13th edition of the Speech Synthesis Workshop
24-27 August 2025, Leeuwarden, The Netherlands

The impact of stress and boundary information in the input to neural TTS

Christina Tånnander[1,2], *Joakim Gustafson*[1], *Jens Edlund*[1]

[1]Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden
[2]Swedish Agency for Accessible Media, Sweden

christina.tannander@mtm.se, jkgu@kth.se, edlund@speech.kth.se

Abstract

Previous research on input to neural text-to-speech (TTS) has focused on the choice between grapheme and phoneme input, and how additional information such as stress or morphological boundaries may affect output. While grapheme input performs well for general texts in some languages, phoneme input is often more robust, especially for complex or irregular material.

We investigate the effect of adding stress, accent, compound and syllable boundaries to grapheme and phoneme input for Swedish neural TTS. The systems are evaluated using a best-worst preference test across three sentence types: high-frequency, low-frequency, and deliberately ambiguous.

Results confirm the expected hierarchy: human recordings are most preferred, followed by phoneme-based systems, and then grapheme-based systems. The benefit of detailed input increases with sentence complexity, with the largest effects seen for low-frequency material. These results highlight the role of explicit linguistic structure, especially under realistic, pronunciation-challenging conditions.

Index Terms: text-to-speech synthesis, training input, transcription levels, Swedish

1. Introduction

Many modern text-to-speech (TTS) systems perform well on general texts composed of high-frequency words in the target language, even when trained solely on grapheme input. However, factors such as the orthographic depth of the language and the domain in which the TTS is to be used play an important part in how well a TTS system performs.

The effectiveness of grapheme versus phoneme input in TTS systems is closely tied to a language's orthographic depth, that is the regularity of the mapping between graphemes and phonemes. In orthographically shallow languages (e.g. Bulgarian and Spanish), with a largely consistent mapping, grapheme input performs well [1]. In contrast, languages with deep orthographies (e.g. French and English) with less predictable spelling-to-sound correspondences pose greater challenges for grapheme-based models [2]. For writing systems based on more complex characters, such as Chinese sinograms or Japanese kanji, grapheme input becomes even less practical [3]. Swedish, the target language of this study, occupies a middle position on the orthographic depth scale among European languages [4].

The target domain for our TTS systems consists of lengthy, information-rich texts, such as textbooks and Wikipedia pages, containing a higher proportion of low-frequency words such as domain-specific terms, foreign proper names, and other items outside the standard pronunciation patterns of the language. Such words are particularly prone to mispronunciation, even by systems that are otherwise well trained on the language.

The choice of test material plays a critical role in evaluating TTS systems. This was demonstrated by [5], who examined how different input representations handled liaison and enchaînement in French. When tested on a small set of randomly selected sentences, the models showed little difference. However, when the evaluation used targeted sentences designed to expose pronunciation challenges, clear differences emerged. This illustrates that if a system is intended to handle linguistically or phonetically complex material, it must also be evaluated using appropriately complex test data.

In this study, we examine how the level of stress and boundary information in the input to a neural TTS system affects the output. The goal here is not to chase output quality scores, but to disentangle the effects of specific input features and examine how interpretable differences in linguistic input shape system behaviour under controlled and task-relevant conditions. Our experiments are conducted in Swedish, but we believe the findings are applicable to other languages with similar phonological characteristics, and that the overall methodology can generalise more broadly. Using a fixed set of grapheme and phoneme symbols, we vary the presence or absence of stress, accent, compound, and syllable boundaries in the input, and evaluate their impact on listener preferences in a best-worst scaling task.

Our results confirm the expected hierarchy: human recordings are preferred over all synthetic systems, phoneme-based systems are consistently preferred over grapheme-based ones, and systems with more detailed input representations tend to be preferred over those with less detail. These effects are considerably stronger when the test material includes a high proportion of low-frequency words.

2. Background

2.1. Swedish stress patterns

Swedish is a pitch accent language: the primary stress of each word is either accent 1 (acute accent) or accent 2 (grave accent) [6, 7, 8]. Words with primary stress on the last syllable, including monosyllabic words, always get accent 1, while polysyllabic words can be either accent 1 or 2. The vast majority of Swedish compounds belong to the accent 2 class, where the first compound part achieves primary stress and the last secondary stress. For non-compounds, the rule is that the syllable following a primary stressed accent 2 syllable achieves secondary stress. Because of this regularity, the secondary stress of non-compounds is not always explicitly assigned in phoneme transcription systems for Swedish. Secondary stress in accent 1 words is usually not assigned. In Stockholm Swedish, accent 1 is characterised by an LHL contour, and accent 2 by an HLHL contour.

10.21437/SSW.2025-10

2.2. Swedish compounds

Compounds are very common in Swedish and can, theoretically, consist of an infinite number of compound parts. The first and middle parts can contain a linking morpheme, for example -s in *pappers-kopia* (from *papper* and *kopia*, en. *paper copy*), the final vowel can be altered, as in *fog**e**-morfem* (from *foga* and *morfem*, en. linking morpheme) the final vowel can be omitted, as in *flug-papper* (from *fluga* and *paper*, en. *flypaper*), or it can be a specific compound form, for example *musei-* (from *museum*, en. *museum*). In Central Swedish, the vast majority of compounds get accent 2 on the first compound part and secondary stress on the last, although there are exceptions such as *blåbär* (en. *blueberry*) [8].

2.3. Boundaries

Sentence boundaries are natural parts of the input to TTS; most systems are trained on separate sentences and also take one sentence as input for inference. In most alphabetic writing systems, the sentence boundaries are marked with major delimiters, such as full stops or exclamation marks in the grapheme representation of a text. Similarly, major *phrase boundaries* are related to commas, semicolons and other minor delimiters, and *word boundaries* are usually signalled by space. Boundaries at a lower level, such as compound, morpheme or syllable boundaries, are normally not represented in raw text, although some compound boundaries in Swedish and English are explicitly marked with a hyphen (e.g. *long-term*). *Morpheme* and *syllable boundaries* are never explicitly represented in raw Swedish text, while it is possible to include whichever boundaries you'd like, such as compound, morpheme or syllable boundaries in phoneme representations of text.

2.4. Input to neural TTS

Both graphemes (letters) and phonemes can be used as input to neural TTS. End-to-end (E2E) TTS typically takes raw text as input and generates speech without intermediate steps, such as grapheme-to-phoneme (G2P) conversion or prosody control. While E2E TTS is interesting and challenging from a machine learning point of view, where it aims to minimise, or even eliminate manual intervention, the practical use of TTS is put at risk if the input cannot be manipulated to represent correct pronunciations [9, 10, 11]. Using phoneme input, a dictionary and/or a G2P converter are required to provide the system with pronunciation guidelines. These are components that themselves can introduce errors [12], but that makes it possible to control the output of the TTS to a greater extent than using grapheme input. While a G2P makes explicit phone predictions given the graphemes, an E2E TTS with grapheme input makes implicit phone predictions given graphemes.

There are several reasons for including boundaries at a more detailed level. Firstly, and language-specifially, a compound boundary in Swedish signals that the word should have accent 2 and secondary stress. Secondly, compound and morpheme boundaries can guide a G2P towards the correct pronunciation, whether it is standalone or implicitly integrated into a neural TTS system. An English example of this is *loophole*, where a boundary between p and h forces the G2P to pronounce them as separate phonemes instead of /f/ [11]. Yet an example, this time from Swedish, is the separation of word parts in *missträsta* (en. *despair*), which is correctly segmented as *miss-trösta*, preventing the /t/ from being deaspirated as it would if preceded by /s/ in the same syllable.

2.5. Symbol sets for Swedish commercial TTS

Symbol sets for Swedish commercial TTS typically contain representations of phones or phonemes, stress and sometimes syllable boundaries [13, 14, 15, 16]. The systems typically use IPA [17], SAMPA [18], or X-SAMPA [19] as input symbols. Not all systems clearly state in their main documentation whether they use different symbols for accent 1 and accent 2, though this may be specified in more detailed resources. Some symbol sets use secondary stress in all accent 2 words, but it is also common to omit the secondary stress in non-compounds, motivated by the rule explained in section 2.1: the syllable following the primary stressed syllable receives secondary stress.

3. Related work

A number of papers have explored the input representations used in neural TTS systems, with a primary focus on grapheme or phoneme inputs. However, some approaches go beyond this by incorporating mixed input symbol types during both training and inference. These include combinations such as graphemes and phonemes, or graphemes enriched with for example morphological or syntactic boundaries, sometimes refereed to as representation mixing. One of the most explored field of representation-mixing is using both graphemes and phonemes in the input data (see e.g. [9, 10, 11]).

3.1. Grapheme or phoneme input

In Tacotron 2, G2P is implicitly learned if using grapheme input [12, 2, 3]. [20] achieved better results when they compared the G2P results trained on typical E2E TTS data instead of an English pronunciation lexicon, suggesting that the balance of the training data is important when aiming for a full E2E system with implicit letter-to-sound conversion. In the same way, the size of the training data is crucial, since a larger training data also is likely to contain a greater variation in character sequences [21].

There are benefits with both methods: grapheme input requires less preprocessing of the input text than phoneme input, but a phoneme representation of a text is naturally closer to its intended acoustic form and has often been shown to perform better than grapheme models (see e.g. [12, 5]), even though some researchers have achieved similar results with grapheme input, for example [3] for French. [12] showed that using grapheme input gave a relative decrease in naturalness scores of 23,5%, compared to using correct phonological transcriptions of all words. Furthermore, they got similar results for their phone model with correct transcriptions for all words and a phone model with 15% incorrect transcriptions, indicating that it is acceptable with an out-of-vocabulary (OOV) rate of up to 15%. [11] mixed graphemes and phonemes in English, and found that it was enough using 500 phonemicised word types in their training data (LJ Speech) to be able to control pronunciation when synthesising, given that these word types were selected wisely (in this case by choosing words with rich grapheme context). Adding syllable boundary information further increased the performance of the model, resulting in correct pronunciations of compounds such as *loophole*, that was pronounced with a word medial /f/ if left to a grapheme representation. They had no success adding stress information in their model, but note that the evaluation concerned judging words in identical carrier sentences, where there was no context triggering different stress patterns of homographs such as record.

Some researchers have taken the level of detail one step

further, using phonetic or phonological feature input to neural TTS. This approach has primarily been used for under-resourced languages or for multilingual or accented voices (e.g., [22, 23, 24, 25, 26]).

3.2. Augmented input

Another representation-mixing method is to add certain features to the grapheme or phoneme representation of the text to be inferred. [2] added morpheme boundaries to their grapheme representation of their English input data, and found that this model outperformed their model with phoneme input, and got similar results as their phoneme model augmented with morpheme boundaries. However, certain word types were still handled better by the phoneme model, for example foreign words and proper names. [21] reached good results for enhancing prosody of English TTS using features derived from phrase structure, for example phrase labels such as NP (noun phrase) and VP (verb phrase), the relative position of the word in the phrase and syntactic distance metrics.

Yet another way of enhancing prosody was explored by [27], who used syllable duration as a proxy for prominence in Swedish. The speaking rate categories 1 to 5 were assigned each word in the training data, depending on their relative syllable duration (1 for the shortest durations, 5 for the longest), and could then be used to create slower and hyper-articulated speech, or to assign prominence to specific words.

As mentioned above, some languages have features that can make them less suited for E2E TTS than others, for example non-alphabetic writing systems like Mandarin. Another example of such features is the liaison process, where linking sounds are inserted between words according to complex phonetic, grammatical and stylistic rules and enchâinement, where the last phone of a word is moved to the first position in the next word, which can result in a changed syllable structure [5]. [5] compared phone and grapheme input data for French, as well as phone input compared to phones with syllable boundaries. They found no significant differences in preference between any of the compared sets when using a small amount of random test sentences. However, when they used targeted test sentences, 10 with disallowed liaison, and 10 sentences with enchaînment changing the syllable structure, phone representations were significantly better than graphemes, and there was also a preference for phone input enriched with syllable information over phones only.

4. Method

4.1. Symbol sets

In the current experiment, two different symbol sets were used: one used for all experimental models with grapheme input (**SYMGRAPH**), and another for all models with phoneme input (**SYMPHONE**). To isolate the effect of input structure, all other aspects of training and architecture were kept constant, allowing us to attribute output variation directly to the type and level of linguistic information provided.

Both sets contain 121 symbols representing graphemes/phonemes, stress and/or boundaries, as well as some extra symbols to use for further extensions of the symbol sets (see table 1). This means that we are guaranteed that all models will have exactly the same dimensions, but also that there will be many symbols that have no representations in the training data, so-called empty symbols. We have trained several Tacotron models with empty symbols, and the only noticeable effect has been that the models become slightly larger.

SYMGRAPH contains 10 vowels and 20 consonants, both sets in lower- and uppercase. Only symbols that occur in the training data more than 100 times were included, which means that some graphemes were merged with their nearest general grapheme (e.g. è is rewritten to e). The training data was already preprocessed, and contained no abbreviations, digits or delimiters other than commas, full stops, some hyphens and colons. To match the symbol set of **SYMPHONE**, commas were replaced with the phrase boundary symbol '/', spaces with the word boundary symbol '&' and spaces were inserted between all characters.

SYMPHONE includes all phonemes and allophones that are commonly included in Swedish phone sets for TTS [13, 15, 28, 16, 18], although it is not able to handle xenophones (foreign speech sounds) that can be common in some text types, for example /ð/. This results in 24 vowels and 24 consonants.

4.2. Speech data

The Swedish speech data was originally recorded by the Swedish Agency for Accessible Media (MTM) when building a unit selection voice in 2011 [29]. The speech database consists of around 32,5 hours (30 746 utterances), of which 24,5 hours (22 142 utterances) Swedish and nearly 8 hours (8 604 utterances) English. 19 hours of the Swedish data (17 034 utterances distributed on 12 420 prompts) was used for training, 134 utterances for validation (100 prompts) and 4 530 utterances have been held out for test purposes. The sentences in the speech data were selected according to their phonetic richness, which means that they also contain examples of phoneme sequences that are uncommon in Swedish. We can therefore suspect that this data is phonetically better balanced, and accordingly better suited for E2E TTS than for example data collected from some books, such as the widely used LJ Speech dataset [30]. The speech database is currently not open source, but the owners are looking into the possibility for a public release.

4.3. Training the models

We used Nvidia's PyTorch implementation of Tacotron 2 [31], a known, stable architecture with well-understood behaviour. This allowed us to hold all aspects of training constant and isolate the effect of input structure. We synthesised using the WaveGlow vocoder [32], trained for 650 epochs on the same speaker as the Tacotron models.

4.3.1. Basic models

First, two base models were trained: (**GRAPH-500** and **PHONE-500**) Each base model was trained for 500 epochs (iterations of the entire training data). The only symbol categories that were changed in the additional training sessions were stress and boundaries. The grapheme and phoneme sets were identical throughout the training procedure. The basic models used three boundary levels: sentence, phrase (corresponding to pauses in the recordings) and word, and included no stress symbols. This is motivated by the information levels of a normal Swedish text: major and minor delimiters represent sentences and phrases/pauses, spaces represent word boundaries, and there is no information about stress.

Table 1: *Boundary and stress information in the grapheme (G) and phoneme (P) models. All models include sentence, phrase and word boundaries.*

	G BASE	G COM	G SYLL	P BASE	P STR	P ACC	P ALL
compound	0	1	0	0	0	0	1
syllable	0	0	1	0	0	0	1
primary stress	0	0	0	0	1	1	1
secondary stress	0	0	0	0	1	1	1
accent 1	0	0	0	0	0	1	1
accent 2	0	0	0	0	0	1	1

4.3.2. Additional models

All additional models were based on **GRAPH-500** and **PHONE-500**, and each of them were trained for yet another 500 epochs, resulting in a total of 1 000 epochs. This limit was chosen after comparing models with 1 000 and 1 500 epochs without noticing any major quality difference. The additional models added different combinations of stress, accent and boundary information to the basic models, as illustrated in table 1. **G-BASE** is the same as **GRAPH-500** but was trained for another 500 epochs. In **G-COMP**, compound boundaries were inserted, and in **G-SYLL**, all words were segmented into syllables. **P-BASE** is the same as **PHONE-500**, trained for 500 more epochs. **P-STR** had stress markers (but no information about accent); **P-ACC** had stress and accent; **P-ALL** had all available information: stress, accents, compound and syllable boundaries.

Finally, **HUM**, the human voice, was sentences from the test set of the recordings of the human professional voice talent.

The choice to insert compound and syllable boundaries in the grapheme input is motivated by the fact that this information often is accessible in common dictionaries. Albeit they might not show the phonological representation of the word, they sometimes have information about stress and accent, as well as compound boundaries. We can then assume that it is possible to access information about compound parts and stress patterns in the orthography, also without a pronunciation dictionary. The phonological transcriptions in the phoneme models were obtained from a Swedish pronunciation dictionary for TTS, which also contains information about stress and accent, as well as compound and syllable boundaries. In **G-SYLL**, the syllable boundaries were inserted at every compound boundary, and before each valid syllable onset. The valid orthographic onsets were obtained by creating a list of all word initial onsets in the training data.

4.4. Test sentences

The test data consisted of sentences from the test set of the original TTS training data. Frequency lists from 8 sidor (a newspaper in easy Swedish) and LäsBarT (easy Swedish and children's books) from Språkbanken Text [33], together including more than 5 million tokens, were used to decide whether a sentence was "high-frequency" or "low-frequency" in the selection of our three different test sets:

Ambiguous sentences (AMB): 8 sentences with homographs (words with identical written form but different pronunciations, for example different stress locations (as in the English noun and verb *record*) or Swedish accent 1 or 2.

High-frequency sentences (HIGH): sentences consisting of words that occurred at least 100 times in the easy Swedish corpora. 200 sentences with the highest average word frequencies were manually approved to not include any proper names, acronyms or foreign words, and to have a reasonable syntax. 8 sentences of different length were selected from this collection.

Low-frequency sentences (LOW): 200 sentences with a low average of word frequency were selected, without any restrictions regarding word types such as proper names or foreign words. Again, 8 sentences of different lengths were selected.

Tacotron 2 produces different versions of input sequences every time they are synthesised (if not seeded differently). to decrease the risk of getting random results caused by luck or bad luck when synthesising the input just once, we included 4 different renditions of the same sentence and model in the audio test data. 8 renditions of each model and sentence were synthesised, leaving us with 4 backup sentences to use if a rendition failed due to severe attention errors. To avoid unintentionally selecting the "best" renditions, we did not listen to the synthesised sentences before stimuli selection. Instead, we visually inspected the alignments and replaced renditions due to bad alignment or unexceptionally long audio files.

In addition to the seven models trained for our purposes, the human recording of each sentence (which was not part of the training set) was included in the test sets. The human voice was only available in one version.

4.5. Best-worst and balanced incomplete block design

Comparing all pairs of 8 systems across 24 utterances results in 672 unique system-pair–utterance combinations. With a target of 5 judgements per pair, traditional pairwise testing would require 6720 audio presentations and 3360 listener decisions. We used best-worst scaling (BWS) [34] where participants selected the best and worst of four versions of the same sentence in each trial. Four items is a good compromise between information density and perceptual load: smaller sets give less data, while larger ones increase the risk of listener confusion and subsequenct errors. Four item BWS achieves the aforementioned 3360 pairwise judgments from just 2688 presentations and 1344 decisions, or 40% of the effort, with no loss in coverage.

We constructed a balanced incomplete block design (BIBD) with $v = 8$ systems and $\lambda = 6$ co-occurrences per pair. The 4-item BWS format fixes the BIBD block size to $k = 4$, yielding $b = 28$ blocks and $r = 14$ appearances per system. This structure guarantees that each pair of systems is compared the same number of times within each utterance. One BIBD was generated per utterance, ensuring balanced coverage across all 24 test sentences.

To reduce repetition, we transposed the BIBD: each of its 28 blocks was assigned to a different participant, and the same design structure was reused across all utterances with rotated assignments. This let each participant hear each sentence once, with a different set of systems on each trial, while ensuring that each sentence received a full, balanced set of comparisons.

The BIBD was generated using the `ibd` package in R and converted into per-participant runsheets with a custom Perl script. This script assigned system combinations and synthesis variants to each trial, randomised the order of systems within each trial, and shuffled the sentence order for each participant after system assignment. Each participant heard each system about 12 times, and all system pairs co-occurred equally across the full experiment.

4.6. Procedure

28 native Swedish participants were recruited via Prolific and directed to the test platform at cognition.run. Participants were

instructed to listen to all four systems in each trial and select the version they liked most and the one they liked least ("Listen to all four audio files and select the reading you like most, and the one you like least," translated from Swedish). An audio control question was presented first, requiring identification of an animal sound. Participants could not continue without listening to all versions and making both selections. The test took approximately 12 minutes, and participants were compensated £2.

4.7. Analysis

4.7.1. Pairwise expansion of Best-Worst comparisons

To enable pairwise comparison modelling, each best-worst judgement was expanded into a set of directed win–loss outcomes. Specifically, each four-item trial yields five pairwise preferences: the system selected as "best" is considered to beat the three others, and each of the two non-selected systems is considered to beat the "worst". These five pairwise outcomes per trial were mirrored during processing, so that each comparison appears both as a win for one system and a loss for the other. This transformation retains the comparative information from the original task while producing a dataset that is compatible with models such as Bradley–Terry. Trials where two systems were present but neither selected as best or worst do not produce an outcome and are treated as unobserved.

4.7.2. Bradley–Terry modelling

The resulting pairwise win–loss data was modelled using the Bradley–Terry model [35], a widely used approach for estimating latent system-level *ability scores* from pairwise preference data.

Each TTS system, and the human reference, is assigned a real-valued ability score θ_i such that the probability of system i being preferred over system j is given by:

$$P(i \succ j) = \frac{e^{\theta_i}}{e^{\theta_i} + e^{\theta_j}}.$$

A higher θ value indicates a system that is more likely to win pairwise comparisons. One system must be fixed as a reference to identify the model; we used **G-BASE**, the simplest grapheme-based system, and set its ability score to zero.

Bradley–Terry models were fitted to the full dataset and to each of the three utterance-type subsets: low-frequency (**LOW**), high-frequency (**HIGH**), and ambiguous (**AMB**). This allowed us to examine overall performance rankings as well as variation in comparative system strength by sentence type.

Confidence intervals and all estimated ability scores were obtained directly from the model via logistic regression, and pairwise win probabilities between systems were derived from the same model fit.

5. Results

The 28 respondents all had Swedish as their first language. 11 were female and 17 male, with ages ranging from 23 to 68 years.

The left pane of Figure 1 shows the pairwise win probabilities between all systems, derived from the Bradley–Terry model fit to the full dataset. Each cell represents the estimated probability that the system on the row is preferred over the system on the column. The resulting matrix reflects a broadly consistent ranking of systems by transcription richness. Within the grapheme-based systems, differences are small, as they are within the phoneme-based systems.

Filtering on the ambiguous sentences produces a similar picture, but with somewhat greater variance (not shown here for space reasons), likely reflecting a combination of reduced data volume and the inherently variable nature of the ambiguities and their resolution.

The right pane of Figure 1 shows the same structure based on the low-frequency sentences alone. As expected, this condition brings out stronger distinctions between systems. The human reference maintains a clear lead, while the gaps between synthetic systems widen in line with transcriptional detail.

We conclude the results section by presenting the estimated Bradley–Terry ability scores for the low-frequency sentences in Figure 2, with confidence intervals confirming that the differences are statistically meaningful.

6. Discussion

Our method and analysis serve a broader aim: understanding how symbolic input structure affects system behaviour. Accordingly, we make no claims about model architecture or training volume, and we purposefully avoid common preprocessing techniques such as byte-pair encoding (BPE), which risk obscuring structural effects through fragmentation and normalisation.

The Bradley–Terry model fitted to the full dataset (Fig. 1, left pane) reveals a consistent and interpretable structure. The human reference system clearly outperforms all synthetic systems, as expected. It is a professional voice talent reading held-out test sentences from the same corpus used to train the TTS systems, and thus represents a realistic and meaningful topline. Among the synthetic systems, a broad separation emerges between the phoneme-based and grapheme-based models, with all phoneme systems scoring higher than any grapheme system.

Among both the grapheme-based and phoneme-based systems, the models are tightly clustered. Both groups show improvements with increased information, but these are dwarfed by the shift from grapheme to phoneme, and from phoneme to the human reference.

We now move to the low-frequency subset. This can be seen as targeted test material in the sense that it is especially likely to expose system weaknesses in pronunciation modelling. At the same time, it reflects a task-sensitive approach to evaluation: low-frequency vocabulary is characteristic of the long, information-rich texts our systems are intended to handle.

As shown in the heatmap for the low-frequency subset (Fig. 1, right pane), each system improves consistently over the previous one, with the effect of added boundary and stress information now clearly visible. These differences are confirmed in the ability score plot (Fig. 2): although adjacent systems generally show overlapping 95% confidence intervals, the hypothesised continuous rise of ability score with information is present.

The clearest outcome in the low-frequency condition is the dramatic underperformance of the grapheme-only system, **G-BASE**. It is beaten in 57% of comparisons by the next weakest system (**G-COMP**), and in more than 90% of comparisons by all phoneme-based models. The human system beats it 98% of the time. These margins are not subtle, and exceed many improvements reported between modern TTS systems under general conditions. This reinforces the idea that low-frequency material exposes failures in pronunciation modelling that otherwise remains hidden.

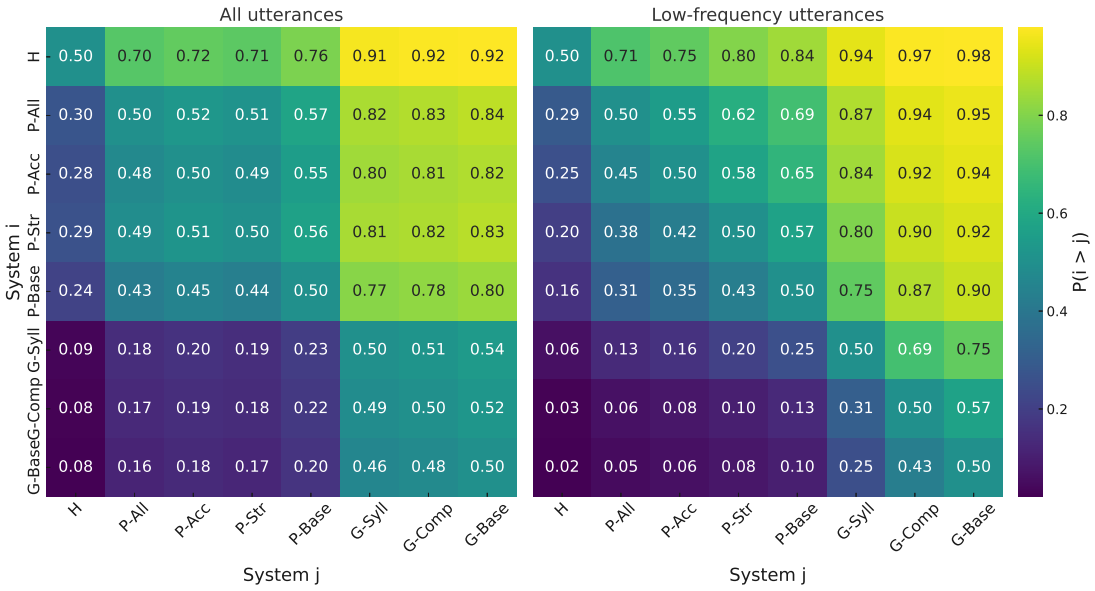

Figure 1: *Estimated pairwise win probabilities for all systems, based on the full dataset (left) and the low-frequency utterances subset (right). Each cell shows the probability that the row-system is preferred over the column-system.*

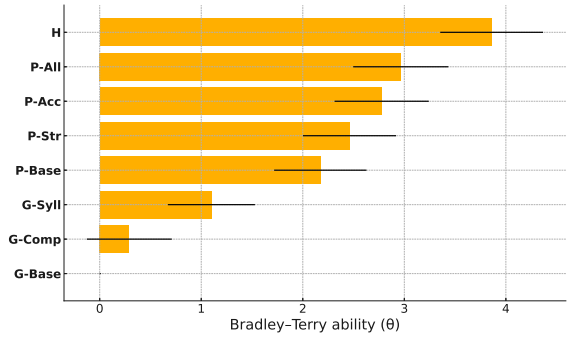

Figure 2: *Bradley–Terry ability scores with 95% confidence intervals, estimated from low-frequency sentences only.*

The human system, **H**, remains clearly ahead in the low-frequency condition, but it is nearly rivalled by the best phoneme-based system when it comes to beating the weakest model. The **G-BASE** system performs so poorly on this material that it is readily beaten by every other system. This highlights a critical nuance behind the oft-quoted "indistinguishable from human performance" of modern TTS: our best system is almost humanlike in its ability to outperform the worst one; a textbook ceiling effect, not a sign of true parity. On the low-frequency data, the best system (**P-ALL**) and **H** even have slightly overlapping 95% confidence intervals overall, meaning they are broadly comparable in how effectively they beat the rest of the systems. However, in direct pairwise comparisons, **H** still beats our **P-ALL** 70% of the time across all sentences, and 71% of the time on the low-frequency subset. To suggest that the system is "indistinguishable from a human" on this basis would be an overreach, if not outright misleading.

An opposite floor effect appears when comparing the human system to the grapheme models. Even the best of these (**G-SYLL**) wins just 6% of its pairwise comparisons against **G-BASE**, with the rest performing much worse. In this range, the gap is so wide that the comparisons offer little insight. More meaningful comparisons occur within the phoneme-based systems, which win 16%, 20%, 25%, and 29% of the time against the human voice **H**, ordered by amount of boundary information provided during training.

7. Conclusion

Our results confirm that adding boundary information to input representations affects listener preference in predictable and interpretable ways. Importantly, these effects only become fully visible under targeted testing, where the material is deliberately selected to include lexical items likely to trigger pronunciation errors. In our case, the targeted testing is not artificial; it reflects the types of lexical challenges that are common when dealing with long and information-rich texts.

We end with two take-home messages: Firstly, if you have a human reader, use them. Otherwise, if you have access to accurate phoneme input, use that. Note, however, that truly accurate phonemes are often difficult to obtain for low-frequency or novel words, and maintaining a high-quality lexicon may require continuous human support. Our results suggest that if you are choosing between adding boundary information to graphemes or using accurate phonemes, phonemes provide greater control over pronunciation—and are likely to have a stronger effect. But if phonemes are out of reach, boundaries will still help.

Secondly, when your goal is to synthesise long materials such as university textbooks, Wikipedia pages, or other information-dense content, testing on high-frequency sentences won't tell you what you need to know. Instead, abandon one-size-fits-all evaluation strategies and choose materials and methods that are aligned with your actual synthesis goals.

8. Acknowledgements

This work is funded in part by the Swedish Vinnova funded project Deep learning based speech synthesis for reading aloud of lengthy and information rich texts in Swedish (2018-02427). The results will be made more widely accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2017-00626).

9. References

[1] M. Wieling, M. Kroon, and G. Bouma, "Write as you speak? A cross-linguistic investigation of orthographic transparency in 16 Germanic Romance and Slavic languages," in *Mining for parsing failures*. University of Groeningen, 2017.

[2] J. Taylor and K. Richmond, "Enhancing sequence-to-sequence text-to-speech with morphology," in *Proc. Interspeech 2020*, 2020.

[3] A. Perquin, E. Cooper, and J. Yamagishi, "Grapheme or phoneme? An analysis of Tacotron's embedded representations," arXiv 2010.10694, 2020.

[4] P. H. K. Seymour, M. Aro, J. M. Erskine, and Network, collaboration with COST Action A8, "Foundation literacy acquisition in European orthographies," *British Journal of Psychology*, vol. 94, no. 2, pp. 143–174, 2003.

[5] J. Taylor, S. L. Maguer, and K. Richmond, "Liaison and pronunciation learning in end-to-end text-to-speech in French," in *Proc. of SSW 11*, Budapest, Hungary, 2021.

[6] C.-C. Elert, *Ljud och ord i svenskan 2*. Stockholm: Almqvist & Wiksell International, 1981.

[7] T. Riad, "Scandinavian accent typology," *STUF - Sprachtypologie und Universalienforschung*, vol. 59, no. 1, 2006.

[8] G. Bruce, *Allmän och svensk prosodi*. Studentlitteratur, 2012.

[9] W. Ping, K. Peng, A. Gibiansky, S. Arık, A. Kannan, and S. Narang, "Deep Voice 3: 2000-speaker neural text-to-speech," in *Proc. of ICLR 2018*, 2018.

[10] K. Kastner, J. F. Santos, Y. Bengio, and A. C. Courville, "Representation mixing for TTS synthesis," in *Proc. of ICASSP 2019*, 2019.

[11] J. Fong, J. Taylor, and S. King, "Testing the limits of representation mixing for pronunciation correction in end-to-end speech synthesis," in *Interspeech 2020*, 2020.

[12] J. Fong, J. Taylor, K. Richmond, and S. King, "A comparison of letters and phones as input to sequence-to-sequence models for speech synthesis," in *Proc. of SSW 10*, 2019.

[13] Acapela Group, "Language manual, Swedish," 2005. [Online]. Available: http://www.acapela-vaas.com/Includes/language_manuals/Swedish.pdf

[14] Nordisk Språkteknologi, "PhonTable Swedish," 2002. [Online]. Available: https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-22/

[15] Amazon, "Amazon Polly: Developer Guide," 2023. [Online]. Available: https://docs.aws.amazon.com/pdfs/polly/latest/dg/polly-dg.pdf

[16] Microsoft, "SSML phonetic alphabets," 2023. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-ssml-phonetic-sets

[17] Wikipedia, "Help:IPA/Swedish," 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Help:IPA/Swedish&oldid=1120169408

[18] UCL Phonetics and Linguistics, "SAMPA for Swedish," 2004. [Online]. Available: https://www.phon.ucl.ac.uk/home/sampa/swedish.htm

[19] Wikipedia, "X-SAMPA," 2022. [Online]. Available: https://sv.wikipedia.org/w/index.php?title=X-SAMPA&oldid=51628114

[20] J. Taylor and K. Richmond, "Analysis of pronunciation learning in end-to-end speech synthesis," in *Proc. of Interspeech 2019*, 2019.

[21] H. Guo, F. K. Soong, L. He, and L. Xie, "Exploiting syntactic features in a parsed tree to improve end-to-end TTS," in *Proc. of Interspeech 2019*, 2019.

[22] M. Staib, T. H. Teh, A. Torresquintero, D. S. R. Mohan, L. Foglianti, R. Lenain, and J. Gao, "Phonological features for 0-shot multilingual speech synthesis," in *Proc. of Interspeech 2020*. ISCA, 2020, pp. 2942–2946.

[23] G. Maniati, N. Ellinas, K. Markopoulos, G. Vamvoukakis, J. S. Sung, H. Park, A. Chalamandaris, and P. Tsiakoulis, "Cross-lingual low resource speaker adaptation using phonological features," in *Interspeech 2021*. ISCA, 2021, pp. 1594–1598.

[24] A. Sanchez, A. Falai, Z. Zhang, O. Angelini, and K. Yanagisawa, "Unify and conquer: How phonetic feature representation affects polyglot text-to-speech (TTS)," in *Interspeech 2022*. ISCA, 2022, pp. 2963–2967. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2022/sanchez22_interspeech.html

[25] P. Do, M. Coler, J. Dijkstra, and E. Klabbers, "Text-to-Speech for under-resourced languages: phoneme mapping and source language selection in transfer learning," in *Proc. of SIGUL 2022*. Marseille, France: European Language Resources Association, 2022, pp. 16–22.

[26] C. Tånnander, S. Mehta, J. Beskow, and J. Edlund, "Beyond graphemes and phonemes: continuous phonological features in neural text-to-speech synthesis," in *Interspeech 2024*. ISCA, 2024, pp. 2815–2819.

[27] C. Tånnander, D. House, and J. Edlund, "Syllable duration as a proxy to latent prosodic features," in *Procs. of Speech Prosody 2022*, 2022.

[28] CereProc Ltd, "CereVoice phone sets," 2023.

[29] C. Tånnander, "Speech synthesis and evaluation at MTM," in *Proc. of Fonetik 2018*, 2018.

[30] K. Ito and L. Johnson, "The LJ Speech dataset," Tech. Rep., 2017. [Online]. Available: https://keithito.com/LJ-Speech-Dataset/

[31] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: towards end-to-end speech synthesis," in *Procs. of Interspeech 2017*, 2017.

[32] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: a flow-based generative network for speech synthesis," in *Proc. of ICASSP 2019*. IEEE, 2019.

[33] L. Borin, M. Forsberg, and J. Roxendal, "Korp-the corpus infrastructure of Språkbanken." in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 2012.

[34] T. Flynn and A. Marley, "Best-worst scaling: theory and methods," in *Handbook of Choice Modelling*. Edward Elgar Publishing, 2014.

[35] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.