success.
Oopsy-daisy Detector	It is a device that alerts you whenever you're about to make a clumsy mistake, by yelling oopsy-daisy.	Just what we need, more machines to do the thinking for us.
Procrastinatron 3000	It is a robot that gently encourages you to keep putting off tasks, by offering endless entertainment, snacks, and distractions.	Finally, a robot that understands my priorities.
Sassy Stapler	It is a cheeky office supply that offers witty comments, as it binds papers together.	Because adding more distractions to your workday is always a great idea.
Schrödinger's Socks	It is socks that are simultaneously mismatched and perfectly paired, depending on whether you look at them.	Just what we need, more confusion in our lives.
Hummus Sapien	It is a human who loves hummus so much, it's practically their identity.	Just what we need, more people claiming to be unique by having the same favorite food.
Philosopher's Stoned	It is a state of mind where profound thoughts seem hilarious and deep simultaneously.	Just what we need, more deep thoughts that make no sense.
AcciDelight Cake	It is a cake that didn't turn out as expected but still tastes delicious, reminding us that life's imperfections can still bring joy.	Nothing like setting the bar low and still managing to trip over it.

5 OBJECTIVE EVALUATION

In [45] a tool called Starmap is introduced for visualizing and exploring the variety of prosodic styles across a corpus using the dimensionality reduction method t-SNE [55] and normalized utterance-level means of prosodic features extracted with the Wavelet Prosody Analyzer [43]. As an objective evaluation, we apply this method to validate the system's ability to produce varying degrees of clarity in articulation, With Starmap, it is possible to estimate speech rate based solely on acoustics, using peaks in the maximum energy scale, which correlate with the locations of syllables. However, in conversational speech, particularly hypo-articulated speech, syllables are often reduced or dropped entirely. This can result in a difference between the number of syllables identified in the speech samples and the number of syllables in the written prompt used as input to the TTS. We can use this metric, the ratio of estimated versus written syllables, as a measure of how much the prosodic features (f0, speech rate, and energy) influence the clarity of articulation. The same 100 utterances are synthesized in different articulatory styles, namely normal (middle of the distribution of all features), hyper-articulated (high f0, slower speech rate, high energy), and hypo-articulated (low f0, faster speech rate, low energy). The dropped syllable ratio (DSR) of each style is shown in Fig. 5. Our hypothesis that hyper- and hypo-articulated speech both significantly alter the DSR of synthetic speech is confirmed by pairwise t-tests on the distribution (hyper vs. normal $p \ll 0.001$, and hypo vs. normal $p \ll 0.001$). The measured prosodic features, as well as the DSR of the evaluation utterances, are visualized in a t-SNE in Fig. 6. A two-dimensional Kolmogorov-Smirnov test is performed to verify that the distribution of hyper- and hypo-articulated utterances are different from the normal utterances and from each other, The results confirm that both the hyper- ($p \ll 0.001$) and hypo-articulated ($p \ll 0.001$) synthesis results in different distribution of prosodic representation compared to the normal population. The same holds true between the hyper- and hypo-articulated populations ($p \ll 0.001$).

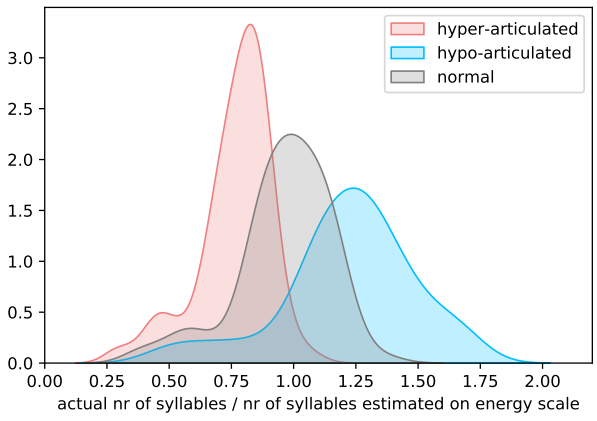

Figure 5: Density graphs for dropped syllable ratio: the ratio of the number of syllables in the input text to the estimated number of syllables from the signal (using the peaks of the energy scale) for the 100 synthesis samples in three styles.

6 SUBJECTIVE EVALUATION

To investigate the effect of the proposed methods, we carried out two online perceptual tests, looking at *joke-delivery* and *audiovisual speech matching*.

6.1 Method

We generated 16 utterances, according to the procedure described in 4.2. We synthesized the new words in a hyper-articulated speaking style, the funny descriptions with an expressive prosodic realisation and the self-defeating comments in a hypo-articulated speaking style. In addition to the conversational TTS voice, the utterances were also synthesized using a commercial TTS voice. In the *joke-delivery* task, we asked subjects to listen to synthesized jokes (audio

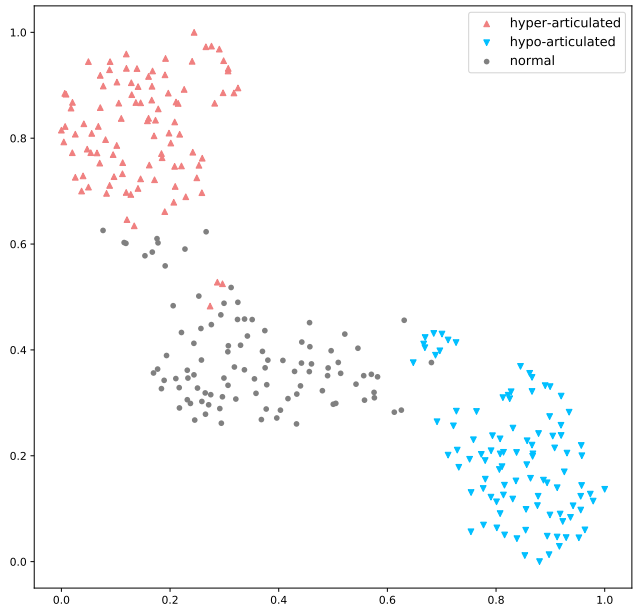

Figure 6: t-SNE visualization of the distribution of the prosody of 100 synthesized utterances in three styles, based on utterance-level normalized mean values of duration, f0, energy, speech rate (syl/s) and the dropped syllable ratio (DSR).

only) and rate how well the joke was delivered on a 5-point scale from *poor delivery* to *great delivery*. Each subject received the 16 jokes, 8 in each voice. The pairing of joke and voice was randomized between subjects, as was the presentation order. At the end of the experiment, we asked follow-up questions about their experience with speaking machines such as Alexa or Siri, what they based their ratings on, if they believe computers should have human traits such as humor or sarcasm, as well as if they had general comments on the study.

For the *audiovisual speech matching* task, we presented animations rendered with the virtual Furhat robot of the same 16 utterances, and asked subjects to rate how well the lip movements match the speech on a 5-point scale from *not matching at all* to *perfect match*. The animations were generated by the speech animation method presented in section 3.2, and using a baseline method (the Furhat systems built-in lipsync). We used the same two voices as in the joke-delivery test, and generated videos representing all four configurations of speech animation method (*new* vs *baseline*) and voice (*commercial* vs *conversational*). Each subject was presented with 16 animations, four in each configuration. Pairing of joke and configuration was randomized between subjects, as was the presentation order. For each of the tests, we recruited 70 subjects on the Prolific crowd-worker platform for the task. An attention check was used in the middle of the sequence. Median completion time was 5:30 minutes and subjects received a 1.50 GPB compensation.

6.2 Results

Ratings from the two experiments were analysed by means of a one-way ANOVA and a post-hoc Tukey multiple-comparisons test for statistical significance. In the joke delivery task, the mean rating and 95% confidence interval for the commercial and conversational voice was 2.5 ± 0.1 and 2.6 ± 0.1 respectively, but the difference was not statistically significant. Results from the audiovisual speech matching task are shown in Fig. 7 top. The *conversational TTS + new animation* configuration got the highest rating, and the *commercial TTS + baseline animation* the lowest. All differences were significant ($p < 0.05$). The joke-delivery test also contained a set of open questions. For the question *What did you base your rating on?*, intonation was most frequently mentioned, followed by timing, funniness, human-likeness and clarity, see Fig. 7 bottom. Often they gave several of these where the most common combinations were funniness and human-likeness, or intonation and timing. When computing the average scores per reason both voices got the same score for all reason accept for clarity, where the commercial TTS got 3.3 and the conversational 2.7. The lowest score for both (2.2) was from the users who based their scores on how funny the actual joke was, and not how it was read. Otherwise the average scores for both TTS voices were 2.7. In response to the question *Do you think computers should have human traits, like humor and sarcasm?* 44 said "yes", 19 "no", and 6 "I doné know".

7 DISCUSSION

Our work aims to advance the development of social robots and embodied conversational agents which can serve as companions or conversational peers. To achieve this goal, we have created an audiovisual speech generation system for expressive conversational characters and developed methods to control the manner of speaking with accompanying facial animation. As pointed out by Wagner at al. when we evaluate our TTS systems we need *"to assess and take into account listeners' application-specific needs and expectations"* [56]. Furthermore, the TTS evaluations should be as contextualized as possible to the participants. In our study, we choose a companion agent which could provide social company as the context and joke-delivery as the test case. We found that our conversational TTS voice performed on par with a state-of-the-art neural commercial TTS voice in the joke delivery task. Notably, several participants' comments revealed that many were impressed by the unique human-like attributes of the conversational TTS voice. Some particularly noteworthy general comments include:

> *"One of the audios sounded very well like a human, I did not expect this technology to be this far in human mannerisms."*
>
> *"Some audio tracks sounded quite close to the way humans deliver a joke! You can still tell they're computer-generated but I'm floored by how much more advanced they sound compared with Siri"*
>
> *"As time went on, the jokes got funnier despite the quality of the jokes not getting any better. The absurdity may have had me rating a bit higher."*.

This tells us that while are not up for an Ebert test just yet, we have might have managed to build a rather capable conversational voice. This is in line with our long-standing goal of building more

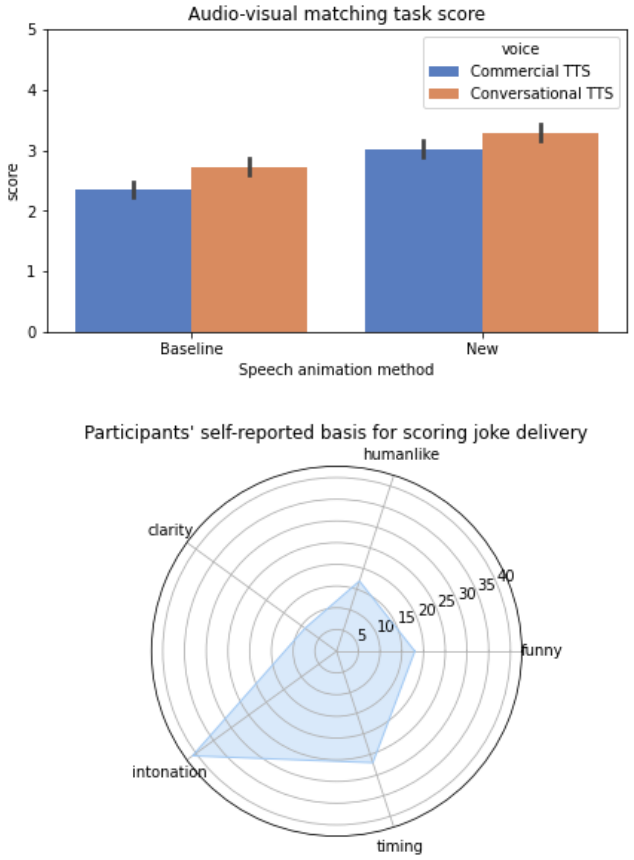

Figure 7: Subjective evaluation results. Top: Score from the audiovisual speech matching task, bottom: joke delivery task, summary of responses to *What did you base your rating on?*.

human-like conversational systems [13]. The ability to change the intonation and articulation effort is crucial in situated interaction and during error handling and grounding. Finally, we found that our new speech animation method consistently outperformed the baseline in the audiovisual speech matching test for both TTS-voices. We also note that the conversational TTS voice, which is considerably more varied in speech rate and articulatory effort than the commercial TTS, also received a higher rating in the multimodal setting.

8 CONCLUSIONS

We presented a system capable of producing conversational synthetic speech and accompanying facial animation with an adjustable degree of clarity of articulation. Since the TTS is probabilistic, the generated speech has an added natural variation. With this functionality we hope to enable virtual agents to exhibit refined social behaviors such as mumbling, muttering, attracting attention, being engaging or talking more clearly during error resolution. A novel speech animation algorithm that allows control over articulatory effort, for varying prominence and hyper-hypo speech production, pairs particularly well with the conversational TTS voice.

In the objective evaluations we show that the system indeed was able to generate speech that vary in articulation effort with accompanying lip movements. In the subjective evaluations we compared our conversational TTS system's capability to deliver jokes with a commercial TTS in an audio-only setting. Both system succeeded moderately good at this task, indicating that today's TTS technology is not on comedian-level yet. In a multi-modal context, we found that the conversational TTS, combined with our novel speech animation algorithm, provided the best overall subjective audiovisual coherence. These findings suggest that our system has potential for creating more natural and engaging conversational agents.

A key contribution in this paper is the development of a TTS voice that is not only grounded in ecologically valid data but also capable of generating spontaneous speech with accompanying lip movements for ECAs. This was achieved by constructing a voice for conversational systems that allows for the manipulation of speaking style, articulatory effort, and prosodic realization. The TTS voice was trained on a diverse speech corpus, which included slow, clear read speech as well as conversational interactions from the same speaker. This training enabled the blending of read speech with spontaneous conversation, while also providing control over pitch and speaking rate. Furthermore, the system's facial animation is driven by time-stamped phonemes and prominence estimates derived from the synthesized speech waveform, allowing for the modulation of lip and jaw movements in sync with the speech. This adds a layer of realism and expressiveness to the ECAs. In addition, we developed a GUI for VUI designers, which facilitates the control of the blend between read and conversational speech, as well as prosody. This GUI is instrumental in pre-generating system prompts with precise prosodic realization and offers insights into the capabilities of the voice in terms of speaking style mix and prosodic realization. VUI designers can utilize this tool to learn the optimal mixes and ranges of pitch and speaking rate to pair with system prompt text for achieving specific pragmatic functions. Lastly, we introduced a novel evaluation paradigm that transitions from relying solely on Mean Opinion Scores (MOS) for naturalness to evaluating the multimodal speech generation system within an application context. A notable application demonstrated is joke delivery. In these evaluations we used chatGPT to generate the jokes, enhancing realism and demonstrating, where the system was found to be on par with commercial TTS systems in terms of performance. This showcases the system's potential for applications such as stand-up robot performances, and highlights the importance of evaluating ECAs in real-world contexts.

9 ACKNOWLEDGMENTS

This research is supported by the Swedish Research Council projects Connected (VR-2019-05003), STANCE (VR-2020-02396) and StyleBot (VR-2018-05409); the Riksbankens Jubileumsfond project CAPTivating (P20-0298) and the Digital Futures project Advanced Adaptive Intelligent Systems (AAIS).

REFERENCES

[1] Elisabeth André, Thomas Rist, and Jochen Muller. 1999. Employing AI methods to control the behavior of animated interface agents. *Applied Artificial Intelligence* 13, 4-5 (1999), 415–448.

[2] Matthew P Aylett, Benjamin R Cowan, and Leigh Clark. 2019. Siri, Echo and performance: You have to suffer darling. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–10.

[3] RS Aylett, S Louchart, J Dias, A Paiva, and M Vala. 2005. FearNot!–an experiment in emergent narrative. In *Proc. IVA*. 305–316.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020), 12449–12460.

[5] Linda Bell and Joakim Gustafson. 2003. Child and Adult Speaker Adaptation during Error Resolution in a Publicly Available Spoken Dialogue System. In *Proc. Eurospeech*. 613–616.

[6] Elisabetta Bevacqua, Ken Prepin, Radoslaw Niewiadomski, Etienne de Sevin, and Catherine Pelachaud. [n. d.]. Greta: Towards an interactive conversational virtual companion. ([n. d.]).

[7] Timothy W Bickmore, Lisa Caruso, Kerri Clough-Gorr, and Tim Heeren. 2005. 'It's just like you talk to a friend'relational agents for older adults. *Interacting with Computers* 17, 6 (2005), 711–735.

[8] Justine Cassell. 1999. Embodiment in conversational interfaces: Rea. In *Proc. SIGCHI conference on Human Factors in Computing Systems*. 520–527.

[9] Justine Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22, 4 (2001), 67–67.

[10] Li-Wei Chen, Shinji Watanabe, and Alexander Rudnicky. 2023. A Vector Quantized Approach for Text to Speech Synthesis on Real-World Spontaneous Speech. *arXiv preprint arXiv:2302.04215* (2023).

[11] Michelle Cohn and Georgia Zellou. 2020. Perception of concatenative vs. neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes. In *Proc. Interspeech*.

[12] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization ability of MOS prediction networks. In *Proc. ICASSP*. 8442–8446.

[13] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech communication* 50, 8-9 (2008), 630–645.

[14] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. 2016. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–11.

[15] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proc. IEEE Int. Conference on Computer Vision*. 5784–5794.

[16] Joakim Gustafson, Johan Boye, Morgan Fredriksson, Lasse Johanneson, and Jürgen Königsmann. 2005. Providing computer game characters with conversational abilities. In *Proc. IVA*. 37–51.

[17] Joakim Gustafson, Nikolaj Lindberg, and Magnus Lundeberg. 1999. The August spoken dialogue system. In *Proc. Eurospeech*.

[18] Joakim Gustafson and Kåre Sjölander. 2004. Voice creation for conversational fairy-tale characters. In *Proc. SSW*.

[19] Mohammed Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proc. UbiComp*. 697–706.

[20] Ahmed Hussen Abdelaziz, Anushree Prasanna Kumar, Chloe Seivwright, Gabriele Fanelli, Justin Binder, Yannis Stylianou, and Sachin Kajareker. 2021. Audiovisual Speech Synthesis using Tacotron2. In *Proc. ICMI*. 503–511.

[21] Patrik Jonell, Mattias Bystedt, Per Fallgren, Dimosthenis Kontogiorgos, José Lopes, Zofia Malisz, Samuel Mascarenhas, Catharine Oertel, Eran Raveh, and Todd Shore. 2018. Farmi: a framework for recording multi-modal interactions. In *Proc. LREC)*.

[22] Heather Knight, Scott Satkin, Varun Ramakrishna, and Santosh Divvala. 2011. A savvy robot standup comic: Online learning through audience tracking. In *Workshop paper (TEI'10)*.

[23] John Kominek and Alan W Black. 2004. The CMU Arctic speech databases. In *Proc. SSW*.

[24] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*, Vol. 33. 17022–17033.

[25] 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*. 119–127.

[26] Stefan Kopp, Lars Gesellensetter, Nicole C Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide–design and evaluation of a real-world application. In *Proc. IVA*. 329–343.

[27] Katharina Kühne, Martin H Fischer, and Yuefang Zhou. 2020. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in Neurorobotics* (2020), 105.

[28] Harm Lameris, Shivam Mehta, Gustav Eje Henter, Joakim Gustafson, and Éva Székely. 2023. Prosody-controllable spontaneous TTS with neural HMMs. *Proc. ICASSP* (2023).

[29] Yinghao Aaron Li, Cong Han, and Nima Mesgarani. 2022. StyleTTS: A Style-Based Generative Model for Natural and Diverse Text-to-Speech Synthesis. *arXiv preprint arXiv:2205.15439* (2022).

[30] Björn Lindblom. 1983. Economy of speech gestures. In *The production of speech*. Springer, 217–245.

[31] Matthew Marge, Carol Espy-Wilson, Nigel G Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé III, Debadeepta Dey, et al. 2022. Spoken language interaction with robots: Recommendations for future research. *Computer Speech & Language* 71 (2022), 101255.

[32] Sven EG Öhman. 1967. Numerical model of coarticulation. *The Journal of the Acoustical Society of America* 41, 2 (1967), 310–320.

[33] Stefan Olafsson, Teresa K O'Leary, and Timothy W Bickmore. 2020. Motivating health behavior change with humorous virtual agents. In *Proc. IVA*. 1–8.

[34] Raquel Oliveira, Patricia Arriaga, Minja Axelsson, and Ana Paiva. 2021. Humor–Robot interaction: a scoping review of the literature and future directions. *International Journal of Social Robotics* 13 (2021), 1369–1383.

[35] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[36] Ye Pan, Ruisi Zhang, Shengran Cheng, Shuai Tan, Yu Ding, Kenny Mitchell, and Xubo Yang. 2023. Emotional Voice Puppetry. *IEEE Transactions on Visualization and Computer Graphics* 29, 5 (2023), 2527–2535.

[37] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

[38] Tuomo Raitio, Ramya Rasipuram, and Dan Castellani. 2020. Controllable Neural Text-to-Speech Synthesis Using Intuitive Prosodic Features. *Proc. Interspeech* (2020), 4432–4436.

[39] Jeff Rickel and W Lewis Johnson. 1999. Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence* 13, 4-5 (1999), 343–382.

[40] Pejman Sajjadi, Laura Hoffmann, Philipp Cimiano, and Stefan Kopp. 2018. On the effect of a personality-driven ECA on perceived social presence and game experience in vr. In *Proc. VS-Games*. 1–8.

[41] Katie Seaborn, Norihisa P Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in human–agent interaction: a survey. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–43.

[42] 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*. 4779–4783.

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

[44] William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, et al. 2010. Ada and Grace: Toward realistic and engaging virtual museum guides. In *Proc. IVA*. 286–300.

[45] Éva Székely, Jens Edlund, and Joakim Gustafson. 2020. Augmented Prompt Selection for Evaluation of Spontaneous Speech Synthesis. In *Proc. LREC*. 6368––6374.

[46] Éva Székely, Gustav Eje Henter, and Joakim Gustafson. 2019. Casting to corpus: Segmenting and selecting spontaneous dialogue for TTS with a CNN-LSTM speaker-dependent breath detector. In *Proc. ICASSP*. 6925–6929.

[47] Éva Székely, Siyang Wang, and Joakim Gustafson. 2023. So-to-Speak: an exploratory platform for investigating the interplay between style and prosody in TTS. In *Proc. Interspeech*.

[48] Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.

[49] Sean Thomas, Ylva Ferstl, Rachel McDonnell, and Cathy Ennis. 2022. Investigating how speech and animation realism influence the perceived personality of virtual characters and agents. In *Proc. IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. 11–20.

[50] Kristinn R Thórisson. 1997. Gandalf: an embodied humanoid capable of real-time multimodal dialogue with people. In *Proc. of the First International Conference on Autonomous Agents*. 536–537.

[51] Guanzhong Tian, Yi Yuan, and Yong Liu. 2019. Audio2face: Generating speech/face animation from single audio with attention-based bidirectional lstm networks. In *Proc. ICMEW*. IEEE, 366–371.

[52] Christiana Tsiourti, Emilie Joly, Cindy Wings, Maher Ben Moussa, and Katarzyna Wac. 2014. Virtual assistive companions for older adults: qualitative field study and design implications. In *Proc. PervasiveHealth*. 57–64.

[53] Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang. 2020. Emotional speech synthesis with rich and granularized control. In *Proc. ICASSP*. 7254–7258.

[54] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. 2020. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *Proc. ICASSP*.

[55] Laurens Van Der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[56] Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, et al. 2019. Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. In *Proc. SSW*.

[57] Siyang Wang, Simon Alexanderson, Joakim Gustafson, Jonas Beskow, Gustav Eje Henter, and Éva Székely. 2021. Integrated Speech and Gesture Synthesis. In *Proc. ICMI*. 177–185.

[58] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*. 5180–5189.

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

[60] Preben Wik and Anna Hjalmarsson. 2009. Embodied conversational agents in computer assisted language learning. *Speech Communication* 51, 10 (2009), 1024–1037.

[61] Rohola Zandie, Mohammad H Mahoor, Julia Madsen, and Eshrat S Emamian. 2021. RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis. In *Proc. Interspeech*. 2751–2755.

[62] Hongbo Zhu, Chuang Yu, and Angelo Cangelosi. 2023. Affective Human-Robot Interaction with Multimodal Explanations. In *Proc. ICSR*. Springer, 241–252.