Acceptance of synthetic speech in South African languages: A comparative study of Afrikaans, isiZulu, and Sepedi in healthcare contexts

Johannes Abraham Louw and Ilana Wilken
Natural Language Processing Research Group
Council for Scientific and Industrial Research
Pretoria
South Africa
jalouw@csir.co.za and iwilken@csir.co.za

Abstract

While text-to-speech technologies have made significant advances in recent years, questions remain about how synthesised speech is accepted in culturally and linguistically diverse settings such as South Africa. This study explores how South Africans perceive synthetic speech in comparison to human-recorded speech across three official languages: Afrikaans, isiZulu, and Sepedi, with healthcare as the application context.

Using a blind and randomised listening test, 65 participants rated audio prompts across four acceptance metrics: trust, knowledgeability, likability, and relatability. Statistical analysis using the Wilcoxon signed-rank test revealed no significant difference between natural and synthesised speech perception among Afrikaans speakers. However, low participation rates prevented meaningful analysis of speech perception for isiZulu and Sepedi speakers. When combining data from all participants, a medium effect size favouring natural speech was observed, though this difference was not statistically significant.

These findings suggest that synthetic speech adapted from natural recordings may be suitable for certain applications in South Africa, though larger and more linguistically representative samples are needed to confirm these results.

1 Introduction

In recent years, the development of synthetic speech generation technologies has made significant strides, driven by advances in machine learning, neural vocoding, and large-scale text-to-speech (TTS) modelling. Modern TTS systems now produce synthetic speech that emulate many of the acoustic and prosodic nuances of natural human speech, leading to growing interest in their deployment across a variety of applications. Although substantial technical progress has been made, a key challenge still hinders mass adoption: determining user attitudes toward and acceptance of synthesized speech in everyday contexts. This is particularly pertinent in the South African context, where uptake of speech technology by large enterprises has been cautious, hindered in part by uncertainties regarding user acceptance, trust, and the perceived authenticity of synthetic speech and voices.

This study presents findings from an empirical investigation into how users respond to synthetic speech within specialized domains. Combining methodologies from human-computer interaction and digital humanities, we conducted controlled experiments comparing user reactions to human versus synthetic-generated audio.

The experiment controlled for speaker identity by applying cross-lingual transfer learning to create synthetic voices that replicated the original speakers. This design isolated speech origin as the sole variable, allowing us to test whether participants responded differently to human recordings versus their synthetic counterparts from the same speakers.

The central research question guiding this work is: *How do South Africans perceive synthesized speech when used in specific use cases, such as healthcare communication?* We conducted a series of blind listening tests with native speakers. In these tests, participants were exposed to audio prompts: some produced using high-quality TTS generated from the same speaker as the natural voice recordings, and others through natural voice recordings themselves. Participants weren't informed of each sample's origin. This methodology allows for a controlled investigation into whether synthesized speech is deemed acceptable in specific practical scenarios.

Our findings aim to contribute to the broader discourse on the intersection of artificial intelligence and the humanities by grounding technical evaluations of TTS systems in human-centric empirical

evidence. Furthermore, this work seeks to support a more informed and confident adoption of speech technologies in South Africa by identifying domains where current synthetic speech quality is already sufficient to meet user expectations.

2 Background

South Africa's cultural and linguistic diversity presents a unique challenge for the development and deployment of speech technologies. With twelve official languages (eleven spoken and South African Sign Language) and a population characterised by varying levels of literacy, education, and technological exposure, it is evident that the acceptance of new technologies, such as synthetic speech, cannot be assumed to be uniform across the population. While technology has become deeply embedded in daily life for many globally, South Africans' historical and socioeconomic context means that digital adoption remains uneven. As such, the successful integration of TTS systems into local applications requires not only technical efficacy but also sociocultural alignment and trust.

Trust and cultural resonance are particularly critical in contexts involving emotionally sensitive information, such as healthcare. As highlighted by (Weber et al., 2008), voice interfaces hold promise for improving accessibility among low-literate populations, given their lower barrier to entry and alignment with widespread mobile phone use. However, building acceptance for such interfaces demands more than accessibility alone. Users' willingness to trust and engage with synthetic voices is shaped by subtle social cues, perceived empathy, and cultural appropriateness (Rau et al., 2009).

Decades of research have shown that humans can exhibit social responses to machines if minimal social cues are present (Rau et al., 2009). This is particularly relevant for speech interfaces, where the synthetic voice acts as the auditory "face" of a system. Yet, this brings the danger of approaching the so-called *uncanny valley*, a term introduced by Mori to describe the discomfort experienced when a system or robot closely, but not perfectly, mimics human behaviour (Mori et al., 2012). In speech, this can occur when a synthetic voice is perceived as nearly human but fails in subtle ways, potentially undermining trust or causing unease, especially among users with limited prior exposure to such systems.

The primary purpose of synthetic speech is to support clear, trustworthy, and culturally respectful communication, not to mislead. This is especially relevant in multicultural societies like South Africa, where emotive responses to speech, whether real or synthesised, can vary significantly across linguistic and cultural groups. As argued by (Wagner et al., 2019), "just like clothes do not fit every person alike," TTS systems cannot be developed with a one-size-fits-all approach. Instead, acceptability and perceived quality must be evaluated in context-specific ways, guided by sociolinguistic sensitivities.

From a methodological perspective, the evaluation of TTS systems has traditionally relied on three categories: objective, subjective, and behavioural assessments (Wagner et al., 2019). While objective metrics provide reproducible benchmarks, they often correlate poorly with human perception. Subjective evaluations, typically involving listener ratings of naturalness, intelligibility, and likability, remain the most reliable for gauging real-world performance. However, these too require careful experimental design. For instance, (Wester et al., 2015) critique existing evaluation practices and propose a checklist of best practices to improve the meaningfulness of subjective TTS testing. Their guidelines highlight the importance of selecting appropriate listener groups, contexts, test types, and questions to ensure robust and interpretable results.

Despite growing attention to TTS quality metrics, few studies have focused on the social acceptance of synthetic speech, particularly within the South African context. The Qfrency TTS engine[1], for example, provides synthetic voices in eleven of the country's official languages, and while technical evaluations of these voices have been conducted, there has been little investigation into how everyday users perceive and respond to these voices in practice.

This lack of empirical user feedback is especially problematic in developing regions, where the acceptance of new technologies tends to lag behind that of developed countries (Weber et al., 2008). Even when the technical barrier is reduced, as with mobile phone interfaces or voice systems, the absence of cultural trust or familiarity with the technology may inhibit adoption. It is therefore important to identify which domains and user profiles are more receptive to synthetic speech, and to determine the conditions under which its deployment

[1] http://qfrency.com

can be socially and ethically appropriate.

The study reported in this work seeks to address this gap by exploring the acceptability of synthetic speech in domain-specific applications, using a blind listening test methodology. This investigation aims to provide a more nuanced understanding of how South Africans perceive synthesised speech, with a focus on emotionally sensitive settings such as healthcare.

3 Methodology

The methodology is structured around four components: (a) the selection and formulation of acceptance metrics and their associated questions; (b) the preparation of text prompts and audio recordings; (c) the voice adaptation process used to generate synthetic speech; and (d) the experimental setup, including participant exposure and response collection.

3.1 Acceptance Metrics and Questionnaire Design

The evaluation of synthetic speech in this study is grounded in four perceptual metrics, chosen to reflect key socioemotional and cognitive factors that influence speech acceptability: *trust*, *knowledgeability*, *likeability*, and *relatability*. These metrics are informed by prior work on human-robot interaction and technology-mediated communication, notably the study by (Rau et al., 2009), which explored how communication style and cultural context affect receptivity to artificial agents.

Each metric is associated with two semantically related but syntactically distinct questions to ensure response reliability while minimising response priming. Participants are asked both questions associated with each metric, once after listening to a synthetic voice and once after listening to a human-recorded voice. The parallel phrasing is designed to elicit comparable information while reducing the likelihood that participants recognise repeated constructs.

Trust

- Question 1: "I would feel comfortable sharing this audio with family."
- Question 2: "I feel like I can rely on the speaker to tell the truth."

Knowledgeability

- Question 1: "The speaker probably went to university."
- Question 2: "The speaker sounds experienced on the topic."

Likability

- Question 1: "I think I like the speaker."
- Question 2: "The speaker sounds like a nice person."

Relatability

- Question 1: "The speaker sounds similar to the people from my community."
- Question 2: "The speaker sounds like someone I could have a friendly conversation with."

3.2 Text Selection and Audio Preparation

The selected domain for the initial experimental phase is healthcare, motivated by the research group's prior involvement in health-related TTS applications, such as the AwezaMed platform (Marais et al., 2020) and ongoing collaboration with the African Health Research Institute. This domain is particularly relevant for studying acceptability due to the emotive and sensitive nature of healthcare communication.

To reduce bias and avoid content-based influence on user perception, all textual prompts were carefully curated to remain domain-relevant yet emotionally neutral. Based on established practices in TTS evaluation, the texts are kept short, ranging between 10 and 15 words in English, to avoid listener fatigue. Evaluation sessions are designed to last no longer than 30 minutes per participant, balancing the need for sufficient exposure with practical time constraints.

3.3 Recording of Human Voices and Synthesis Process

To ensure linguistic authenticity and dialectal fidelity, native speakers of the target languages who are also members of the research group were selected to record the speech prompts. The data collection phase focused on three languages: Afrikaans, isiZulu, and Sepedi.

The recorded speech material forms the basis for constructing synthetic voices using cross-lingual transfer learning methods (Louw, 2023). These techniques make it possible to develop high-quality synthetic voices from a relatively small amount of speech data per speaker, which is especially valuable in resource-constrained contexts. Because

each synthetic voice is generated from the same individual who provided the original recordings, the approach supports a controlled experimental design in which the effects of synthesis on perception can be isolated from other variables. The methodology used for cross-lingual voice construction is described in the next sub sections.

3.3.1 Step 1: Phonological Feature Representation

The first step consists of transforming the linguistic input to the TTS model into a phonological feature (PF) representation, where phonological features refer to a set of articulatory and acoustic properties that together provide a more abstract and linguistically grounded encoding of speech sounds. Each input utterance is processed through a phonological encoder, which maps either graphemic or phonemic transcriptions to a binary vector of phonological features. These features are selected to capture articulatory and acoustic aspects of speech sounds, including place and manner of articulation, voicing, nasality, and other language-independent attributes. This step ensures that the input to the model is structured in a way that facilitates cross-lingual generalization and enhances the model's ability to learn transferable phonological patterns.

3.3.2 Step 2: Pretraining on a Resource-Rich Language

A foundational model is trained on a large-scale corpus from a resource-rich language, in this case English. The training utilizes a modified VITS (Kim et al., 2021) architecture, in which the standard text encoder is replaced with a sequential feed-forward network capable of embedding binary PF vectors. The decoder adopts a Multi-Band inverse Short-Time Fourier Transform (MB-iSTFT) mechanism for efficient waveform reconstruction. The model is trained using both adversarial (Goodfellow et al., 2020) and reconstruction objectives, incorporating normalizing flows (Rezende and Mohamed, 2015) to improve output quality.

3.3.3 Step 3: Fine-Tuning on Resource-Scarce Data

The pre-trained model is fine-tuned using human-recorded samples of the speech prompts. Recordings corresponding to the domain-specific text (see Section 3.2) were excluded from the data used for synthetic voice building. Table 1 summarizes the total duration and number of utterances for each language and data subset employed in building the voices in this study.

Table 1: Number of utterances and total duration (hh:mm:ss.ms) per language used to build the synthetic voices.

Language	# Utterances	Duration
Afrikaans	341	00:22:44.81
isiZulu	341	01:01:43.39
Sepedi	352	00:26:48.63

The sentences of the recordings were transcribed in International Phonetic Alphabet (Association, 1999) (IPA) and converted to PF vectors using the same encoding scheme as in the pretraining phase. The fine-tuning process enables the model to learn target-language-specific prosody and phonetic realization while leveraging shared PF representations to bootstrap performance.

Once the model is fine-tuned, the system is capable of synthesizing speech from IPA or grapheme-to-PF converted text.

3.4 Experimental Design

During the evaluation sessions, participants were exposed to a randomised sequence of audio prompts. Each prompt was presented either in its original human-recorded form or in a synthesised version, with careful attention given to balancing the number of samples across both conditions. Participants were not informed whether they were listening to a natural or a synthetic voice, thereby enabling a blind assessment of speech acceptability across the defined perceptual metrics.

The experimental setup was designed to collect both quantitative and qualitative data for comparative analysis. Participant responses were linked to the type of voice (synthetic or human) and to the relevant acceptance metric. In addition to these responses, demographic data were also collected. This included information such as participants' home language, age, and prior exposure to speech technologies, which allowed for subgroup analyses and contextual interpretation of the results.

Before the experiment could be conducted, it was first implemented on a web-based evaluation platform. The questionnaire was deployed using `webMUSHRA` (Schöffler et al., 2018), a *Multiple Stimuli with Hidden Reference and Anchor* (MUSHRA) (ITU-R, 2001) experimental framework that operates through a Web Audio application programming interface (API). This platform allowed for

the creation of multi-page experimental workflows, including sections for informed consent and the core experimental questions.

Participants were provided access to a questionnaire that corresponded to the home language they selected at the start of the experiment. Both the audio prompts and the associated evaluation questions were randomised using the functionality provided by webMUSHRA. This meant that no two participants were likely to hear the audio files in the same order. The randomisation also applied to the distribution of the two questions (Section 3.1) per metric across human and synthetic voice conditions, ensuring balanced exposure and reducing order bias.

Each participant listened to a total of eight audio files and answered one acceptance metric question per file. The randomisation process ensured that all four metrics were evaluated for both the human-recorded and synthetic voices, without alerting the participants to any pattern in the experimental design.

3.5 Participant Selection

For the study, the selected evaluation languages were Afrikaans, isiZulu, and Sepedi. These languages were prioritised based on the availability of native speakers within the research group as well as existing collaborations in related projects. The remaining official South African languages may be evaluated in subsequent phases of the study.

Participants were selected based on whether their home language matched one of the three evaluation languages. Eligible participants were required to be aged 18 years or older, and no restrictions were placed on gender.

According to (Wester et al., 2015), more than 30 participants are generally required for the robust evaluation of synthetic speech systems. In line with this recommendation, the study aimed to recruit a minimum of 30 participants per language group, resulting in a target sample size of at least 90 participants across the three languages.

Recruitment was conducted using a single primary channel: email invitations were circulated to colleagues within the Council for Scientific and Industrial Research (CSIR) community, requesting their voluntary participation and encouraging them to share the invitation further.

Each invitation included a brief description of the study's aims and a link to the online evaluation platform. Upon following the link, participants were presented with a detailed information page outlining the purpose of the study, ethical considerations, and data handling practices. This enabled individuals to make an informed decision regarding their participation.

4 Results

A total of 65 participants had completed the evaluation, distributed across the three selected languages as follows: 27 Afrikaans-speaking participants, 22 isiZulu-speaking participants, and 16 Sepedi-speaking participants.

Participant responses were automatically logged by the webMUSHRA platform in comma-separated values (CSV) files. These files contained raw response data, which were subsequently processed and structured for analysis.

4.1 Data Processing and Analysis

To enable structured analysis, a set of Python scripts was developed to parse the CSV files and associate each response with the corresponding voice condition (natural or synthesised) and acceptance metric. This ensured accurate alignment of perceptual responses with the correct utterance types.

Responses were captured using a four-point Likert-type scale (Likert, 1932) with the following numeric mappings:

- Strongly Agree: 2
- Agree: 1
- Disagree: -1
- Strongly Disagree: -2

Each response was converted to its corresponding numeric value and categorised according to whether it was associated with a human-recorded or synthesised audio stimulus. For each of the 16 evaluation statements (comprising four perceptual metrics, two phrasing variants per metric, and two voice types), responses were tallied separately by voice condition.

The median value across all participants was calculated per condition to provide a descriptive measure of central tendency in perceptual agreement with the statements. The median was selected because Likert-type responses are ordinal rather than interval data; although the categories have an inherent order, the perceptual distance between adjacent points cannot be assumed to be equal. The median thus provides a robust representation of

central tendency without requiring interval-scale assumptions, unlike the mean.

Importantly, the Wilcoxon Signed-Rank test was conducted using the individual paired response values, not the medians. The medians served as a descriptive summary of participant tendencies, while the inferential analysis used all paired ordinal data to determine whether statistically significant differences existed between the natural and synthesised conditions.

4.2 Statistical Analysis

To test whether there were significant differences in participant perceptions between natural and synthesised voices, the Wilcoxon signed-rank test (Wilcoxon, 1945) was applied. This non-parametric test was selected due to its suitability for analysing paired samples where normal distribution cannot be assumed. A two-tailed test was used with a significance level of $\alpha = 0.05$.

The test produces three key statistics:

- **Z-value**: A standardized test statistic indicating how far the observed difference is from zero. A positive value ($Z > 0$) suggests that synthetic speech was generally rated higher than natural speech, while a negative value ($Z < 0$) suggests that synthetic speech was generally rated lower.

- **p-value**: The probability of observing the data assuming there is no true difference. A p-value below 0.05 is typically considered statistically significant.

- **Effect size (r)**: Measures the magnitude of the difference, calculated as $r = \frac{Z}{\sqrt{N}}$, where N is the number of pairs. Values around 0.1 indicate a small effect, 0.3 a medium effect, and 0.5 or higher a large effect. The sign of r mirrors that of Z and indicates the direction of the effect.

These statistics together inform whether a difference exists, its significance, and its practical importance.

4.2.1 Afrikaans Group Results

For the Afrikaans-speaking participants, the test revealed no statistically significant difference between natural and synthetic speech conditions. The Wilcoxon signed-rank test yielded a test statistic of $Z = 0.11$, a p-value of $p = 0.915$, and an effect size of $r = 0.04$. These results indicate a negligible difference in participant perceptions between the two voice types, suggesting that, for Afrikaans listeners, the synthetic voices were broadly comparable to their natural counterparts in terms of acceptability.

4.2.2 isiZulu and Sepedi Group Results

The number of responses from isiZulu-speaking and Sepedi-speaking participants was insufficient for statistical testing. As such, no Wilcoxon signed-rank tests were performed for these two groups. Future iterations of the study will prioritise increasing sample sizes for these languages to enable robust statistical evaluation.

4.2.3 Combined Group Results

When combining data across all participants, the analysis again showed no statistically significant difference between natural and synthetic speech. The Wilcoxon test yielded $Z = -1.31$, $p = 0.190$, and an effect size of $r = -0.38$. While the p-value remains above the threshold for significance, the moderate effect size suggests that there may be perceptual differences warranting further investigation with a larger and more balanced sample.

5 Conclusion

The findings suggest that, for Afrikaans-speaking participants, there is only a very small and statistically non-significant difference between natural and synthesized speech in terms of perceived trust, knowledgeability, likability, and relatability. While the results for isiZulu and Sepedi speakers could not be analysed due to an insufficient number of responses, the combined dataset across all participants indicated a medium effect size in favour of natural speech, but this too was not statistically significant.

These preliminary results imply that current state-of-the-art synthetic speech, when adapted from natural recordings, may already be suitable for certain high-impact applications in South Africa, particularly in linguistically and technologically diverse contexts. However, limitations in participant numbers, especially for underrepresented languages, restrict the generalisability of these conclusions.

Moreover, it must be acknowledged that the participant group likely shared similar demographic characteristics, such as education level, employment type, and technological exposure. This in-

troduces potential bias and limits the extent to which the findings can be generalised to the broader speaker communities.

Future work should focus on expanding the sample size and improving representation from all official South African languages. Moreover, the synthesis techniques themselves could be refined further and evaluated under a broader range of use cases. A larger and more diverse study would allow for greater statistical power, more reliable comparisons across linguistic groups, and deeper insight into the sociolinguistic factors influencing the perception of synthetic speech in multilingual societies.

Ethical Considerations

Prior to commencement, ethical clearance for the study was obtained from the CSIR Research Ethics Committee. Given that the study involved human participants and data collection in an online environment, informed consent was a requirement.

Participation in the experiment was entirely voluntary and anonymised. No identifiable personal information was collected, and no compensation was offered for participation. To ensure transparency, the informed consent form was embedded within the first page of the web-based questionnaire. Participants were required to indicate their consent before continuing to the rest of the survey.

References

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, 1 edition. Cambridge University Press, Cambridge.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative Adversarial Networks. *Communications of the ACM*, 63(11):139–144.

Recommendation ITU-R. 2001. BS. 1534-1. Method for the subjective assessment of intermediate sound quality (MUSHRA). *International Telecommunications Union, Geneva.*

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Rensis Likert. 1932. A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22(140):1–55.

Johannes Abraham Louw. 2023. Cross-lingual transfer using phonological features for resource-scarce text-to-speech. In *Proceedings of the 12th Speech Synthesis Workshop (SSW)*, Grenoble, France.

Laurette Marais, Johannes A Louw, Jaco Badenhorst, Karen Calteaux, Ilana Wilken, Nina Van Niekerk, and Glenn Stein. 2020. AwezaMed: A multilingual, multimodal speech-to-speech translation application for maternal health care. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.

Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The Uncanny Valley. *IEEE Robotics & automation magazine*, 19(2):98–100.

PL Patrick Rau, Ye Li, and Dingjun Li. 2009. Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior*, 25(2):587–595.

Danilo Rezende and Shakir Mohamed. 2015. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, Lille, France. PMLR.

Michael Schöffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. 2018. webMUSHRA – A Comprehensive Framework for Web-based Listening Tests. *Journal of Open Research Software*, 6(1):8.

Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, G Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, and 1 others. 2019. Speech synthesis evaluation-state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th Speech Synthesis Workshop (SSW)*, Vienna, Austria.

Frederick Weber, Kalika Bali, Roni Rosenfeld, and Kentaro Toyama. 2008. Unexplored directions in spoken language technology for development. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1–4, Goa, India. IEEE.

Mirjam Wester, Cassia Valentini-Botinhao, and Gustav Eje Henter. 2015. Are we using enough listeners? no! – an empirically-supported critique of interspeech 2014 tts evaluations. In *Proc. INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association*, pages 3476–3480, Dresden, Germany.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.