Speech Prosody 2024
2-5 July 2024, Leiden, The Netherlands

Prosodic characteristics of English-accented Swedish neural TTS

Christina Tånnander[1,2], *Jim O'Regan*[1], *David House*[1], *Jens Edlund*[1] *and Jonas Beskow*[1]

[1]KTH Royal Institute of Technology
[2]Swedish Agency for Accessible Media
christina.tannander@mtm.se, {joregan,davidh,jonas}@kth.se, edlund@speech.kth.se

Abstract

Neural text-to-speech synthesis (TTS) captures prosodic features strikingly well, notwithstanding the lack of prosodic labels in training or synthesis. We trained a voice on a single Swedish speaker reading in Swedish and English. The resulting TTS allows us to control the degree of English-accentedness in Swedish sentences. English-accented Swedish commonly exhibits well-known prosodic characteristics such as erroneous tonal accents and understated or missed durational differences.

TTS quality was verified in three ways. Automatic speech recognition resulted in low errors, verifying intelligibility. Automatic language classification had Swedish as the majority choice, while the likelihood of English increased with our targeted degree of English-accentedness. Finally, a rank of perceived English-accentedness acquired through pairwise comparisons by 20 human listeners demonstrated a strong correlation with the targeted English-accentedness.

We report on phonetic and prosodic analyses of the accented TTS. In addition to the anticipated segmental differences, the analyses revealed temporal and prominence-related variations coherent with Swedish spoken by English-speakers, such as missing Swedish stress patterns and overly reduced unstressed syllables. With this work, we aim to glean insights into speech prosody from the latent prosodic features of neural TTS models. In addition, it will help implement speech phenomena such as code switching in TTS.

Index Terms: foreign-accented text-to-speech synthesis, neural text-to-speech synthesis, latent prosodic features

1. Introduction

Current *neural text-to-speech* (TTS) capture latent prosodic features that permit synthesis with unprecedented quality of prosody. In this lies the promise that TTS is instrumental to our understanding of prosody and of its interaction with other phenomena. In this work, we attempt to capture language dependent segmental and prosodic features of a single bilingual speaker to create accented speech, allowing us to understand and replicate its patterns of subtle variation. To this end, we train a Swedish neural TTS system on a single speaker's recordings of Swedish and English. The model is controlled in such a manner that it allows us to synthesise Swedish with different degrees of *targeted English-accentedness* (TEA). We verify the intelligibility and degree of accentedness of the resulting TTS renditions and report on an initial series of qualitative and quantitative analyses to describe and interpret the model output in terms of prosody. In addition to a clear association between the system's renditions of the TEA and the expected segmental variation, the results show clear prosodic aspects of accented speech.

2. Background

2.1. Accented TTS

Neural TTS is the generation of an acoustic signal (the synthesized speech) conditioned on a symbol sequence (e.g. orthographic text, graphemes, symbols for phonemes or phones). The current praxis is to use graphemic or phonemic input. More detailed phonemic input can provide better results than graphemic representations (e.g. English, [1]; French, [2]). The symbol sequence can also be enriched with other information: morphological boundaries [3], a mixed variety of levels such as both graphemes and phonemes [4], or spelling variation to target correct pronunciations [5].

[6] synthesized unseen phonemes in a new language and [7] showed that phonological features outperformed phonemes in a cross-lingual transfer learning context. The published phoneme inventories of commercial Swedish TTS voices from corporate players such as Amazon and Microsoft do not contain symbols representing English phonemes such as /w, θ, ð/ (e.g. [8, 9]), while others incorporate a full set of English phonemes in the Swedish phoneme set [10].

TTS models demonstrably capture prosodic interactions (e.g. between pitch, loudness, and duration) referred to as "latent prosodic features". The designation stems from the models' hidden handling of prosody, whose exact mechanisms remain largely unknown. The latent prosodic features provide an opportunity in that analysis-by-synthesis [11, 12] may allow us to glean new insights.

2.2. English-accented Swedish

[13] investigated the perceived accent strength of various deviant phonetic features in accented Swedish, by introducing stereotypical features of British English (and Finnish) accents in natural spoken Swedish. As the deviations presented in the study give an indication of what to expect from English-accented Swedish, they are summarized here in the order of strength from naïve judgements of accentedness when used as single feature manipulations (similar features from the study have been aggregated for brevity). We note that the feature list is a mixture of mainly segmental and mainly prosodic characteristics.

- /r/ replaced by preceding vowel lengthening, omitted or replaced by another rhotic (ranks 1, 4, and 5).
- Vowel quality changed (ranks 2 and 8).
- Vowels in unstressed syllables are reduced (rank 3).
- Acute and grave accents (i.e. tone accents 1 and 2) are confused (ranks 6 and 7).
- The distinction between long and short vowels and consonants is not made (rank 9).

10.21437/SpeechProsody.2024-209

3. Method

3.1. Objective and overall method

Our objective is to investigate the prosodic characteristics of English-accented Swedish as manifested in a purpose-built monolingual Swedish TTS model capable of varying degrees of English accent. The aim is not to develop a bilingual system but to investigate how English influences can be integrated and controlled within a Swedish phonological framework, providing insights into the interplay of foreign accent and prosody.

The Swedish phonology implemented in our TTS is adapted to cover English-accented speech by mapping English phonemes to corresponding categories within the Swedish phonological framework and by adding a *language* feature that distinguishes Swedish from English speech in the training. The former affords us segmental control while the latter provides overall control of language features.

3.2. TTS system

The training utilized the neural TTS system OverFlow [14], featuring an auto-regressive encoder-decoder architecture akin to Tacotron 2 [15]. Instead of the attention map of Tacotron, OverFlow uses a neural HMM for input-output frame alignment, enhancing robustness and training speed. It also uses a normalizing flow-based post filter for improved acoustic modeling. The spectrogram output was converted to waveforms using the HiFi-GAN vocoder [16]. The input module was modified to accept an additional continuous language symbol that was set to 0.0 (for Swedish) or 1.0 (for English) during training, allowing us to set language as a continuous value when synthesizing. The fully Swedish OverFlow voice compared successfully to a prior Tacotron 2 voice.

3.3. Training data

Our voice was trained on ~17 000 Swedish (language symbol set to 0.0 during training) and ~8 200 English (language symbol set to 1.0) sentences spoken by a female Swedish professional speaker. This speech database was originally recorded in 2011 for the Swedish *Filibuster* unit selection TTS system for university textbooks, capable of pronouncing intra-sentential English words or phrases, along with more extended English passages like text references [17].

The original segmentation of the speech database was automatic, although large parts have since been manually controlled and corrected, as have the initial lexicon-based transcriptions. Thus, the symbol set is primarily phonological, with some added phonetic detail, such that the more open variants of /ɛ/ and /œ/ that occur before /r/ and retroflexes in Swedish have separate symbols, and word pronunciations are not always fixed to fit lexicon pronunciations, but can vary according to the speaker's actual pronunciation.

The TTS system's language flag made it possible to use the same phoneme symbol in the two languages, even when their realization differs significantly. The symbol set can be said to be Swedish with the addition of English variants. For the voice used in the current work, English phonemes were mapped to their closest counterpart in the Swedish phonological system. /t/ for example was given the same symbol in Swedish and English, ignoring that in English, /t/ is alveolar rather than dental as in Swedish.

3.4. Test data

Two sets of sentences composed entirely of Swedish words were synthesized: **TestNorth**, The Swedish translation of *the North wind and the sun* as transcribed in [18] (5 sentences) and **TestGenSci**, two general and two scientific sentences constructed by ChatGPT and edited by the authors (4 sentences). In each couple, one sentence contained no /r/ sounds at all in order to ensure that human listeners in the listening test didn't solely base their perception of English-accentendness on /r/ sounds, which are a strong indicator of English accent in Swedish. The remaining sentences all contained /r/'s.

The five sentences in *the North wind and the sun* were synthesized at 9 degrees of English accent (TEA), where the first and last two represent values below and above what the system saw during training. These extremes were included as a sanity test. Each sentence was synthesized with a random seed 100 times for each TEA, resulting in a total of 9 * 5 * 100 = 4 500 renditions. For **TestGenSci**, we used five degrees of TEA where the end-points 0.0 and 1.0 represent fully Swedish and fully English versions. Each sentence was synthesized 32 times for each TEA, resulting in 4 * 5 * 32 = 640 renditions.

In the following, we label synthesized Swedish sentences with varying degrees of English accent with **s*N***, where **s** denotes synthesis, and ***N*** the TEA from 0 to 100 (e.g. **s0** for no TEA, **s25** for some, and **s100** for maximal TEA).

3.5. Verification of intelligibility

An intelligibility test was performed by running all 5 140 renditions through OpenAI's Whisper [19] using the *large-v3* model. Whisper was prompted with the language, to ensure better output quality, but otherwise default settings were used. Before calculating Word Error Rate (WER), we first normalized both the ASR output and the reference text by removing punctuation and lowercasing the output, as only the words were relevant to our task.

3.6. Verification of English-accentedness

Automatic language classification was also conducted using Whisper, in a separate pass through the model where we retained only the probabilities for Swedish and English for each rendition.

In addition, an AB test with human listeners was conducted. Four sets with 20 sentence pairs from **TestGenSci** were constructed such that in each pair, one of the sentences had a higher degree of TEA than the other. All different degrees from 0 to 100 were paired, resulting in pairs with varying expected distance (e.g. **s0** vs. **s25**, **s0** vs. **s50**, and **s0** vs. **s75**). The AB test was run on a web-based platform, *Lyss*. The 20 respondents were asked to wear headphones, and listened to the sentence pairs and answered the question (translated from Swedish): *Listen to both files and select the one you think has the highest degree of English accent.* The test took about 5 minutes to complete.

3.7. Quantitative analysis

Quantitative analyses were conducted on the entire human speech database (**REC**) and on all 5 140 synthesized test sentences (**TTS**). For phoneme durations, the segmentation already in place for the *Filibuster* TTS system was used for **REC**. Since silence-adjacent segments include some of the silence in this database, these were omitted. For **TTS**, the first

rendition of the **s0** and **s100** variants of all sentences in **TestNorth** and **TestGenSci** was segmented manually. The phonemes that had the greatest difference in average duration between **s0** and **s100** were further analyzed: a narrow window around the deviant phone in the intermediate **s25**, **s50** and **s75** renditions was also segmented manually. For f0, *REAPER* [20] was used to extract f0 for all of **REC** and **TTS**.

3.8. Qualitative analysis

Two trained phoneticians listened to a single rendition of the **s0** and **s100** variants of all nine test sentences in **TestNorth** and **TestGenSci** and took careful notes of segmental and prosodic differences.

4. Results

4.1. Verification of intelligibility and accentedness

Figure 1 (left pane) shows average WER for the different TEA. For the extraordinary targets (**s-50**, **s-25**, **s125**, **s150**), WER starts at just above 10% and increases with the target extremity, but does not reach above 40%, indicating that even with exaggerated targets, the synthesis does work to some extent. No further analysis of these targets was undertaken. For **s0** through **s100**, WER starts under 10% and increases with TEA, but does not reach 20%, indicating that the synthesis is intelligible.

Figure 1 (right pane) shows Whisper's average probability for Swedish and English for each TEA. The probability for Swedish remains over 99% for **s0, s25, s50**, then falls to nearly 96% for **s100**. In other words, the language is correctly and robustly identified as Swedish. The probability for English increases steadily from **s0** through **s100**, indicating that the English features indeed increase with increased TEA.

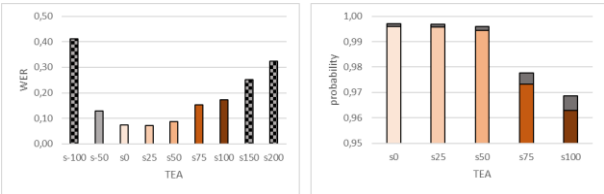

Figure 1. *WER for varying TEA (left pane) and estimated probability of Swedish (lower) and English (upper), with truncated X axis (right pane).*

As a final verification, the AB test with human listeners further confirmed that our synthesis indeed produced renditions of the intended TEA. Overall, respondents chose the rendition with the highest TEA as the most accented in 89% of the comparisons, with an individual variation between 80 and 100% "correct" judgments. A detailed inspection revealed that the pairs with the greatest distance between TEA (i.e. **s0-s100**, **s0-75**, **s25-s100**) were judged in accordance with the TEA in 99% of the cases, while for the pairs with the least difference (i.e. **s0-s25**, **s25-s50**, **s50-s75**, and **s75-s100**), the number dropped, but still reached 79% against the 50% baseline. We concluded from this that the speech was intelligible, and that the TEA was achieved, warranting further segmental and prosodic analysis of the accented synthesis. We found no difference between sentences with or without /r/.

4.2. Quantitative analysis

The number of phonemes in the nine sentences in **TestNorth** and **TestGenSci** amounts to 657, distributed on 47 combinations of 21 vowels and their stress possibilities (accent 1, accent 2, focal accent and unstressed), and 21 consonants. Only 10 vowel-stress combinations and 11 consonants occur at least 10 times, which makes phoneme averages largely meaningless. In Table 1 we report instead the average percentage difference of English/**s100** as compared to Swedish/**s0** for phonemes aggregated over feature categories. These differences exhibit a strong correlation between **REC** and **TTS** (Pearson r = 0.726, p < 0.001).

Table 1. *Durational differences in* **s100***/English compared to* **s0***/Swedish aggregated over feature categories. The Dur. diff. columns show how much shorter or longer the English phones are compared to Swedish (recordings) and how much shorter or longer* ***s100*** *is compared to* ***s0*** *(TTS sentences).*

	Dur. diff.(%)			Dur. diff.(%)	
Feature	**REC**	**TTS**	**Feature**	**REC**	**TTS**
consonant	15.54	8.46	vowel	-3.24	-5.09
stop	0.99	3.92	long	4.78	-4.70
fricative	19.67	23.96	short	-3.65	-5.35
nasal	11.15	2.88	prim.stress	-1.42	-5.86
approximant	24.02	8.46	accent 1	-0.09	-3.50
bilabial	2.46	-4.16	unstressed	-5.41	-12.81
labiodental	17.92	9.53	unrounded	-9.84	-8.29
dental	16.69	12.65	rounded	-5.58	-1.46
palatal	-12.63	10.57			
velar	2.23	8.79			
glottal	31.16	20.61			

The results from the REAPER f0 analysis of **TTS** showed a steady decrease in average f0 with increased TEA, starting from 155 Hz in **s0** and ending at 148 Hz in **s100** (Figure 2). We found no corresponding difference of averages in Swedish and English **REC** (average 152 Hz), but closer inspection revealed a left-leaning tilt of the f0 *distributions* in both English **REC** and **s100** that was not present in the corresponding Swedish **REC** and **s0**.

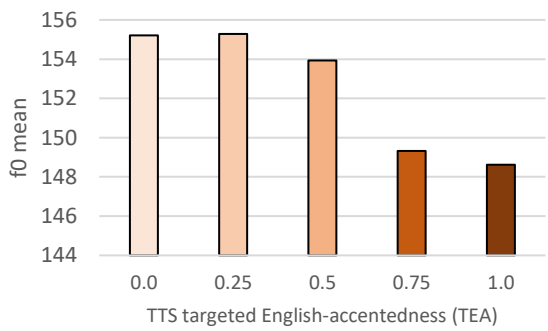

Figure 2. *Average f0 of the different degrees of English accentedness in the test data* **TestNorth** *and* **TestGenSci.** *Note that the y axis is truncated.*

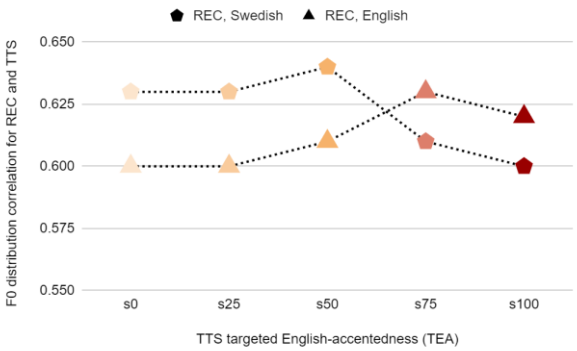

Figure 3. *Correlation (y axis; cropped at the lower end) of TTS at each TEA (x axis) Swedish and English REC (pentagons and triangles, respectively).*

Figure 3 shows correlations of f0 distributions of f0 in **s0** through **s100** in **TTS** compared to English and Swedish **REC**.

4.3. Qualitative analysis

Both phoneticians noticed a wide range of anticipated segmental and prosodic features of English accented Swedish in the **s100** renditions, and none of them in s0**.** Several of the specific features seemed to interact, which we return to in the discussion.

5. Discussion and future work

Firstly, we note that the modified OverFlow TTS framework can be used to create English-accented Swedish from Swedish and English training data. The automated verification studies show that the accented speech is intelligible (this is also confirmed by straight-forward listening) and that the degree of English features indeed increases with TEA. The latter is also quite strikingly confirmed by the AB perception test, which also shows that we are able to render varying degrees of English accent in Swedish.

Secondly, we note that our analyses of durations and f0 indicate that the TTS model captures coherently general differences in the distribution of quantity and f0. The former is evidenced by the strong correlation of durational differences aggregated over feature categories between Swedish and English in **REC** on the one hand and unaccented (**s0**) and English-accented (**s100**) in **TTS** on the other. The latter is supported by the fact that the correlation of the shape of the distribution is clearly higher when language in **REC** matches TEA in **TTS**, which the shift occurring somewhere between **s50** and **s75** (see Figure 3).

Zooming in on some detail, we see for example that the consonants in the English recordings are 15.54% longer than in the Swedish recordings, and that this phenomenon has been transferred to **TTS**, where **s100** consonants are 8.46% longer than in **s0**. This difference is mainly caused by longer fricatives, and especially so by /s/, where **s100** is 38.64% longer than **s0**.

We now turn to our anticipated English-accent features in order of their relevance as reported in 2.2.

/r/ is replaced by lengthening of the preceding vowel, is omitted, or is replaced by another rhotic (ranks 1, 4, and 5)
The absence of /r/ had no adverse effect on listeners ability to judge TEA. We note however that deletion of /r/ is highly unlikely due to the way the system is implemented; the input phonetic transcriptions to the TTS allowed no /r/-dropping. Instead, /r/ is 25.33% longer in **s100** than in **s0**. In **REC**, the English /r/'s in English sentences are roughly 44% longer than Swedish /r/'s in the Swedish sentences. From listening to the **s100** renditions, we can state unequivocally that the /r/'s in those renditions do indeed sound quite English-accented.

Vowel quality is changed (ranks 2 and 8)
Here, we rely mainly on listener judgements. The expected vowel quality changes are largely present, but the most noticeable differences have to do with vowel reduction affecting both quality and quantity (see below).

Vowels in unstressed syllables are reduced (rank 3)
This is quite noticeable, and if anything exaggerated by the model, as indicated by the differences on human readings (5% shorter in English) and **TTS** (13% shorter in **s100**). The listeners also perceive a change in quality towards a more central vowel, as expected.

Acute and grave accents (i.e. accents 1 and 2) are confused (ranks 6 and 7)
The professional listeners agree that this is often the case. Interestingly, it seems to be linked to the previously mentioned, less complex features. The combination of heavy vowel reduction where there should be none seemingly makes it difficult to contrast accent 1 and 2. We find this particularly interesting, and worthy of future investigation.

The distinction between long and short vowels and consonants is not made (rank 9)
Although this may be present in the consonant durations, it was not captured by the analyses we performed, nor is it a striking feature when listening to **s100**, where it gets overshadowed by the effects of the reduction of unstressed vowels. It is likely that we would need to render specifically designed sentences to capture this distinction.

6. Conclusions

Our Swedish OverFlow voice with an added linear layer of English language features is on par with a traditionally trained Tacotron 2 voice with phoneme input. By adjusting the language layer along the 0 to 1 continuum, we can successfully produce not only sentences with no or full English-accentedness, but also intermediate degrees of English-accentedness in Swedish sentences.

We have shown that this provides an opportunity to study specific interactions between Swedish and English, and we believe the approach is also suitable for a range of applications: for simulating other accents than English, dialects or speech disorders, or in multi-lingual TTS systems.

7. Acknowledgements

This work is funded in part by the 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).

8. References

[1] 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*, Vienna, Austria: ISCA, Sep. 2019, pp. 223–227. doi: 10.21437/SSW.2019-40.

[2] 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: ISCA, Aug. 2021, pp. 195–199. doi: 10.21437/SSW.2021-34.

[3] J. Taylor and K. Richmond, 'Enhancing sequence-to-sequence text-to-speech with morphology', in *Proc. Interspeech 2020*, Shanghai, China: ISCA, Oct. 2020, pp. 1738–1742. doi: 10.21437/Interspeech.2020-1547.

[4] K. Kastner, J. F. Santos, Y. Bengio, and A. C. Courville, 'Representation mixing for TTS synthesis', in *Proc. of ICASSP 2019*, Brighton, UK, 2019. doi: 10.1109/ICASSP.2019.8682880.

[5] J. Fong, H. Tang, and S. King, 'Spell4TTS: Acoustically-informed spellings for improving text-to-speech pronunciations', in *12th ISCA Speech Synthesis Workshop (SSW2023)*, ISCA, Aug. 2023, pp. 8–13. doi: 10.21437/SSW.2023-2.

[6] M. Staib *et al.*, 'Phonological features for 0-shot multilingual speech synthesis', in *Interspeech 2020*, Oct. 2020, pp. 2942–2946. doi: 10.21437/Interspeech.2020-1821.

[7] D. Wells and K. Richmond, 'Cross-lingual transfer of phonological features for low-resource speech synthesis', in *Proc. of SSW 11*, ISCA, Aug. 2021, pp. 160–165. doi: 10.21437/SSW.2021-28.

[8] Amazon, 'Amazon Polly - Developer Guide'.

[9] Microsoft, 'SSML phonetic alphabets'. Accessed: Nov. 22, 2023. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/speech-ssml-phonetic-sets

[10] CereProc Ltd, 'CereVoice phone sets'.

[11] C. Tånnander, D. House, and J. Edlund, 'Analysis-by-synthesis: phonetic-phonological variation indeep neural network-based text-to-speech synthesis', in *20th International Congress of Phonetic Sciences (ICPhS)*, 2023, pp. 3156–3160. Accessed: Mar. 21, 2024. [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1797227

[12] D. Hirst, 'The analysis by synthesis of speech melody: from data to models', *J. of Speech Sci.*, vol. 1, no. 1, Art. no. 1, Jul. 2011, doi: 10.20396/joss.v1i1.15011.

[13] U. Cunningham-Andersson and O. Engstrand, 'Perceived Strength and Identity of Foreign Accent in Swedish', *Phonetica*, vol. 46, pp. 138–154, Jul. 1989, doi: 10.1159/000261839.

[14] S. Mehta, A. Kirkland, H. Lameris, J. Beskow, É. Székely, and G. E. Henter, 'OverFlow: Putting flows on top of neural transducers for better TTS', in *Interspeech 2023*, ISCA, Aug. 2023, pp. 4279–4283. doi: 10.21437/Interspeech.2023-1996.

[15] Y. Wang *et al.*, 'Tacotron: towards end-to-end speech synthesis', in *Procs. of Interspeech 2017*, Stockholm: ISCA, Aug. 2017, pp. 4006–4010. doi: 10.21437/Interspeech.2017-1452.

[16] J. Kong, J. Kim, and J. Bae, 'HiFi-GAN: generative adversarial networks for efficient and high fidelity speech synthesis', in *In procs. of NeurIPS 2020*, Virtual event: Curran Associates, Inc., 2020, pp. 17022–17033. Accessed: Mar. 11, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html

[17] C. Tånnander, 'Speech synthesis and evaluation at MTM', in *Proc. of Fonetik 2018*, Gothenburg, June 7–8: Gothenburg University, 2018, pp. 75–80.

[18] 'Swedish phonology', *Wikipedia*. Oct. 31, 2023. Accessed: Jan. 11, 2024. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Swedish_phonology&oldid=1182856646

[19] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, 'Robust speech recognition via large-scale weak supervision'. 2022. doi: 10.48550/ARXIV.2212.04356.

[20] 'REAPER: Robust Epoch And Pitch EstimatoR'. Google, Dec. 11, 2023. Accessed: Dec. 12, 2023. [Online]. Available: https://github.com/google/REAPER