No More Mumbles: Enhancing Robot Intelligibility through Speech Adaptation

Qiaoqiao Ren[1], Yuanbo Hou[2], Dick Botteldooren[2], and Tony Belpaeme[1]

***Abstract*—Spoken language interaction is at the heart of interpersonal communication, and people flexibly adapt their speech to different individuals and environments. It is surprising that robots, and by extension other digital devices, are not equipped to adapt their speech and instead rely on fixed speech parameters, which often hinder comprehension by the user. We conducted a speech comprehension study involving 39 participants who were exposed to different environmental and contextual conditions. During the experiment, the robot articulated words using different vocal parameters, and the participants were tasked with both recognising the spoken words and rating their subjective impression of the robot's speech. The experiment's primary outcome shows that spaces with good acoustic quality positively correlate with intelligibility and user experience. However, increasing the distance between the user and the robot exacerbated the user experience, while distracting background sounds significantly reduced speech recognition accuracy and user satisfaction. We next built an adaptive voice for the robot. For this, the robot needs to know how difficult it is for a user to understand spoken language in a particular setting. We present a prediction model that rates how annoying the ambient acoustic environment is and, consequentially, how hard it is to understand someone in this setting. Then, we develop a convolutional neural network model to adapt the robot's speech parameters to different users and spaces, while taking into account the influence of ambient acoustics on intelligibility. Finally, we present an evaluation with 27 users, demonstrating superior intelligibility and user experience with adaptive voice parameters compared to fixed voice.**

***Index Terms*—Human-Centered Robotics, Design and Human Factors, Social HRI**

I. Introduction

THERE is the expectation that social robots will one day be capable of engaging in human-like spoken language interaction with people. However, achieving effective spoken interaction with devices such as robots and digital assistants presents a significant challenge, especially in environments with background noise and suboptimal acoustic properties. The parameters that govern a robot's vocal characteristics—such as volume, speech rate, pitch, and emphasis—can impact speech intelligibility and the overall user experience [1].

Earlier research has underscored the detrimental effects of background noise on speech intelligibility [2]. The reverberation time of a space, which strongly correlates with its acoustic quality, has been found to negatively affect speech intelligibility [3]. Additionally, the distance between the listener and the sound source has a crucial impact on speech intelligibility and on the user experience[4]. Earlier studies have identified pitch, pitch range, volume, and speech rate as fundamental vocal characteristics conveying personality [5]. For instance, voice pitch has been effectively employed to model a robot's personality and emotion [6]. Speech speed influences speech intelligibility, and Jones *et al.* found that a faster speaking rate lowers comprehension for listeners [7]. Except for the critical role of acoustic factors, individual differences in listeners are also important. Different listeners vary in their ability to understand speech in noisy environments [8]. As such, it is crucial to consider the listener's personal information when designing the robot's adaptivity.

This work advances the field of data-driven adaptive speech, building upon a foundation of prior research that has explored various facets of the challenge, which has distinct advantages, particularly when addressing complex, variable environments and tailoring individual user needs. People adapt their speech during spoken conversation. This is known as the Lombard effect, which is an involuntary speech adaptation where speakers naturally modify their speech in noisy environments to improve communication. In response to loud environments, speakers will increase their speech volume, adjust pitch, and enunciate more, thus optimising interaction between interlocutors. However, robots lack this natural adaptability, highlighting the need for advanced systems to mimic the Lombard effect to improve human-robot interaction (HRI). Earlier studies have looked into adapting speech to noisy environments and the distance between the user and the robot [9], addressing the needs of individuals with hearing impairments [10], or scaling the loudness of the voice dynamically when approaching a user [11]. Niculescu *et al.* find that pitch adaption of a robot's voice impacted the users' rating of overall interaction quality [6]. However, most robots (and other interactive devices for that matter) do not take the acoustic properties of the environment, the ambient sound, or the user's characteristics into account during the conversation. The large majority of spoken language interactions with robots use fixed voice parameters, often ignoring the dynamic nature of environmental and user factors.

Our study addresses three research questions (RQs) - RQ1: Is there a relationship between robot speech parameters, user characteristics, acoustic quality of the environment and ambient sound, and the robot's intelligibility as well as user experience? RQ2: Can the robot use information about the environment and the user to adapt its speech parameters? RQ3: Can the robot's speech parameters be dynamically adapted to optimise the user's experience and the robot's intelligibility?

To address RQ1, we collect data to map how intelligibility suffers under different conditions in Section II. From this, Section III responds to RQ2 by building an automated system to adapt the robot's voice to changing user and environmental factors. Finally, we evaluate our system in a user study in Section IV to address RQ3. Discussion is given in Section V.

[1]Qiaoqiao Ren and Tony Belpaeme are with the Faculty of Engineering and Architecture, IDLab-AIRO, Ghent University – imec, Technologiepark 126, 9052 Gent, Belgium (e-mail: {Qiaoqiao.Ren, Tony.Belpaeme}@UGent.be).

[2]Yuanbo Hou and Dick Botteldooren are with the Faculty of Engineering and Architecture, WAVES, Ghent University, Belgium (e-mail: {Yuanbo.Hou, Dick.Botteldooren}@UGent.be).

The source code, pre-trained model and demos are available at: https://github.com/qiaoqiao2323/robot-speech-intelligibility

II. Intelligibility assessment

A. Experimental design

1) Materials: To reveal how ambient sound, the environment's acoustic quality, the distance between the user and the robot, and the user's hearing can affect both the robot's intelligibility and the user's experience, we first set up a data collection campaign. We used a speech recognition task using a paradigm from [12]. The task relies on open-set recognition of four lists of English consonant-nucleus-consonant (CNC) monosyllabic words (e.g. *hash*, *dodge*, *should*, ...) from the Northwestern University Auditory Test 6 (NU-6) [13]. The words were spoken by an NAO V5 humanoid robot (United Robotics Group, formerly known as Aldebaran Robotics). During the task, a recording of ambient sound (ranging from soft chatter to the sound of power tools) was played through a speaker placed near the robot. The voice parameters of the robot (i.e., volume, speech rate, pitch, and emphasis) were drawn using a uniform distribution from predetermined parameter ranges, thus sampling the voice parameter space.

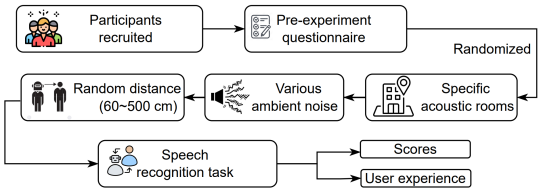

Fig. 1. Experimental procedure.

2) Procedure: Participants were recruited via a social media campaign in the local area, excluding those with self-reported reading-related learning difficulties, to ensure data quality. The data collection and study adhered to the ethics procedures of the *Universiteit Gent*. Participants gave informed consent and were told they could withdraw at any time without providing a reason. The experiment process is shown in the Fig. 1. A pre-test questionnaire gathered personal information, including gender and age, their self-evaluated hearing difficulties in noisy environments, and their self-reported Common European Framework of Reference for Languages (CEFR) levels as their English levels. 39 participants (17 identified as female, 20 as male, 2 as other; 28.0 ± 3.7 years old) took part in the data collection exercise. Each participant was randomly assigned to one of six different rooms, each having its own acoustic quality. The participants were exposed to three different ambient sounds drawn from a total of 11; complete silence was used as a baseline condition. After, the participants were placed at a random interaction distance to the robot, ranging from 60 to 500 *cm*. Each participant is exposed to 50 words per round; During the task, the robot spoke a word picked randomly from the NU-6 list and the participant was asked to type the word using a keyboard, along with providing feedback on the user experience. Fig. 2 shows examples of different settings in which data was collected.

Next, we build a model to examine how different factors influence the intelligibility of the robot's speech and the user experience. As independent variables, we have two participant characteristics, four robot voice parameters, and three environmental factors. As dependent variables, we collected similarity scores and user experience from the participants per word. Therefore, each word is a tuple consisting of the independent variables and dependent variables. We collected 5442 tuples in total.

Fig. 2. Illustration of different environments in which data was collected.

B. Factors influencing intelligibility

1) Participant characteristics: We collect data on participants' hearing loss and English language level. Individuals with normal hearing already experience difficulties in comprehending speech in noisy environments, and this is exacerbated when people experience hearing loss. We therefore ask participants to self-report hearing loss. Moreover, listening in adverse conditions presents more challenges for non-native listeners in a foreign language context [14]. Thus, we also record their English proficiency level. Specifically, participants rated their difficulty hearing in noisy environments, such as restaurants or social gatherings, using a Likert scale (1: No, never, to 5: Yes, always), which has been used in previous research [15] and rated their English proficiency using the CEFR scale, which rates English proficiency from A1 (1: beginner) to C2 (6: fluent). This yields distributions for hearing difficulty (2.2 ± 1) and English proficiency level (4.8 ± 0.8).

2) Robot voice parameters: The second data category relates to four specific parameters associated with the robot's vocal characteristics: speed, pitch, emphasis, and volume. These features play a crucial role in determining the clarity and intelligibility of the robot's speech. We study their impact on participants' experiences and speech intelligibility.

The words are produced using the built-in US English speech engine of the NAO robot, which is supplied by Nuance. The designers of NAO state that the default voice is the voice of a machine but with a fluidity that makes it very comfortable to listen to [16], and the default voice has been widely used in research studies [17]. We modulate four key parameters in the voice engine: volume, speed, pitch, and emphasis (which is implemented by double voice synthesis and controlled by the *doubleVoiceLevel* parameter, which adds a second voice over the first voice). The volume ranges between [0, 2], and the pitch ranges between [0.5, 2], with a default of 1. Speed is set as a multiple of the default rate of 100%, ranging between [0.4, 4], and governs the pace at which the speech is articulated. Lastly, the *doubleVoiceLevel*, ranging between [0, 4], grants control over the gain of a second voice; 0 disables the effect. In the word recognition task, all the voices are played with NAO's loudspeakers set to 100%. The default voice settings for the voice parameters—volume, speed, and pitch are all set at 1, with the double voice effect deactivated.

3) Environmental factors: The third category of data comprises three environmental factors. The reverberation time (T30), the participants' annoyance rating (AR) of the ambient noise, and the distance of the individuals to the robot.

Reverberation time is a key metric for assessing a room's acoustic quality. The T30 is a standard metric and reports the time it takes for the sound pressure level to decrease by 60 decibels [18], as detailed in the ISO3382 standard. To measure the T30, an impulse noise (a popping balloon)

was used, and the reverberation time was measured using a Svantek SVAN 959 Sound and Vibration Analyser. During the measurement, the impulse source and microphone were 1.0m above floor level, and we measured at three positions for each room (the middle of the room and two diagonal corners). The T30 measures were averaged over a frequency range from 250Hz to 2kHz based on the three measurements per room [19]. The measured results are given in Fig. 3. For the data collection and evaluation, we used different spaces, each with varying reverberation times (in seconds(s)). Data was collected in an office setting, which includes an open-plan area (T30 = 0.78s), meeting room (T30 = 0.56s), and a home setting (the imec Homelab) consisting of a living room (T30 = 0.43s), bedroom (T30 = 0.19s) and meeting room (T30 = 0.32s), and an anechoic chamber (T30 = 0.04s) used for echo-free sound assessment. Those environments offer a stable setting without disturbance from uncontrollable variables like crowds.

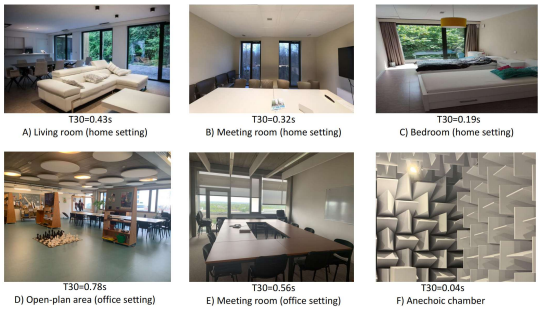

Fig. 3. Reverberation time (T30) of different rooms.

Next to the different rooms, we also make use of different ambient sounds. The subjective perception of these sounds– and specifically the extent to which these sounds are perceived as annoying, which is correlated with the volume of the sounds– is expected to influence the participant's performance on tasks [20]. The ambient noise played in the experiment is from the DeLTA dataset, which includes 2980 samples of 15-second binaural audio recordings with 24 classes of sound sources and human-annotated annoyance ratings [21]. The annoyance rating ranges from 1 (not annoying at all) to 10 (very annoying). All the ambient sounds are played at the same speaker volume, guaranteeing a consistent playback level. For the robot to quantify the subjective experience of ambient sounds, it needs a model that maps a sound recording into an annoyance rating. Important here is that the model not only looks at the amplitude of the ambient sound but also takes a holistic view of how annoying the sound is.

Finally, the distance participants keep from the robot will impact the robot's intelligibility. During data collection, we sampled distances ranging from 60 to 500 *cm*.

C. Intelligibility metrics

The intelligibility metrics encompass both user experience and speech intelligibility.

1) User experience: Participants' user experience related to the robot's voice is also collected. This is rated from 1 (not pleasant at all) to 10 (very pleasant). These measures help gauge the affective and subjective aspects of the interaction.

2) Speech intelligibility: Participants were tasked with typing the words spoken by the robot. We evaluated the robot's speech intelligibility by measuring the phonetic similarity score between the spoken words and the words typed by the participants. This compensates for homophones, such as *flour* and *flower*, and gives us a scalar metric for how poorly the word was understood. We used a phonetic embedding method to calculate the phonetic similarity of test and input words from participants, as in [22]. This phonetic embedding represents each phonetic component as a vector in a continuous vector space, thus encoding phonetic information, considering phoneme alignment in the words and capturing intricate phonetic characteristics and enhancing similarity assessment.

D. Generalised linear mixed-effect models

Based on the independent variables collected in Section II-B and dependent variables from Section II-C, we build a generalised linear mixed-effect model (GLMM) to study the contributions of various factors to spoken language intelligibility and to the user experience, including fixed effects, interact effects and random effects. We use the method by [23], which can handle non-normal distributed response variables. A Variance Inflation Factors analysis revealed an absence of multicollinearity among the independent variables - all factors were below the threshold (≤ 2). In addition, the correlations between the annoyance rating and acoustic quality, as well as the user information, are all smaller than 0.1.

The gamma distribution is a continuous probability distribution to model data that is skewed, continuous and non-negative [24], such as the residuals of the phonetic scores and user experience that we observed. To choose the most accurate model, we compare the prediction error using the Akaike Information Criterion (AIC) of the reverse link function and the log link function. We chose the log link function, as it obtained the lowest AIC.

The generalised model examines the relationship between predictor variables and both the intelligibility and user experience, as represented by Eq 1 and Eq 2. To adequately account for the non-independence of observations stemming from repeated measurements across subjects and the inherent variability in word difficulty, we have integrated subject ID and word ID as random effects in our models to control for intra-subject variability and variable difficulty levels of the words tested.

$$
\begin{aligned}
&\mathrm{SP}_{ij} \sim \mathrm{Gamma}(\mu_{ij}, \phi) \\
&\log(\mu_{ij} + \mu) = \beta_0 \cdot \mathrm{v}_{ij} + \beta_1 \cdot \mathrm{p}_{ij} + \beta_2 \cdot \mathrm{e}_{ij} + \beta_3 \cdot \mathrm{s}_{ij} \\
&\quad + \beta_4 \cdot \mathrm{A}_{ij} + \beta_5 \cdot \mathrm{T}_{ij} + \beta_6 \cdot \mathrm{D}_{ij} + \beta_7 \cdot \mathrm{E}_{ij} + \beta_8 \cdot \mathrm{H}_{ij} \\
&\quad + \beta_9 \cdot (\mathrm{v} \cdot \mathrm{T}) + 1|pn_{ij} + 1|w_{ij} + \epsilon_1 \\
&\epsilon_1 \sim \mathrm{Normal}(0, \sigma^2)
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
&\mathrm{UX}_{ij} \sim \mathrm{Gamma}(\lambda_{ij}, \phi) \\
&\log(\lambda_{ij}) = \lambda_0 \cdot \mathrm{v}_{ij} + \lambda_1 \cdot \mathrm{p}_{ij} + \lambda_2 \cdot \mathrm{e}_{ij} + \lambda_3 \cdot \mathrm{s}_{ij} \\
&\quad + \lambda_4 \cdot \mathrm{A}_{ij} + \lambda_5 \cdot \mathrm{T}_{ij} + \lambda_6 \cdot \mathrm{D}_{ij} + \lambda_7 \cdot \mathrm{E}_{ij} + \lambda_8 \cdot \mathrm{H}_{ij} \\
&\quad + \lambda_9 \cdot (\mathrm{v} \cdot \mathrm{T}) + 1|pn_{ij} + 1|w_{ij} + \epsilon_2 \\
&\epsilon_2 \sim \mathrm{Normal}(0, \sigma^2)
\end{aligned} \tag{2}
$$

Eq 1 and Eq 2 use the coefficients β_0 to β_9 and λ_0 to λ_9 respectively to represent the estimated impact of predictor variables on the response variables (speech intelligibility, user experience). They show how each predictor is expected to log-linearly impact the response, and these effects are fixed across the dataset, not subject to random variation within the model's framework. Specifically, the coefficients β_0 to β_3 , λ_0 to λ_3 correspond to the voice attributes (volume, pitch, emphasis, and speed) of the robot on speech intelligibility and user

experience, respectively, while β_4 to β_6 , (λ_4 to λ_6) represent factors such as room characteristics (AL, T30) and distance to the robot. English proficiency of participants and their auditory challenges in noisy environments are signified by β_7 to β_8 and λ_7 to λ_8 for user's intelligibility and user experience, respectively. The coefficients β_9 pertain to the interaction of volume and the room's acoustic quality and the influence on both speech intelligibility and user experience. Similarly, λ_9 represents the interactions in the context of user experience. Simultaneously, the unaccounted-for residuals that evade fixed and random effects are encapsulated by the residual component ϵ. The subscript ij indicates the jth observation within the ith group. The response variable SP_{ij}, UX_{ij} embodies speech intelligibility and user experience in this context, respectively, and μ is 0.0001.

We used the SIMR package in R for post-hoc power analysis, based on the 1000 Monte Carlo simulations [25]. As the log link function was used in GLMM, we did the exponential transform of the coefficients (β_0 to β_9 , λ_0 to λ_9) to get the estimate of the fixed effect for speech intelligibility and user experience.

E. Intelligibility model results

1) Speech intelligibility model analysis: The negative estimate of -0.75 associated with *annoyance rating* underscores the adverse impact of heightened annoyance on speech intelligibility ($p < 0.001$), the post-hoc power is 98.60%. Additionally, the *acoustic quality of the space (T30)* has a significant negative estimate of -0.77 with speech intelligibility ($p < 0.01$), the post-hoc power is 66.90%; this finding signifies that good acoustic conditions of a space are conducive to comprehension of the robot's speech. Additionally, the *speed* has a significant negative estimate of -0.21 ($p < 0.001$) of speech intelligibility, which obtained 100% post-hoc power. Finally, *pitch* has a negative estimate of -0.71 ($p < 0.001$), which implies that the high pitch leads to decreased intelligibility with 91.70% post-hoc power. We observed some weak estimates without statistical significance ($p > 0.05$). For instance, there is a negative correlation between the *distance*, *emphasis* as well as *hearing difficulty* and speech intelligibility. In terms of *English proficiency level*, which serves as an indicator of language skills, our analysis revealed a marginally positive coefficient.

We explored the interplay between the *T30* of spaces and the *volume* of the robot's voice and its impact on speech intelligibility, the results suggest a significant positive estimate of 1.44 interaction effect between room characteristics and the volume of the robot's voice ($p < 0.001$) with 76.20% power, indicates that the relationship between volume and speech intelligibility is influenced by the quality of room acoustics, and the optimal volume for improve the robot speech intelligibility depending quality of the spatial acoustic, which might explain why the post-hoc power for *T30* is relatively lower. This may indicate that the effect size of it is trivially small compared with the interaction effect. The random effect attributed to subjects and words ID accounts for smaller than 1% of the overall variance for speech intelligibility. This indicates that the influence of the random effect, including the subject and word ID, on the total variability of speech intelligibility is modest.

2) User experience model analysis: The results showed that the predictor *annoyance rating* exhibited a negative estimate of -0.74 ($p < 0.001$) and the post-hoc power is 83.00%, signifying that heightened annoyance ratings due to ambient sounds correspond to a decrease in user experience. Likewise, the predictor *pitch* displayed a significant negative coefficient of -0.89 ($p < 0.05$) with 65.6% post-hoc power, indicating that elevated pitch levels are associated with lower user experience scores. This suggests that an increased pitch might have an adverse impact on user experience during such interactions. The predictor *distance* also yielded a significant negative coefficient of -0.74 ($p < 0.01$) with 95.7% post-hoc power, implying that greater distances from the sound source are linked to lower user experience scores. Furthermore, the predictor *T30* of space revealed a negative coefficient of -0.67 with statistical significance ($p < 0.01$) and the post-hoc power is 81.30%. This implies that the quality of the room environment significantly influences user experience, with a decrease in room quality corresponding to a decrease in user experience. Surprisingly, the *English proficiency level of users exhibited a significant negative* estimate of -0.54 with user experience ($p < 0.01$) with 75.20% post-hoc power. This implies that individuals with higher language proficiency experienced the robot's speech as less pleasant. This might be attributed to their heightened expectations regarding word intelligibility. In addition, the predictor *speed* ($p < 0.001$) displayed a significant negative coefficient of -0.56 with 99.6% post-hoc power, indicating that faster speech rates result in a poorer user experience.

Interestingly, the predictor *hearing difficulty* in noisy environments did not predict user experience ($p > 0.05$). This suggests that participants who reported higher hearing difficulty experienced a significantly better user experience listening to robot speech. In contrast, the predictor *emphasis* showed a negative estimate with user experience in this analysis ($p > 0.05$), although it was not statistically significant. In the context of user experience, the direct effect of volume on user experience showed a significant negative relationship (fixed estimate of -0.92, $p < 0.01$), suggesting that higher volume levels are generally perceived negatively. However, this finding comes with an important caveat as the statistical power for this effect was notably low (18.8%), indicating potential limitations in the robustness of this result. Our finding indicated the significant interaction between room acoustics and the robot's volume (coefficient 1.32, $p < 0.001$), with a robust post-hoc power of 83.3%. The positive interaction coefficient indicates that the optimal volume for enhancing user experience shifts depending on the acoustic quality of the room. This interaction effect potentially explains the low power observed in the direct effect of *volume*, which shows that the robot voice should be adapted to the environment to get better user experience and intelligibility. The random effect attributed to individual subjects and word ID accounts for less than 5% of the overall user experience variance. This suggests the user experience is similar for all subjects and all words.

III. Building an adaptive robot voice

One of the key takeaways from our analysis is the importance of adapting the robot's voice to the environment and user. To create an adaptive robot voice, we required a model to assess ambient sound annoyance. Then, we constructed an adaptive speech model that leverages the annoyance rating, and together with other parameters– the distance to the robot, the T30 reverberation of the room, the user's English proficiency, and hearing difficulties– adjusts the robot's speech parameters.

A. Ambient sounds' annoyance rating prediction

As ambient sound is known to strongly impact the intelligibility of speech and user experience, we require an automated method to rate the annoyance of ambient sounds. This not

only takes into account the amplitude of the sound, but also takes into account the frequency spectrum of ambient sound to predict the annoyance rating (AR). This is a scalar value used as an overall evaluation metric to reflect the impact of ambient sounds on participants' ability to understand speech. Fig. 4 shows the convolutional neural network (CNN)–based annoyance rating prediction (ARP) model we use for this.

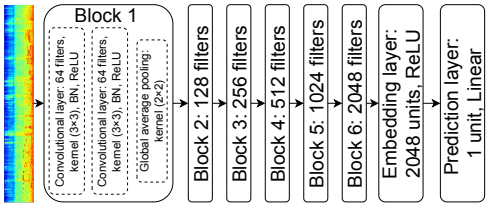

Fig. 4. The CNN-based ARP model for overall evaluation of ambient sounds.

Inspired by the outstanding performance of the convolution-based sound source classification model [26], the ARP model uses 6 convolutional blocks of the same structure but with a sequentially increasing number of filters. Each convolutional block consists of two convolutional layers equipped with ReLU activation functions and a global average pooling layer [27]. To stabilise and speed up the neural network training, batch normalisation (BN) [28] is introduced. After the convolutional blocks, an embedding layer consisting of fully connected layers with a dimension of 2048 is used to transform ambient sound representations extracted by convolutional blocks into high-level embeddings suitable, for the final prediction layer of the whole environmental sound.

1) Ambient sound ARP model training: To train the ARP model to successfully infer the impact of environmental sounds on participants, i.e., the annoyance rating for participants, based on acoustic features containing amplitude, frequency, and category information of ambient sounds, we trained the model on the DeLTA real-life polyphonic audio dataset [21]. The DeLTA contains 2980 samples of 15-second binaural audio recordings with 24 classes of sound sources from European cities for noise annoyance detection. Participants rated 15-second binaural recordings of urban environments by providing an annoyance rating on a scale of 1 to 10.

As shown in Fig. 4, given the prediction output from the final prediction layer is $\hat{y_{ar}}$, and its corresponding label is $y_{ar} \in [1, 10]$, the mean squared error (MSE) is used as the loss function for the ambient sound ARP model, $Loss = MSE(\hat{y_{ar}}, y_{ar})$. To comprehensively consider polyphonic audio information such as the amplitude and frequency of ambient sound sources, the acoustic feature *log Mel* that performs well in sound-related tasks is adopted. *Log Mel*-filter 64-bank spectrogram is extracted by the Short-Time Fourier Transform with a Hamming window length of 46*ms* and a window overlap of $1/3$ [29], which means that the duration of each frame is $0.031s$, and the acoustic features can capture the acoustic information of sounds with a time resolution of $0.031s$. In training, dropout and normalisation are used to prevent over-fitting of the model. The Adam optimiser [30], with an initial learning rate of 0.005, minimises the loss function, with a batch size of 64. The model is trained for 100 epochs. In inference, using a Tesla V100-SXM2-32GB GPU and Intel(R) Xeon(R) CPU E5-2698 v4@2.20GHz as an example, the response time of the model from inputting acoustic features to outputting prediction results is 1.87*ms*, which means the model can process input sounds in real-time.

2) Ambient sound ARP model performance: We used 2200 clips from the DeLTA dataset to train the model, retaining 245 clips for validation and 445 for testing. Table I shows the performance of the proposed CNN-based ARP model on the test set, and also shows the performance of several other typical neural network models. Among them, the deep neural network (DNN) consists of four fully connected (FC) layers and the ARP layer. Each FC layer's number of units is 64, 128, 256, and 512, respectively. The final FC layer's output is flattened and input to the ARP layer. The Simple CNN consists of 4 convolutional layers, each equipped with (3×3) kernels, and the ARP layer. Each convolutional layer's filter numbers are 64, 128, 256, and 512, respectively. The final convolutional layer's output is flattened and input to the ARP layer. The CNN-Transformer consists of 3 convolutional layers with (3×3) kernels, a Transformer encoder [31], and the ARP layer. In Table I, the results of the simple CNN are slightly better than those of the DNN, illustrating the effectiveness of the convolutional layer in extracting information such as the amplitude and frequency of ambient sounds from the input acoustic features. The CNN-Transformer, equipped with a Transformer encoder suited for temporal modelling, performs better than a simple CNN. Finally, the proposed CNN model achieved more competitive results with lower MSE and MAE. This shows that the proposed CNN-based ARP model can effectively predict the impact of ambient sound based on its comprehensive information on the participants.

TABLE I
THE ARP PERFORMANCE ON THE TEST SET OF THE DELTA DATASET.

Model	DNN	Simple CNN	CNN-Transformer	Proposed CNN
MSE	1.73	1.67	1.45	**1.10**
MAE	1.01	1.00	0.97	**0.83**

B. Proposed adaptive speech model

Our findings emphasise that it is crucial to tailor the robot's voice to the specific environment to achieve both optimal speech intelligibility and a positive interaction experience for different users. We trained a neural network to predict optimal robotic voices based on user and environment characteristics. The network is trained only on data for which participant speech recognition was completely accurate and user satisfaction was greater than 5. This excludes data of negative user experiences or where the robot's speech was poorly understood. We call this model the Environment-to-Voice model (ETV). It takes environmental factors as input, including annoyance ratings of ambient sound (reported by the ARP model), the distance to the robot, the English proficiency of the user, the user's difficulty hearing in noisy environments, and the room's acoustic quality. The model's output consists of robot voice parameters, which consist of volume, speed, pitch, and emphasis. The network has 5 input nodes, 2 hidden layers with a ReLU activation function (16 and 32 neurons, respectively) and a 4-neuron output layer with a linear activation function.

Model training: Data augmentation is used to expand the dataset and introduce noise. We use the Adam optimiser [30], with a default learning rate of 0.0001. The model trains for 200 epochs, achieving a minimal MSE of 0.15.

We evaluated the adaptive robot voice using the method reported in section II-A. The ARP model is used to predict environmental annoyance ratings via the robot. Notably, due to the use of a lightweight model, the computation, despite running on the CPU, only requires 2 seconds. Subsequently,

the predictions are communicated to a pre-trained ETV model. This predictive data, together with user-provided information on hearing difficulties and English proficiency and assessments of acoustic quality, serves as input for the ETV model, which then generates adaptive voice parameters. The robot speech adaptive system is visually depicted in Fig. 5.

IV. Evaluation

A. Evaluation experiment design

To assess the generalisation of our proposed ETV model with unseen data, we conducted an evaluation experiment involving 27 participants (14 identified as female, 13 as male; 26.6 ± 2.4 years old, Hearing difficulty: 2.4 ± 0.6, English proficiency level: 4.8 ± 0.92). The experiment followed a within-subject design, using two conditions - *(1) Adaptive Speech*: The robot's speech parameters were dynamically adjusted based on the environmental and spatial settings. *(2) Fixed Speech*: The robot uses its default voice parameters.

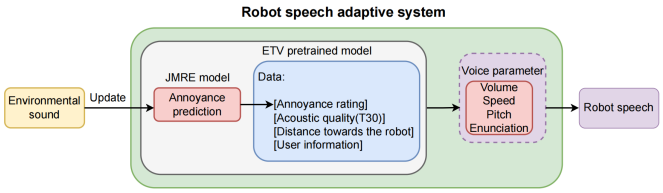

Fig. 5. Robot speech adaptive system.

Each participant experienced both conditions in a balanced random order. The evaluation experiment took place in two rooms with T30 values of 0.04 and 0.56, respectively. Participants were asked to type the words spoken by the robot within one of these two rooms. Within each condition, there were two sessions featuring distinct ambient sounds, which are high annoyance rating (AR) ($AR \geq 5$) and low AR ($AR < 5$). In each session, the robot spoke 15 words randomly drawn from the NU-6 list.

B. Fixed robot voice versus adaptive voice

In our within-subject design study, the phonetic similarity scores and user experience data did not follow a normal distribution; hence we used the non-parametric Wilcoxon Signed-Rank Test; this analysis yielded an effect size (r), demonstrating that the users' intelligibility of the robot's speech is significantly better ($Z = -3.79, p < 0.001$) for the adaptive robot voice than for the fixed robot voice. The adaptive robot voice also leads to a significantly better user experience ($Z = -2.99, p = 0.003$). This can be seen in Fig. 7, Fig. 6 and Tab II; The post-hoc power is calculated using G*Power [32] with 27 participants. The phonetic similarity scores post-hoc power is 97.5% with $\alpha = 0.05$ and $r = 0.73$, and user experience post-hoc power is 88.8% with $\alpha = 0.05$ and $r = 0.58$. The horizontal line that splits the box in two is the median, which in Fig. 7 coincides with the top line; the mean is indicated by the red dot. The top and bottom boundaries of the box indicate the 25th and 75th percentiles, respectively.

TABLE II
Means and Standard Deviations for Phonetic Similarity with Fixed and Adaptive Robot Voice Parameters

AR (1-10)	Voice parameters adaptive Vs fixed	Scores Mean ± Std	User experience Mean ± Std
All	fixed	0.73 ± 0.31	6.27 ± 2.42
	adaptive	0.79 ± 0.29	6.60 ± 2.43
Low AR	fixed	0.78 ± 0.28	6.60 ± 2.36
	adaptive	0.81 ± 0.28	6.86 ± 2.32
High AR	fixed	0.67 ± 0.33	5.93 ± 2.43
	adaptive	0.76 ± 0.30	6.34 ± 2.52

C. Fixed robot voice under different annoyance ratings environment sound

As all the participants experienced the two different kinds of ambient sounds, i.e. low annoyance and high annoyance sounds, the Wilcoxon Signed-Rank Test was used to evaluate the differences between fixed robot voice and adaptive voice under high AR and Low AR, respectively. The results can be seen in the Tab. II. For high AR, we observed that the users rated significantly higher pleasantness for adaptive voice than the fixed default voice setting ($Z = -4.01, p < 0.001$) and significantly better scores are observed for the adaptive voice than the default voice ($Z = -3.82, p = 0.009$). However, there is no significant difference between the adaptive voice and the default voice on the scores under the low AR. Interestingly, the users rate the adaptive voice to be significantly more pleasant than the default voice ($Z = -2.33, p = 0.043$). This indicates that in the case of low ambient noise, although the adaptive sound does not improve the intelligibility, it does improve the user experience.

V. Discussion

A. Recommendations

As expected, there is a relationship between robot speech parameters, user characteristics, environment factors, and the robot's intelligibility as well as user experience. Specifically,

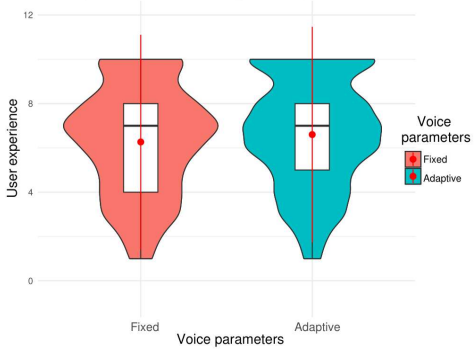

Fig. 6. User experience ratings for default robot voice (left) and the adaptive voice (right).

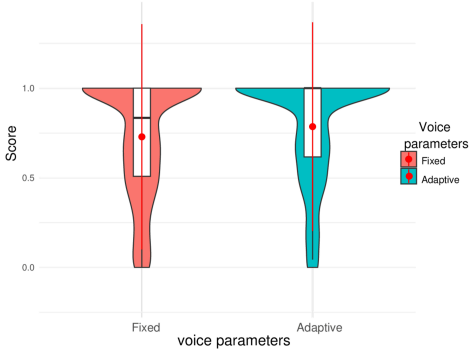

Fig. 7. Intelligibility of the robot, expressed as phonetic similarity scores, when it used the default voice (left) or its adaptive voice (right).

we find that annoying sounds in the environment make it harder to understand the robot's speech, which in turn hurts the user experience. This emphasises the need for adaptive robot speech in noisy settings. In addition, the room's acoustic quality influences the intelligibility and the user experience, highlighting the need to take the room into account. Moreover, increased distance will worsen the user experience. Therefore, it might be interesting to further explore the proper distance for spoken conversation between a robot and its user. The results show that, as expected, the default and therefore nonadaptive robot voice will lead to subpar speech intelligibility and user experience when there is ambient sound in the environment. In brief, a robot voice that does not adapt to the user and the environment is a missed opportunity.

Regarding the factors related to the user, unsurprisingly, English proficiency matters. People understand the English words spoken by the robot better when they are more proficient in English. Surprisingly, they report having a lower interaction experience, which we hypothesise might be explained by their having stricter pronunciation expectations for the robot. Similarly, the user's hearing difficulties will exacerbate their understanding of the robot. These outcomes strongly suggest that user characteristics should be taken into consideration when setting the robot's voice and speech parameters.

As for robot speech parameters, a higher volume will help intelligibility, but worsen the user experience. Unsurprising, as shouting makes you heard, but it is not pleasant. Slow speech will improve both the user experience and the intelligibility. Surprisingly, robots with higher pitch values lead to poorer satisfaction and lower intelligibility. One possible explanation is that users may find sharp voices unpleasant or may not appreciate the emotional implications conveyed by higher pitch levels. For instance, earlier research noted that menacing voices often exhibit an increase in both volume and pitch. Additionally, increased pitch levels are commonly associated with expressions of anger [33]. Surprisingly, better emphasis —implemented through adding a double voice — had a small negative influence on both intelligibility and user experience. This effect could be attributed to the anomaly of emphasising a single word without a context.

Designing adaptive, user-centric robot systems that consider the individual user and environmental conditions is essential for effective communication in HRI. This requires the robot to be informed about the environment and the user, ideally in an automated way. The distance between the user and the robot was now manually entered into the model, but could, of course, be extracted from an RGBD camera. T30 reverberation estimates could also be approximated by the robot through speech [34] during the conversation or by playing an impulse sound when entering a new environment and analysing the reverberation. The ARP model predicts the impact of ambient noise. Other factors, such as the user's language proficiency or potential hearing difficulties are harder to automatically extract but could be collected during the making of a user profile. In addition, from the evaluation experiment, we can conclude that the adaptive robot voice optimises the user's experience and the robot's intelligibility than its default voice.

B. Limitations

This study has certain limitations. First, the data we collected suffers from a class imbalance across predictors, as people with hearing difficulties or very low English proficiency levels are underrepresented. This imbalance can introduce bias into the model's outcomes. Secondly, the participants may not be representative of the broader population, thus limiting the generalizability of the findings. Future studies could draw from a more diverse user population. The experimental design, involving both within-subject and between-subject factors, introduces an additional complexity as not all group conditions for environmental factors are fully covered within the study. In addition, even though the self-reported hearing difficulties or English levels has been used in previous research, it might still be influenced by individuals's perception.

The speech recognition task relies on recognising single words, which is a difficult task. Embedding words in a sentence would improve intelligibility due to the availability of context. Our data was collected under controlled conditions, and a range of factors typical of real-world HRI scenarios, such as the presence of other speakers or contextual cues, were not taken into account. Our findings may not fully extend to more complex scenarios. However, it serves as a foundation for optimizing robots for more complicated scenarios.

One participant noted that it might be advantageous to inquire about participants' listening and English *writing* abilities specifically, as the task exclusively involves listening and writing rather than an assessment of overall English proficiency; in addition, the different participant's accent or background could influence their perception of words, especially for non-native speakers, who might struggle with understanding short phrases and single words. Finally, the position of the robot, whether on the left or right side of the participants, may influence their intelligibility of the robot's speech.

C. Conclusion

Given our results, it is difficult to understand why virtually all current robot voices do not adapt to the environmental context and the diversity in users. Our research sheds light on how adaptive voices can have a positive impact on both the user experience and speech intelligibility. Simply put, shouting louder to be better understood is not the solution. Instead, a judicious balance is needed between making the robot understood and not irritating the user. In light of these, we presented the ETV model that seeks to enhance robot speech intelligibility by predicting and setting user-adapted and contextually suitable robot voice parameters, tailored to users, and to spatial and environmental conditions. Important to note is that the use of a data-driven model alleviates the need to use heuristics to change the robot's voice.

REFERENCES

[1] Robinson *et al.*, "Designing sound for social robots: Candidate design principles," *International Journal of Social Robotics*, vol. 14, no. 6, pp. 1507–1525, 2022.

[2] K. Ishikawa, S. Boyce, L. Kelchner, M. G. Powell, *et al.*, "The effect of background noise on intelligibility of dysphonic speech," *Journal of Speech, Language, and Hearing Research*, vol. 60, no. 7, pp. 1919–1929, 2017.

[3] M. Klatte *et al.*, "Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting," *Noise and Health*, vol. 12, no. 49, p. 270, 2010.

[4] Walters and o, "Human approach distances to a mechanical-looking robot with different robot voice styles," in *Proc. of RO-MAN*, 2008, pp. 707–712.

[5] W. Apple *et al.*, "Effects of pitch and speech rate on personal attributions." *Journal of personality and social psychology*, vol. 37, no. 5, p. 715, 1979.

[6] A. Niculescu, B. van Dijk, A. Nijholt, H. Li, and S. L. See, "Making social robots more attractive: the effects of voice pitch, humor and empathy," *International Journal of Social Robotics*, vol. 5, pp. 171–191, 2013.

[7] C. Jones, L. Berry, and C. Stevens, "Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners," *Computer Speech & Language*, vol. 21, no. 4, pp. 641–651, 2007.

[8] A. Heinrich, H. Henshaw, and M. A. Ferguson, "The relationship of speech intelligibility with hearing sensitivity, cognition, and perceived hearing difficulties varies for different speech perception tests," *Frontiers in psychology*, vol. 6, p. 782, 2015.

[9] E. Martinson and D. Brock, "Improving human-robot interaction through adaptation to the auditory scene," in *Proceedings of ACM/IEEE HRI*, 2007, pp. 113–120.

[10] G. Lindley, "Adaptation to loudness: Implications for hearing aid fittings." *The Hearing Journal*, vol. 52, no. 11, pp. 50–52, 1999.

[11] K. Fischer, L. Naik, *et al.*, "Initiating human-robot interactions using incremental speech adaptation," in *Proceedings of ACM/IEEE HRI*, 2021, pp. 421–425.

[12] J. R. Dubno, D. D. Dirks, and D. E. Morgan, "Effects of age and mild hearing loss on speech recognition in noise," *The Journal of the Acoustical Society of America*, vol. 76, no. 1, pp. 87–96, 1984.

[13] T. W. Tillman and R. Carhart, *An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University Auditory Test No. 6*. USAF School of Aerospace Medicine Brooks Air Force Base, TX, 1966.

[14] M. L. G. Lecumberri *et al.*, "Non-native speech perception in adverse conditions: A review," *Speech communication*, vol. 52, no. 11-12, pp. 864–886, 2010.

[15] T. M. Mikkola, H. Polku, E. Portegijs, M. Rantakokko, T. Rantanen, and A. Viljanen, "Self-reported hearing status is associated with lower limb physical performance, perceived mobility, and activities of daily living in older community-dwelling men and women," *Journal of the American Geriatrics Society*, vol. 63, no. 6, pp. 1164–1169, 2015.

[16] Y. Kali, M. Saad, J.-F. Boland, J. Fortin, and V. Girardeau, "Walking task space control using time delay estimation based sliding mode of position controlled nao biped robot," *International Journal of Dynamics and Control*, vol. 9, pp. 679–688, 2021.

[17] Tozadore *et al.*, "Wizard of oz vs autonomous: Children's perception changes according to robot's operation condition," in *IEEE RO-MAN*, 2017, pp. 664–669.

[18] Paini *et al.*, "Is reverberation time adequate for testing the acoustical quality of unroofed auditoriums?" *Proceedings of Ins. Ac.*, vol. 28, no. 2, pp. 66–73, 2006.

[19] M. S. Dobreva, W. E. O'Neill, and G. D. Paige, "Influence of aging on human sound localization," *Journal of neurophysiology*, vol. 105, no. 5, pp. 2471–2486, 2011.

[20] D. Västfjäll, "Influences of current mood and noise sensitivity on judgments of noise annoyance," *The Journal of psychology*, vol. 136, no. 4, pp. 357–370, 2002.

[21] A. Mitchell, M. Erfanian, C. Soelistyo, T. Oberman, *et al.*, "Effects of soundscape complexity on urban noise annoyance ratings: A large-scale online listening experiment," *International Journal of Environmental Research and Public Health*, vol. 19, no. 22, p. 14872, 2022.

[22] R. Sharma, K. Dhawan, and B. Pailla, "Phonetic word embeddings," *arXiv preprint arXiv:2109.14796*, 2021.

[23] J. A. Nelder *et al.*, "Generalized linear models," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 135, no. 3, pp. 370–384, 1972.

[24] Moran *et al.*, "New models for old questions: generalized linear models for cost prediction," *Journal of evaluation in clinical practice*, vol. 13, no. 3, pp. 381–389, 2007.

[25] P. Green and C. J. MacLeod, "Simr: An r package for power analysis of generalized linear mixed models by simulation," *Methods in Ecology and Evolution*, vol. 7, no. 4, pp. 493–498, 2016.

[26] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM TASLP*, vol. 28, pp. 2880–2894, 2020.

[27] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with relu activation function," *The Annals of Statistics*, vol. 48, no. 4, pp. 1875–1897, 2020.

[28] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 7705–7716.

[29] Y. Hou, B. Kang, W. Van Hauwermeiren, and D. Botteldooren, "Relation-guided acoustic scene classification aided with event embeddings," in *Proc. of International Joint Conference on Neural Networks*, 2022, pp. 1–8.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, *et al.*, "Attention is all you need," in *Proc. of International Conference on Neural Information Processing Systems*, 2017, pp. 5998–6008.

[32] M. E. Bartlett, C. Edmunds, T. Belpaeme, and S. Thill, "Have I got the power? analysing and reporting statistical power in HRI," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 11, no. 2, pp. 1–16, 2022.

[33] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *The Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–1108, 1993.

[34] R. Ratnam, D. L. Jones, B. C. Wheeler, W. D. O'Brien Jr, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.