Usability of Text-to-Speech Synthesis to Bridge the Digital Divide in South Africa: Language Practitioner Perspectives

Georg I. Schlünz
Human Language Technology Research Group
CSIR Meraka Institute
Pretoria, South Africa
gschlunz@csir.co.za

***Abstract*—We report on two sets of perceptual evaluations of our South African text-to-speech voices by language practitioners. In the first evaluation, we measure baseline quality in terms of how understandable and human-like the voices sound. We also determine baseline usability by asking a series of questions related to accessibility and mainstream application settings. In the second evaluation, we employ the same criteria to compare pronunciation improvements against the baseline. The results indicate success in many areas, but also illuminate room for improvement in others, especially in the cases of the African languages.**

***Index Terms*—Digital divide, accessibility, usability, text-to-speech synthesis, South African languages, perceptual evaluation, language practitioners**

I. Introduction

Text-to-speech (TTS) synthesis, the conversion of electronic text into artificial speech [1], is an important component in assistive technology (AT) to make information and communication more accessible to persons with disabilities. In augmentative and alternative communication (AAC) systems it vocalises messages on behalf of users with little or no functional speech [2]. In screen readers and augmented ebooks it provides an accessible audio narrative of text material for print-disabled users to read [3], [4].

However, the usability of TTS integrated into AT can extend its reach to include not only the disabled community, but also the broader population in a developing country like South Africa that faces multilingual literacy and communication barriers [5]. Riding the wave of the digitalisation of literacy and communication in the 21st century (via web pages, ebooks, apps, etc.), TTS can provide (part of) a solution to reading and learning problems by providing speech as an alternative modality to conventional text. When it is able to cater for the 11 official languages of the country, it becomes an even more empowering tool towards bridging the digital divide.

Essential to the user experience is the ability of TTS voices to synthesise quality speech that is as understandable and natural as possible [1]. The former criterion is formally termed "intelligibility" and measures the correctness with which individual words are pronounced, including word-level (lexical) stress and tone, towards conveying the meaning of the sentence as a whole. The latter criterion, "naturalness", constitutes human-like features such as sentence-level stress (emphasis), intonation (pitch from start to end in phrases), tempo, pausing and breathing.

In previous work [6], we reported on the initial integration of our Speect TTS system [7] (commercialised as "Qfrency TTS" [8]) into an AAC system, screen reader and sample augmented ebooks. We engaged with a small sample of end-users of these applications to determine the usability of the TTS voices in the South African languages. The overall qualitative response was positive, with a pragmatic willingness to adopt the technology, despite certain criteria not yet being met. In the particular case of AAC we were able to quantify the performance more finely using a structured, closed-form survey. Intelligibility scored high for simple test sentences, whereas naturalness were more distributed between the two poles of robotic and human-like. Local language accent was deemed acceptable to good.

In our continued research and development to improve our TTS offering, we aim to increase the number of evaluators for a more representative sample space when doing quality control. We are building feedback mechanisms into the commercial product to obtain input from the market in a more organic fashion, as well as extending networks of collaborators in the accessibility, education, health and publishing sectors to help us out behind the scenes. In the meantime, for this round, we make use of contracted local language practitioners to evaluate our TTS voices. The practitioners can, by virtue of their job description, be viewed as professional custodians of the standards and trends in local language usage. Therefore, we hope that, even in their limited numbers, they can approximate a broad spectrum of the speech and language preferences we can expect to meet with new customers, due to their experience in terms of knowledge and exposure.

This paper relates the perspectives of the language practitioners on the quality of 2 versions of our TTS voices: the baseline and one iteration of pronunciation

TABLE I
TTS VOICE CATALOGUE

Language	Gender	Name	Prompts	Hours
Afrikaans	Female	Maryna	5746	10h58m
Afrikaans	Male	Kobus	4657	07h49m
English	Female	Candice	5251	11h04m
English	Male	Tim	4117	06h43m
Sepedi	Female	Mmapitsi	2609	05h59m
Sepedi	Male	Tshepo	2040	04h41m
Sesotho	Female	Kamohelo	1447	03h25m
Setswana	Female	Lethabo	2337	04h56m
isiXhosa	Female	Zoleka	1705	05h44m
isiXhosa	Male	Vuyo	1580	04h47m
isiZulu	Female	Lindiwe	1708	05h58m
isiZulu	Male	Sifiso	1429	03h30m
isiNdebele	Male	Banele	1454	03h58m
siSwati	Female	Temaswati	1492	05h06m
Tshivenda	Male	Rabelani	2539	04h55m
Xitsonga	Female	Sasekani	0942	01h53m

improvements. Section II describes our voice catalogue, the baseline and the improved versions. Section III is the focus of the paper and expounds on the evaluation methodology and results. Section IV concludes with a summary of findings and recommendations for future work.

II. Implementation

A. Voice Catalogue

We build TTS voices from aligned text and speech training data. The text comprises phonetically-balanced prompts that are read out loud by the voice artists in a professional recording studio. Table I lists the voices in our catalogue with corresponding training data set sizes. They total 16 and cover all 11 official South African languages, with some languages in both genders.

B. Baseline

The Speect/Qfrency TTS system implements statistical parametric speech synthesis. The text frontend performs tokenisation and normalisation [9], grapheme-to-phoneme (G2P) conversion and syllabification, as well as basic prosody prediction using punctuation for phrase breaks and positional and counting features for intonation. The speech backend incorporates the hidden Markov model-based HTS engine [10] to model and synthesise excitation (fundamental frequency), spectrum, voicing strength and duration parameters with a mixed-excitation vocoder.

In particular, the pronunciation lexica of seen words of the baseline TTS voices are based on the Lwazi pronunciation dictionaries [11] that contain roughly 5000 words for each language. The English lexicon is a subset of its roughly 65000-word Lwazi dictionary counterpart. The G2P rules for unseen words are trained on these dictionaries using the Default&Refine algorithm [12].

C. Pronunciation Improvements

In order to increase coverage of seen words, we enlarge the pronunciation lexicon of our English TTS voices using the full 65000-word Lwazi pronunciation dictionary and our Afrikaans TTS voices using the modified RCRL pronunciation dictionary [13], [14] that contains roughly 27000 words. We also extend the lexicons of our African language TTS voices using the NCHLT dictionaries [15], [16] of roughly 15000 words each. This necessitates the standardisation of the phonesets used in the voices, including the creation of phone mappings between the International Phonetic Alphabet (IPA) representations, the Lwazi and NCHLT Speech Assessment Methods Phonetic Alphabet (SAMPA) representations, and old and new internal TTS representations. The resources and methodologies developed in [14], [17] serve as a cross-referencing quality control mechanism.

New G2P rules are trained using the standardised phonesets and updated pronunciation lexica to enable the improved TTS voices to predict the pronunciation of unseen words with greater accuracy.

III. Evaluation

A. Methodology

We perform two sets of evaluations using the contracted language practitioners. The first evaluation requires the practitioners to adjudicate the quality of the baseline TTS voices in an absolute benchmark test. For each language and gender TTS voice, 20 unique phonetically-balanced test sentences are selected from the same domain as the training data and synthesised into audio. Each voice test set is assigned to 5 practitioners, although the same practitioner has to be used in the cases where a language has both a female and a male voice. This bias in the observations is partly mitigated by the unique constitution of each test set. The practitioners must listen to the audio using headphones and cast binary votes (Yes, No) for each of the 20 sentences in an offline spreadsheet, according to certain criteria. The criteria explores aspects of the intelligibility and naturalness of the voices, and are phrased as questions in an absolute sense:

- **Word correct** - Could you hear (distinguish, recognise, identify) most of the words in the sentence? (even if they were not the best quality)
- **Stress/tone** - Are most of the words in the sentence clearly pronounced with the correct lexical stress and/or tone?
- **Sent correct** - Could you understand (follow, interpret) the meaning conveyed by the whole sentence? (even if you had to fill in some gaps)
- **Noise** - Are most of the words in the sentence clearly pronounced with no noisy disturbances like scratches, pops, whistles, etc.?
- **Tempo** - Are most of the words in the sentence pronounced at an acceptable, natural speaking tempo (rhythm, rate)?
- **Intonation** - Does the pitch (intonation) of the voice have an acceptable, natural flow for most of the words from beginning to end in each phrase, as well as most of the phrases from beginning to end in the sentence?

- **Pausing** - Are most of the pauses that are present in the sentence located at appropriate places for an acceptable, natural flow of the speech?
- **Breathing** - Do the breathing sounds of the voice (if any) contribute to an acceptable, natural flow of the speech?
- **Human** - Overall, when the voice speaks this sentence, does it sound more human-like than robotic?

Once the practitioners have finished voting for the sentences, they are required to state their overall conclusions about the voices. These take the form of more binary decisions on the usability of the voices in particular application settings of the technology:

- **First lang** - Overall, is this voice an acceptable representation of a first language speaker?
- **Second lang** - Overall, is this voice an acceptable representation of a second language speaker?
- **News** - If you had to listen to the news (that reports on people, places and events) on a daily basis, instead of reading it silently, would you use this voice to read it out loud for you?
- **Work/School** - If you had to listen to a document at work or school (that contains facts and figures) on a daily basis, instead of reading it silently, would you use this voice to read it out loud for you?
- **Leisure** - If you had to listen to a novel (that dramatises a story for leisure) on a daily basis, instead of reading it silently, would you use this voice to read it out loud for you?
- **No speech** - Imagine you were a person who has no speech (due to a physical impairment or stroke): would this voice be acceptable to use as your communication tool, if it were the only option available? (you type your message in text and the voice speaks it out loud to another person)
- **Blind** - Imagine you were a person who is blind: would this voice be acceptable to use as an alternative reading tool, if it were the only option available? (you use the voice to read your news, documents and novels out loud)
- **Dyslexic** - Imagine you were a person who is dyslexic (you can see written words on a page, but you struggle to take in their meaning): would this voice be acceptable to use as an assistive reading tool, if it were the only option available? (you use the voice to read your news, documents and novels out loud)
- **Illiterate** - Imagine you were a person who can see, but is illiterate: would this voice be acceptable to use as an assistive reading tool, if it were the only option available? (you use the voice to read your news, documents and novels out loud)
- **Learning to read** - Imagine you were a person who can see, but is only starting out in learning to read (like a child in an early grade or an adult at an ABET centre): would this voice be acceptable to use as a learning tool towards reading, if it were the only option available? (you use the voice to read your lesson material, including prescribed books and homework assignments)
- **Preoccupied** - Imagine you were a person who can see and read, but whose visual sense is otherwise preoccupied (you want to read, but you are busy driving to work or exercising at the gym, etc.): would this voice be acceptable to use as an alternative reading tool, if it were the only option available? (you use the voice to read your news, documents and novels out loud)

The second evaluation is a comparative/relative benchmark test between the baseline and the improved TTS voices, in order to confirm that the pronunciation improvements do, in fact, result in better quality voices. We employ the same 20 unique phonetically-balanced test sentences for each language and gender TTS voice and synthesise two samples, one with the baseline and one with the improved voice. This time each test set is only assigned to 3 language practitioners due to cost containment measures and, again, a practitioner is reused over the two genders within a language. The practitioners must listen and select blindly which sample sounds better, or whether they sound the same, though now using a more efficient online web-based survey tool. They must also give reasons for their choice if it is one or the other, according to the same criteria used in the absolute benchmark test, though the questions are rephrased in a comparative/relative sense:

- **Word correct** - It pronounces individual words better
- **Stress/tone** - It has more accurate lexical stress and/or tone
- **Sent correct** - It conveys the meaning of the whole sentence better
- **Noise** - It has fewer noisy disturbances like scratches, pops, whistles, etc.
- **Tempo** - It speaks at a more acceptable, natural tempo (rhythm, rate)
- **Intonation** - It speaks with a more acceptable, natural pitch (intonation) from word to word and phrase to phrase
- **Pausing** - It places pauses in more appropriate, natural places
- **Breathing** - It places breathing sounds in more appropriate, natural places
- **Human** - It sounds more like a human
- **Other** - Other reason

B. Results

We illustrate the absolute benchmark test results for the baseline TTS voices with stacked bar charts in Figures 1, 2, 3 and 4. The counts of votes (horizontal axis) for each criterion (vertical axis) are grouped per Yes (blue), No (red) and N/A (missing) (orange) category. In the analysis that follows, we deem a voice "acceptable" to use when the

number of Yes votes are greater or equal to the number of No votes (#Yes $\geq$ #No).

For the sake of brevity in the exposition, we cluster the Word correct, Stress/tone and Sent correct criteria into "intelligibility" and the Noise, Tempo, Intonation, Pausing, Breathing and Human criteria into "naturalness" where appropriate. The same applies to the use cases. We cluster First lang, Second lang, News, Work/School and Leisure criteria into "mainstream" and the No speech, Blind, Dyslexic, Illiterate, Learning to read and Preoccupied criteria into "accessibility".

The Afrikaans female and male baseline voices in Figure 1 are judged acceptable according to all the intelligibility and naturalness criteria. They are also perceived acceptable in all the accessibility use cases and all the mainstream use cases, except for Work/School in the case of the male voice.

The English female and male baseline voices in Figure 1 are judged acceptable according to all the intelligibility and naturalness criteria, except for Human in both cases. However, these latter scores have a number of N/A (missing) values that could swing the votes otherwise. The voices are perceived acceptable in most accessibility and mainstream use cases. The female voice is not suitable for News, Work/School and Leisure. The male voice is not suitable for News, Work/School and Learning to read.

The Sepedi female and male baseline voices in Figure 2 are judged acceptable according to all the intelligibility and naturalness criteria, except for Pausing in the case of the male voice. Both voices are perceived acceptable in all the accessibility use cases, but only some mainstream use cases. They are not suitable for First lang, News and Leisure. Furthermore, the male voice is not suitable for Work/School.

The Sesotho female baseline voice in Figure 2 is judged acceptable according to all the intelligibility and naturalness criteria. It is perceived acceptable for all the accessibility use cases, but only one mainstream use case. It is not suitable for First lang, News, Work/School and Leisure.

The Setswana female baseline voice in Figure 2 is judged acceptable according to all the intelligibility and naturalness criteria. It is perceived acceptable for all the accessibility use cases, but only some mainstream use cases. It is not suitable for News, Work/School and Leisure.

The isiXhosa female and male baseline voices in Figure 3 are judged acceptable according to two out of the three intelligibility criteria. Neither has adequate Stress/tone. Both voices are judged acceptable according to few naturalness criteria. Neither has adequate Tempo, Pausing and Breathing. The female voice does not have adequate Intonation. Their Human scores are suspect due to the many N/A (missing) values that could swing the votes otherwise. Both voices are perceived acceptable for only some accessibility and mainstream use cases. Neither is suitable for First lang, News, Work/School, Leisure, Dyslexic and Learning to read. Furthermore, the female voice is not suitable for No speech, Blind, Illiterate and Preoccupied.

The isiZulu female and male baseline voices in Figure 3 are judged acceptable according to all the intelligibility and naturalness criteria. They are perceived acceptable in all the accessibility use cases and most mainstream use cases. The female voice is not suitable for First lang, News and Work/School. The male voice is not suitable for Leisure.

The isiNdebele male baseline voice in Figure 4 is judged acceptable according to most intelligibility and naturalness criteria. It does not have adequate Stress/tone nor Breathing. There are a number of N/A (missing) values that could swing the Breathing vote otherwise, but not the Human vote. The voice is perceived acceptable in all the accessibility use cases and all the mainstream use cases, except for First lang.

The siSwati female baseline voice in Figure 4 is judged acceptable according to two out of the three intelligibility criteria. It does not have adequate Stress/tone. The voice is judged acceptable according to only one naturalness criterion. It does not have adequate Tempo, Intonation, Pausing and Breathing, nor is it Human. The latter score has a lot of N/A (missing) values, but they cannot swing the vote otherwise. The voice is perceived acceptable in all the accessibility use cases, except Preoccupied. It is suitable for only one mainstream use case. It is not suitable for First lang, News, Work/School and Leisure.

The Tshivenda male baseline voice in Figure 4 is judged acceptable according to all the intelligibility and naturalness criteria. However, the Human score is suspect due to the many N/A (missing) values that could swing the vote otherwise. The voice is perceived acceptable in all the accessibility use cases, except Learning to read. It is suitable for only one mainstream use case. It is not suitable for First lang, News, Work/School and Leisure.

Finally, the Xitsonga female baseline voice in Figure 4 is judged acceptable according to all the intelligibility and naturalness criteria. It is perceived acceptable in all the accessibility use cases and most mainstream use cases. It is not suitable for First lang and News.

We illustrate the results of the comparative/relative benchmark test between the baseline and the improved TTS voices with stacked bar charts. In Figure 5, the counts of votes (horisontal axis) for each voice (vertical axis) are grouped per votes for the Improved (blue) version, votes for the Baseline (red) version and votes where the 2 versions are perceived Equal (orange). In the analysis that follows, we deem the pronunciation improvements in a voice as "successful" when the number of Improved votes are greater or equal to the number of Baseline votes (#Improved $\geq$ #Baseline).

Figure 6 gives a zoomed out and zoomed in view on the reasons behind preferences towards the Improved versions. Once again, we cluster the Word correct, Stress/tone and Sent correct criteria into "intelligibility" and the Noise,

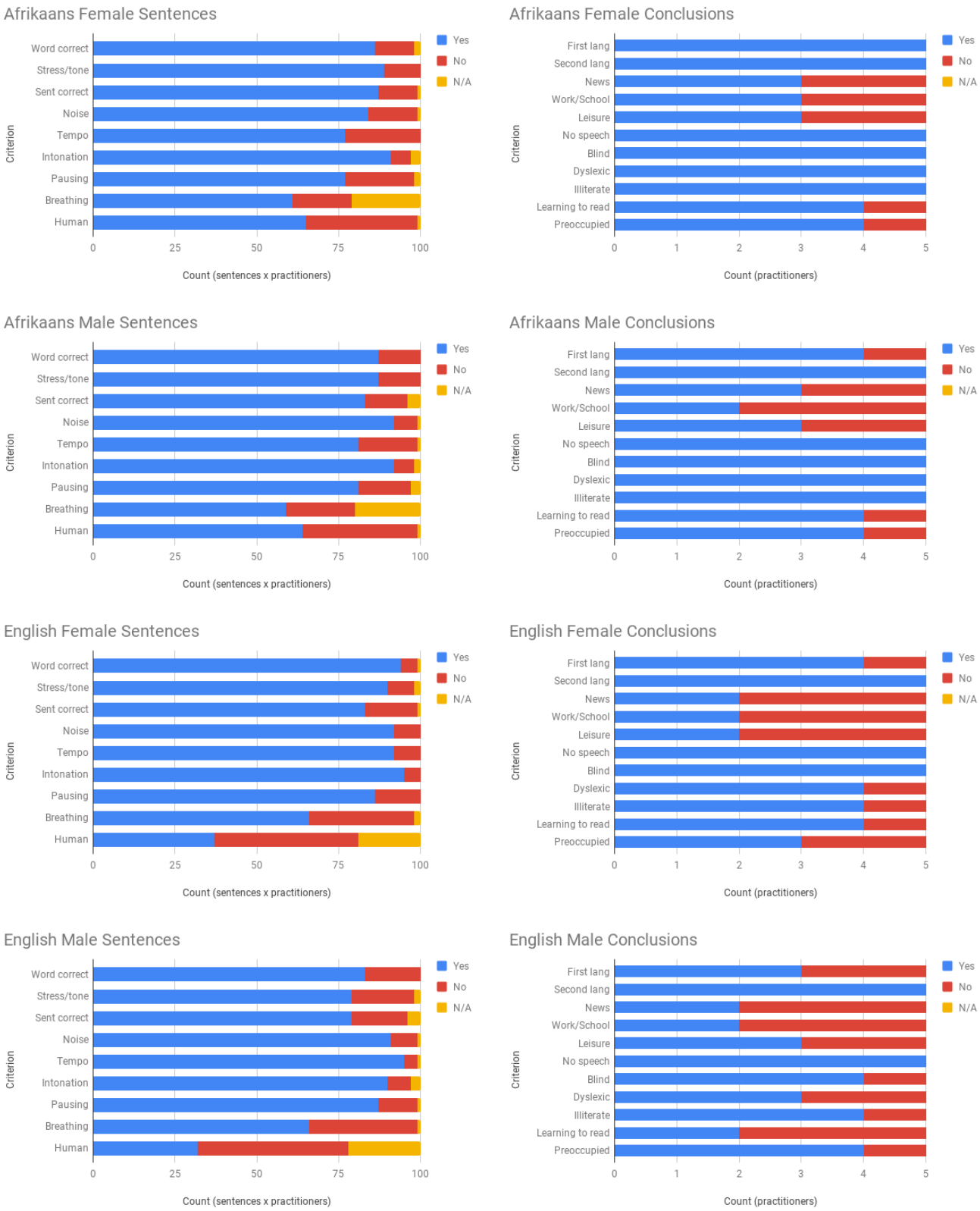

Fig. 1. Afrikaans and English baseline results

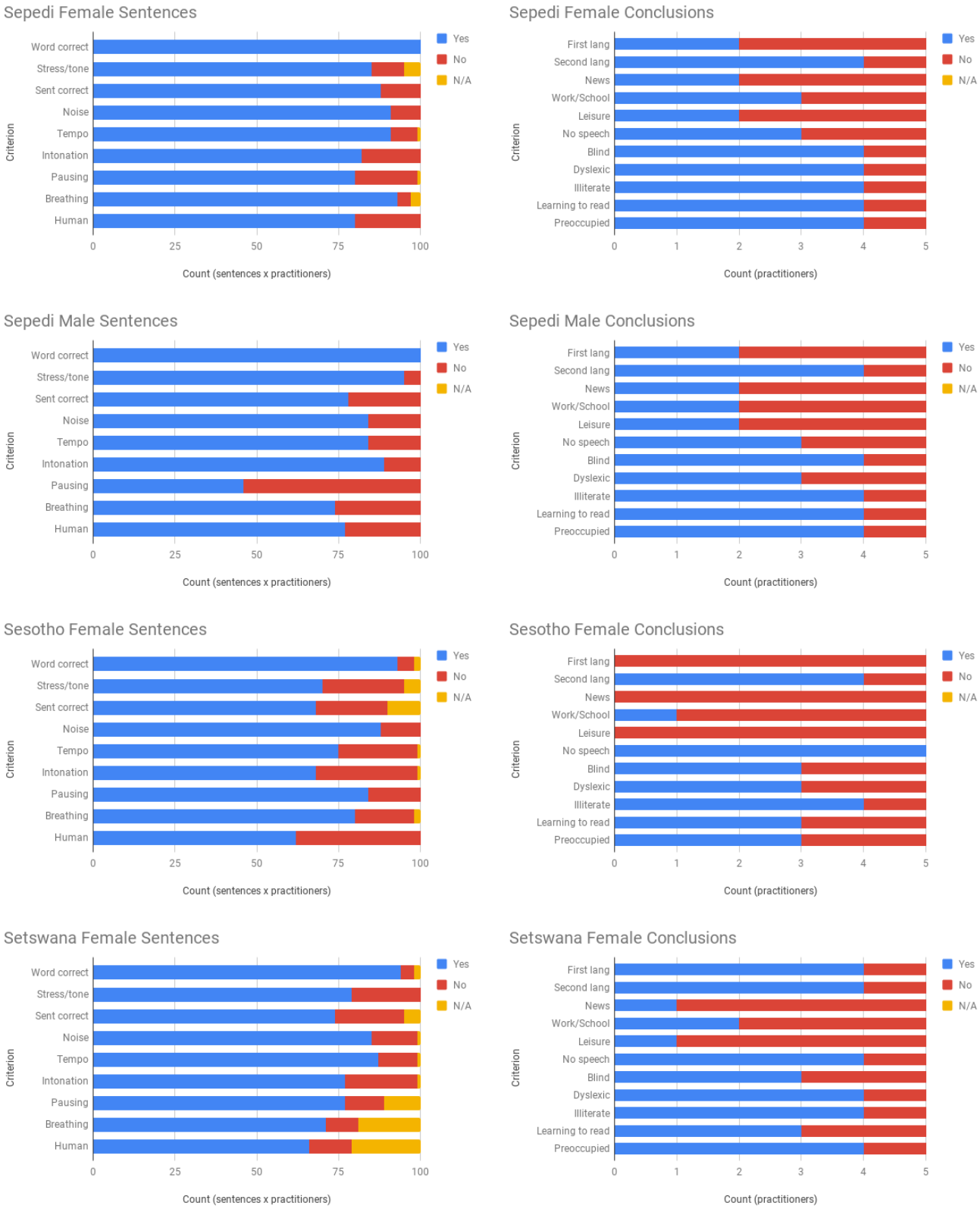

Fig. 2. Sepedi, Sesotho and Setswana baseline results

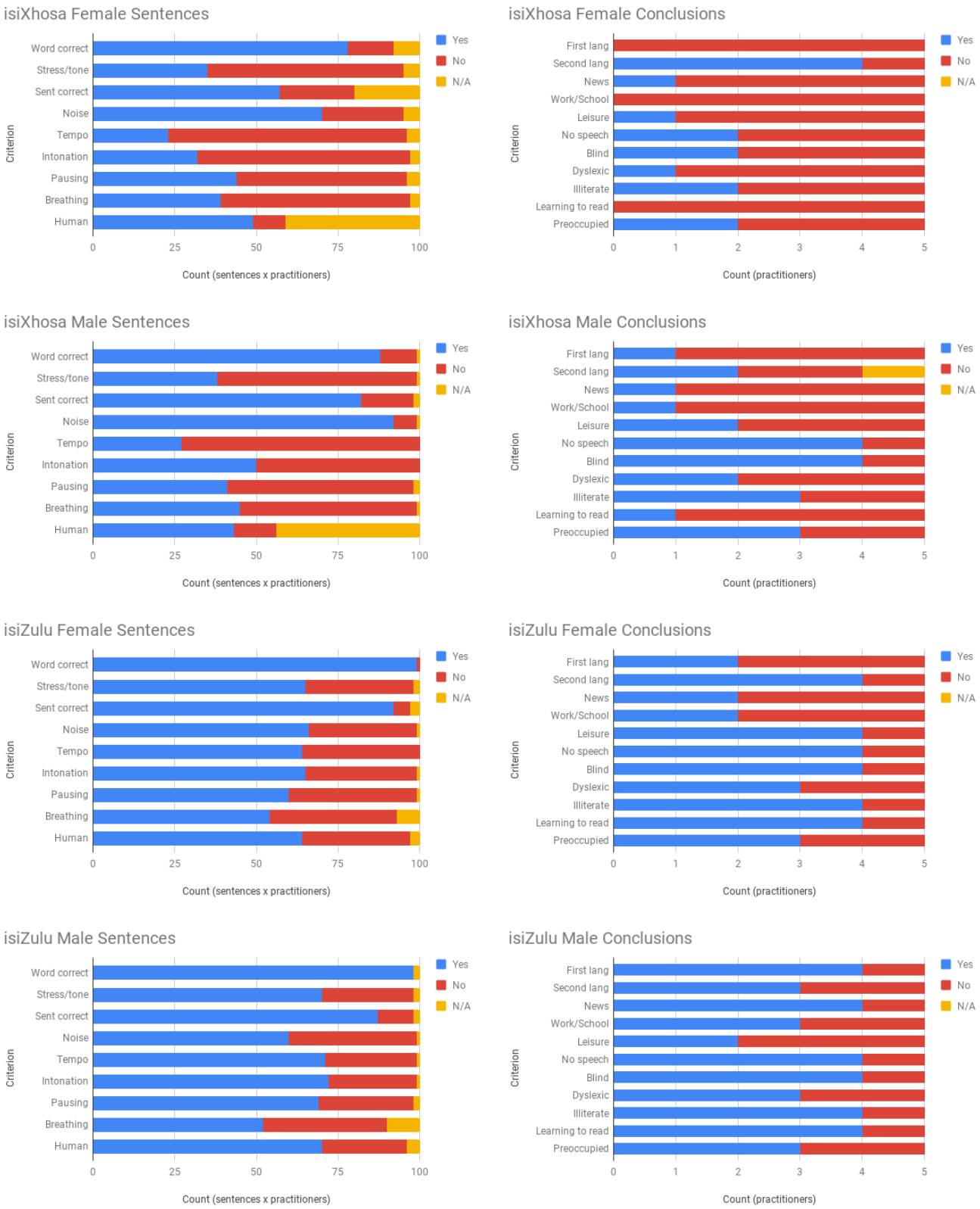

Fig. 3. isiXhosa and isiZulu baseline results

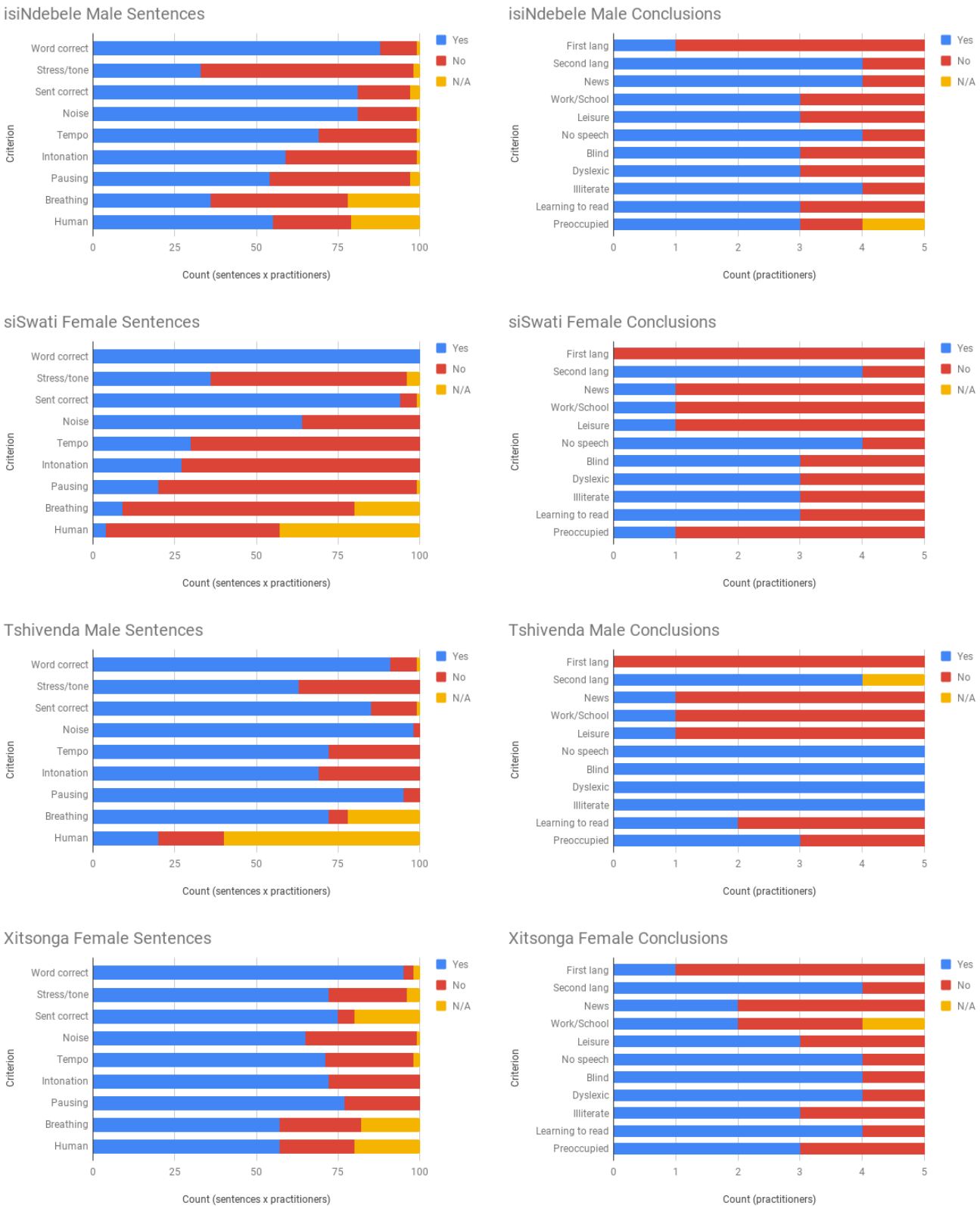

Fig. 4. isiNdebele, siSwati, Tshivenda and Xitsonga baseline results

Tempo, Intonation, Pausing, Breathing and Human criteria into "naturalness" where appropriate.

The pronunciation improvements in the Afrikaans female and male voices are judged successful. The intelligibility reasons given for the preferences towards the Improved versions are Word correct, Stress/tone and to a lesser extent Sent correct. The prominent naturalness reasons are Tempo, Intonation and to a lesser extent Pausing.

The improvements in the English female and male voices are judged successful, although borderline in the case of the male voice. Across both voices, the intelligibility reasons are mostly Word correct and the naturalness reasons Tempo and Pausing. In particular, the female voice also includes Stress/tone and Intonation.

The improvements in the Sepedi female and male voices are judged successful. The seemingly outlying large number of reasons could possibly be explained by the large winning margin of the Improved versions. All the intelligibility and naturalness criteria feature prominently among the reasons.

The improvements in the Sesotho female voice are judged successful, notwithstanding the many Equal votes. The prominent intelligibility reasons are Word correct and Sent correct. The naturalness reasons include all but breathing.

The improvements in the Setswana female voice are judged successful. The major intelligibility reasons are Word correct and Sent correct, while the naturalness counterparts are Noise, Tempo, Intonation and Human.

The pronunciation improvements in the isiXhosa female and male voices are judged successful. The intelligibility and naturalness reasons show similar behaviour to the Sepedi cases in number and criteria coverage.

The improvements in the isiZulu female and male voices are judged successful. The intelligibility reasons are Word correct, Stress/tone and to a lesser extent Sent correct. The prominent naturalness reasons are Tempo, Intonation, Pausing and to a lesser extent Breathing and Human.

The improvements in the isiNdebele male voice are judged successful, despite the many Equal votes. All intelligibility and naturalness reasons feature equally.

The improvements in the siSwati female voice are judged successful, notwithstanding the many Equal votes. Few reasons are given, compared to the isiNdebele case that has a similar winning margin. Only Word correct in intelligibility and Pausing in naturalness are prominent reasons.

The improvements in the Tshivenda male voice are judged successful. The major intelligibility reasons Word correct and Sent correct. The naturalness reasons all feature strongly, except for Intonation.

Finally, the improvements in the Xitsonga female voice are judged successful, although borderline. Relatively, the only prominent intelligibility reason is Sent correct and corresponding naturalness reason is Pausing.

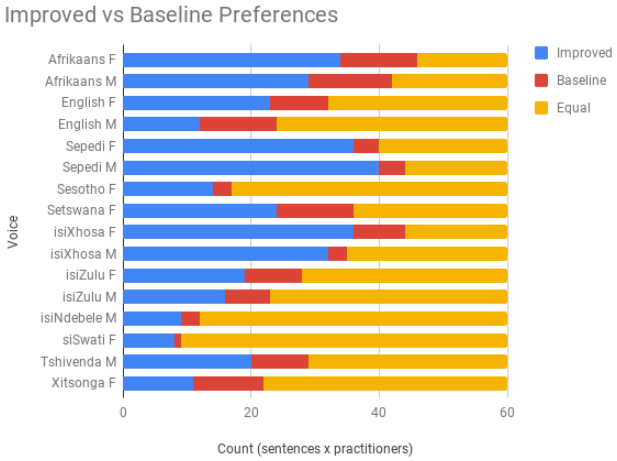

Fig. 5. Improved vs baseline preferences

IV. Conclusion

In the first absolute evaluation, most baseline voices are judged favourably for intelligibility and naturalness. The exceptions are the isiXhosa female and male voices and the siSwati female voice. Most baseline voices are deemed acceptable for use in the accessibility application settings. This confirms the historic market trend of early adoption of TTS in assistive technologies. The exceptions are the isiXhosa female and male voices.

In contrast, the language practitioners are much more conservative in their judgments over use in mainstream application settings. Only the Afrikaans female and male voices, the isiZulu male voice and the isiNdebele male voice have overly positive pronouncements. A particular point of interest is the comparison between the First lang and Second lang use cases. The Germanic language voices score well for both, but the African language voices are not deemed of sufficient quality to pass as first language speakers, only second language ones. This is confirmed by the qualitative feedback from some practitioners. We use the exact same modelling technique for all our voices, hence the reason must lie in the different linguistic structures among the language families. From a theoretical point of view, we are of the opinion that it is the lack of explicit tone modelling that hurts the quality of the tonal African languages. However, overly negative scores for Stress/tone are only reflected for the isiXhosa female and male voices, the isiNdebele male voice and the siSwati female voice.

In the second comparative/relative evaluation, the pronunciation improvements are successful for all the voices, although borderline for the cases of the English male voice and the Xitsonga female voice. The Word correct criterion features consistently as a prominent reason for the preferences towards the Improved versions, with the exception of the Xitsonga female voice. This confirms the theory that larger pronunciation lexica and corresponding

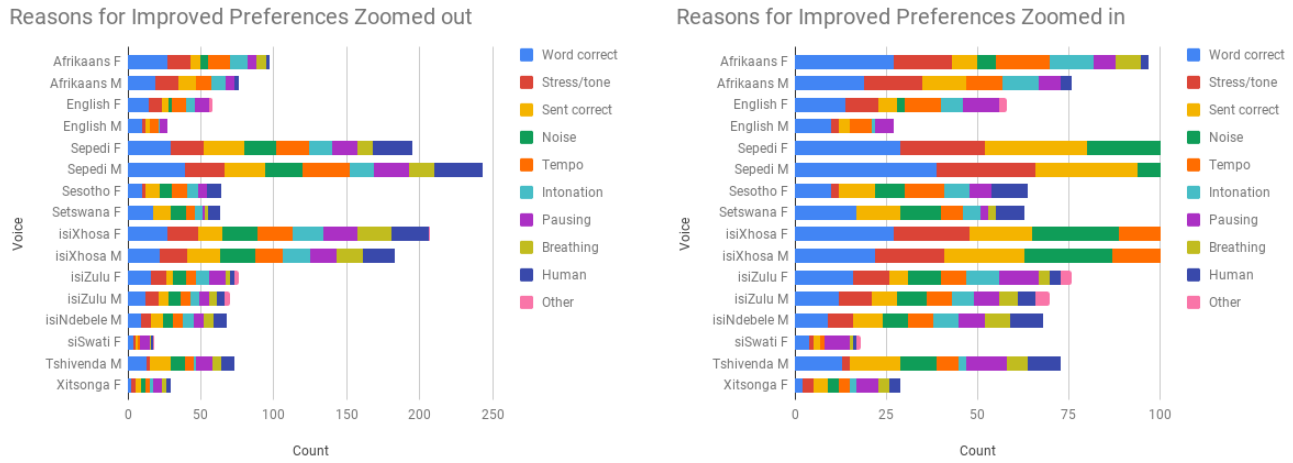

Fig. 6. Reasons for improved preferences

G2P rules should result in better quality pronunciations. Owing to the fact that the phonemes are fundamental units in the implicit modelling technique of our TTS system, the effects of the improved phonemic representations propagate indirectly to other criteria such as Stress/tone, Sent correct, Tempo and Intonation, though they do vary in prominence among the voices.

The pronunciation improvements are being released in our commercial Qfrency TTS offering. The next step on our research and development roadmap is word-level (lexical) stress and tone modelling, in an attempt to bring the quality of the African language voices up to first language standard. There are mainly two approaches to solving the problem: either explicit markup in the pronunciation lexica with an accompanying classifier in the text frontend, or implicit modelling using more advanced deep learning in the speech backend. A combination of the two is also possible. As mentioned in Section I, we also want to expand our sample base for evaluations to allow us to make stronger statistical inferences about our results.

References

[1] P. Taylor, *Text-to-Speech Synthesis*, 1st ed. Cambridge University Press, 2009.

[2] S. Millar and J. Scott, "What is augmentative and alternative communication? an introduction," *Augmentative communication in practice: An introduction*, pp. 3–12, 1998.

[3] L. Rello, H. Saggion, and R. Baeza-Yates, "Keyword highlighting improves comprehension for people with dyslexia," in *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, 2014, pp. 30–37.

[4] D. Lukeš, "Dyslexia friendly reader: Prototype, designs, and exploratory study," in *Information, Intelligence, Systems and Applications (IISA), 2015 6th International Conference on*. IEEE, 2015, pp. 1–6.

[5] University of Pretoria, "Progress in international reading literacy study (PIRLS) 2016," 2016. [Online]. Available: https://www.up.ac.za/media/shared/164/ZP_Files/pirls-literacy-2016_grade-4_15-dec-2017_low-quality.zp137684.pdf

[6] G. I. Schlünz, I. Wilken, C. Moors, T. Gumede, W. van der Walt, K. Calteaux, K. Tönsing, and K. van Niekerk, "Applications in accessibility of text-to-speech synthesis for South African languages: Initial system integration and user engagement," in *Proceedings of SAICSIT '17*, Thaba Nchu, South Africa, 2017.

[7] J. A. Louw, A. Moodley, and A. Govender, "The Speect text-to-speech entry for the Blizzard Challenge 2016," in *Proceedings of The Blizzard Challenge 2016 Workshop*, 2016.

[8] CSIR, "Qfrency TTS," 2018. [Online]. Available: http://qfrency.com/

[9] G. I. Schlünz, N. Dlamini, A. Tshoane, and S. Ramunyisi, "Text normalisation in text-to-speech synthesis for South African languages: Native number expansion," in *Proceedings of Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, Bloemfontein, South Africa, 2017.

[10] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proceedings of ISCA SSW6*, 2007, pp. 294–299.

[11] M. H. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proceedings of Interspeech*, Brighton, UK, 2009, pp. 2851–2854.

[12] M. H. Davel and E. Barnard, "Pronunciation prediction with Default&Refine," *Computer Speech & Language*, vol. 22, no. 4, pp. 374–393, Oct. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2008.01.001

[13] M. H. Davel and F. de Wet, "Verifying pronunciation dictionaries using conflict analysis," in *Proceedings of Interspeech*, Makuhari, Japan, 2010, pp. 1898–1901.

[14] D. R. van Niekerk, "Syllabification for Afrikaans speech synthesis," in *Proceedings of Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, Stellenbosch, South Africa, 2016.

[15] M. H. Davel, W. D. Basson, C. van Heerden, and E. Barnard, "NCHLT dictionaries: Project report," North-West University, Tech. Rep., 2013. [Online]. Available: https://sites.google.com/site/nchltspeechcorpus/home

[16] E. Barnard, M. H. Davel, C. J. van Heerden, F. de Wet, and J. A. C. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proceedings of the Workshop on Spoken Language Technologies for Under-resourced languages (SLTU)*, St. Petersburg, Russia, 2014, pp. 194–200.

[17] D. R. van Niekerk, C. J. van Heerden, M. H. Davel, N. Kleynhans, O. Kjartansson, M. Jansche, and L. Ha, "Rapid development of TTS corpora for four South African languages," in *Proceedings of Interspeech*, Stockholm, Sweden, 2017.