

The effects of phonological content, sentence information, and vocoding on voice cue perception

Supplementary Materials S1

Thomas Koelewijn^{1,2}, Etienne Gaudrain^{1,2,3}, Thawab Shehab^{1,4}, Tobias Treczoks^{1,5}, & Deniz Başkent^{1,2}

¹ Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, Netherlands

² Research School of Behavioural and Cognitive Neurosciences, Graduate School of Medical Sciences, University of Groningen, Groningen, Netherlands

³ Lyon Neuroscience Research Center, CNRS UMR5292, Inserm U1028, Université Lyon 1, Lyon, France

⁴ University of Groningen, faculty of Arts, Neurolinguistics, Groningen, Netherlands

⁵ Medical Physics and Cluster of Excellence "Hearing4all", Department of Medical Physics and Acoustics, Faculty VI Medicine and Health Sciences, Carl von Ossietzky Universität Oldenburg, Germany

The stimuli used in Experiment 1 were extracted from the VariaNTS corpus, which is freely available on Zenodo (<https://doi.org/10.5281/zenodo.3932038>). The stimuli used in Experiment 2, both sentences and isolated words, were extracted from the “VU zinnen” corpus (Versfeld et al., 2000). The f_0 and vocal-tract length (VTL) of these stimuli were then manipulated using the WORLD vocoder (Morise, 2015, 2016; Morise et al., 2016, 2017; see Morise & Watanabe, 2018 for an evaluation of WORLD’s output quality) through the PyWORLD wrapper (J. Hsu; <https://github.com/JeremyCCHsu/Python-Wrapper-for-World-Vocoder>).

Spectrograms of the example audio stimuli are provided as supplementary materials. In addition, GIF animations of the spectrograms have been produced to further illustrate the effect of these manipulations on the stimuli. Finally, WAV and FLAC files for Experiment 1 are provided in Supplementary Material S2. The audio files of Experiment 2 were not included for copyright reasons.

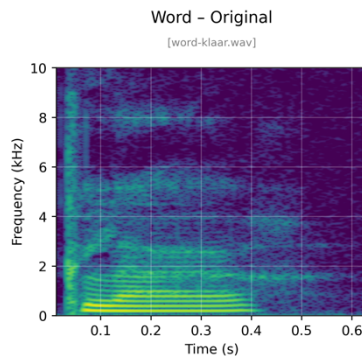
All spectrograms were generated using Scipy (v1.7.3, Virtanen et al., 2020) using 1024-long Hann windows overlapping by 921 samples. In each window, a 2048-point FFT (zero-padded) was used to estimate the power spectral density.

Experiment 1 – Phonological Content

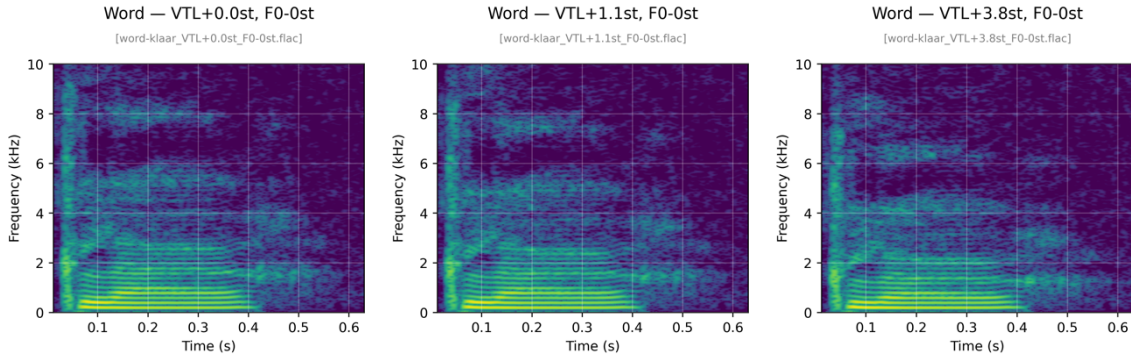
In this experiment, words, non-words and time reverse words were used. Only the VTL was manipulated, while the f_0 remained unchanged. For each condition, a single item is used as example, but keep in mind that every trial of the adaptive just-noticeable difference (JND) task was made of three different items.

Words

In this example, we chose the Dutch word “klaar” (which translates to “ready”). The average VTL JND for words was 1.1 semitone (st). Below is first displayed the spectrogram of the original sound file `world-klaar.wav`.



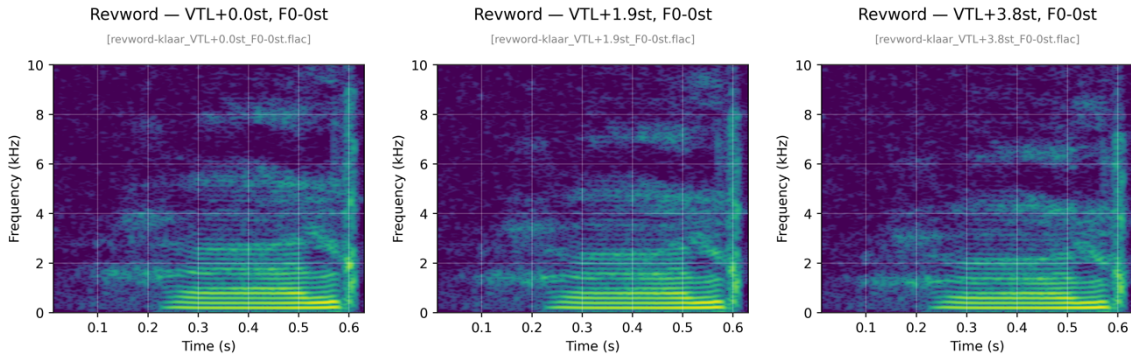
All sound files were processed with WORLD, even when the VTL and f_0 were unchanged. The spectrograms of the resynthesized sound without change, at the VTL JND, and for a +3.8 st VTL shift are displayed below.



One can observe that the frequencies of the harmonic structure remain unchanged while the spectral envelope, which represents formants, shifts down with VTL, affecting the magnitude of all harmonics. This effect can be more easily visualized in `word-klaar_VTL.gif`.

Time reversed-words

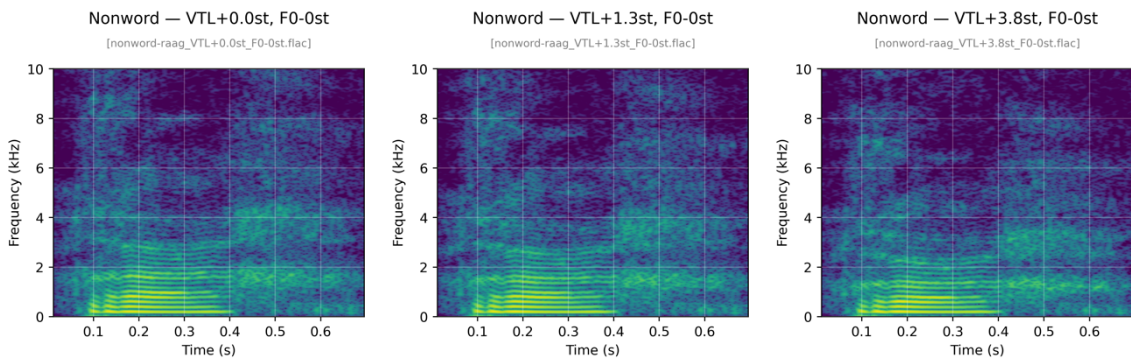
We chose the same word to illustrate the time-reversed condition. In that condition, the average VTL JND was 1.9 st. The same three panels are shown below.



The corresponding animation is `revword-klaar_VTL.gif`.

Non-words

Finally, we chose “raag” as non-word. In this condition, the average VTL JND was 1.3 st.



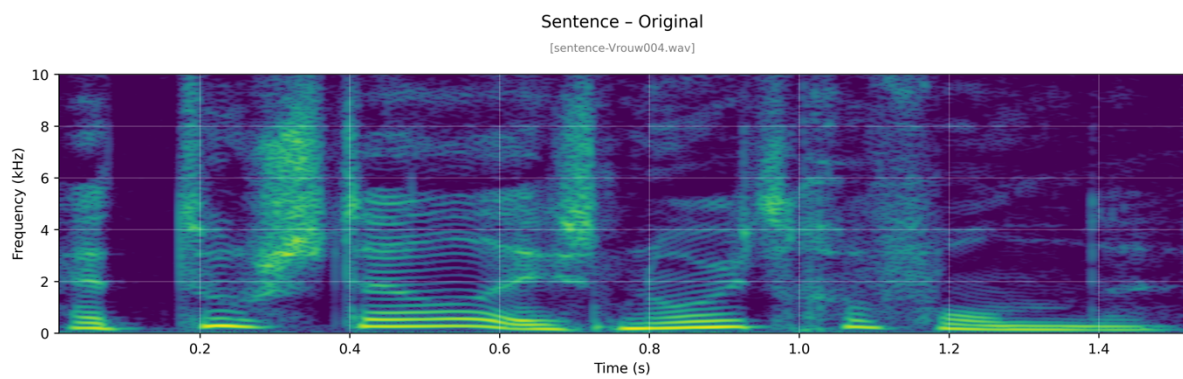
The corresponding animation is `nonword-raag_VTL.gif`.

Note that, for all examples, we also included an audio sample with a VTL shift of +3.8 st and an f_0 shift of -12 st, which corresponds to a convincing female to male voice transformation.

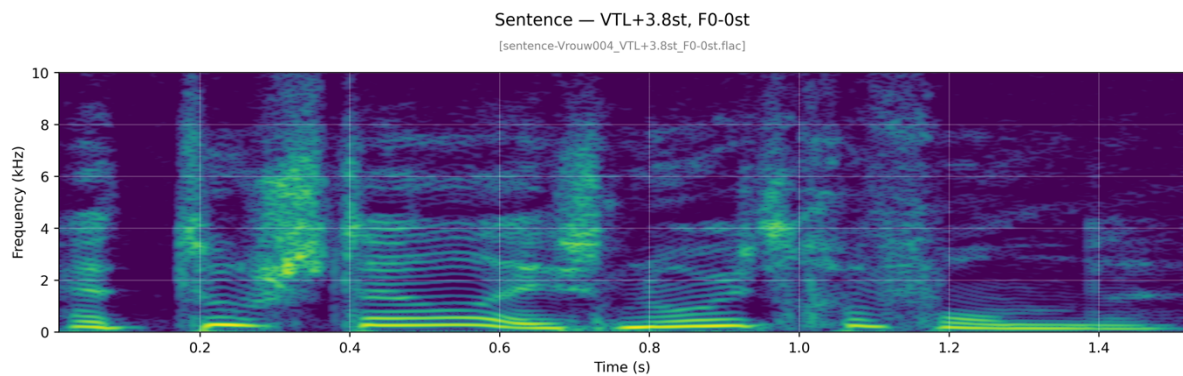
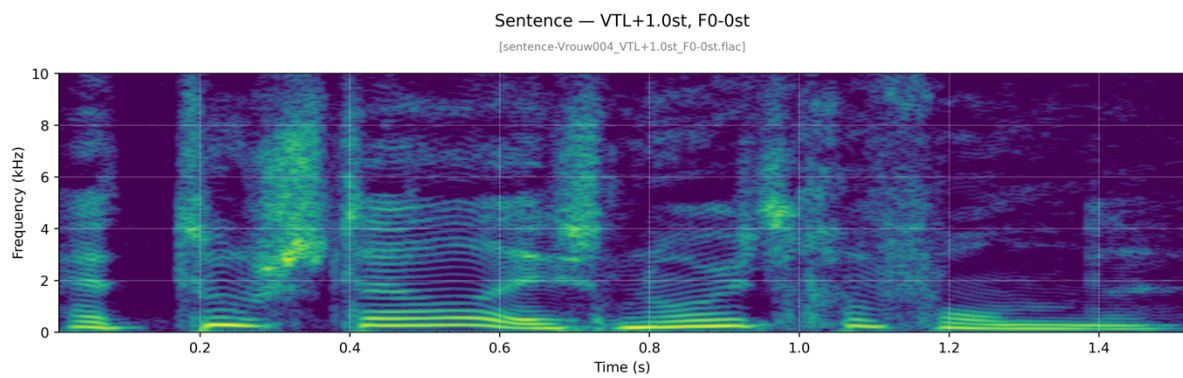
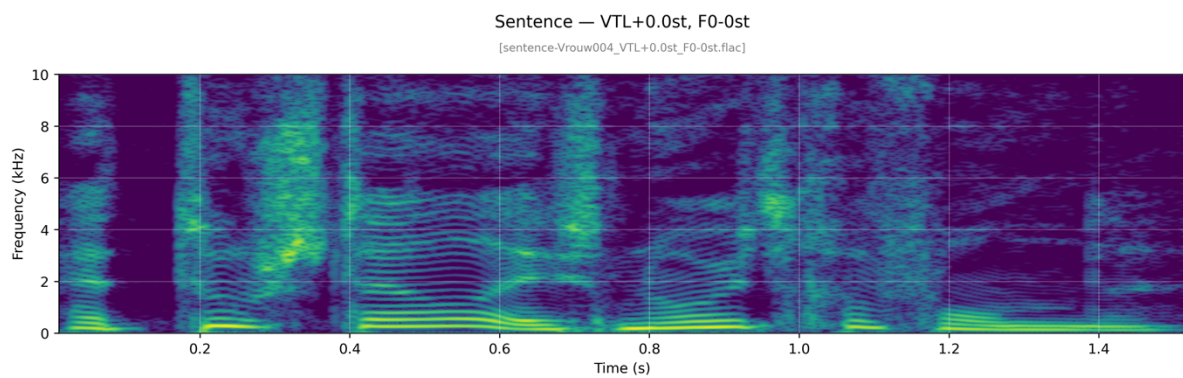
Experiment 2 – Words vs. Sentences

In the second experiment, words were compared to sentences, and both the f_0 and the VTL JNDs were measured.

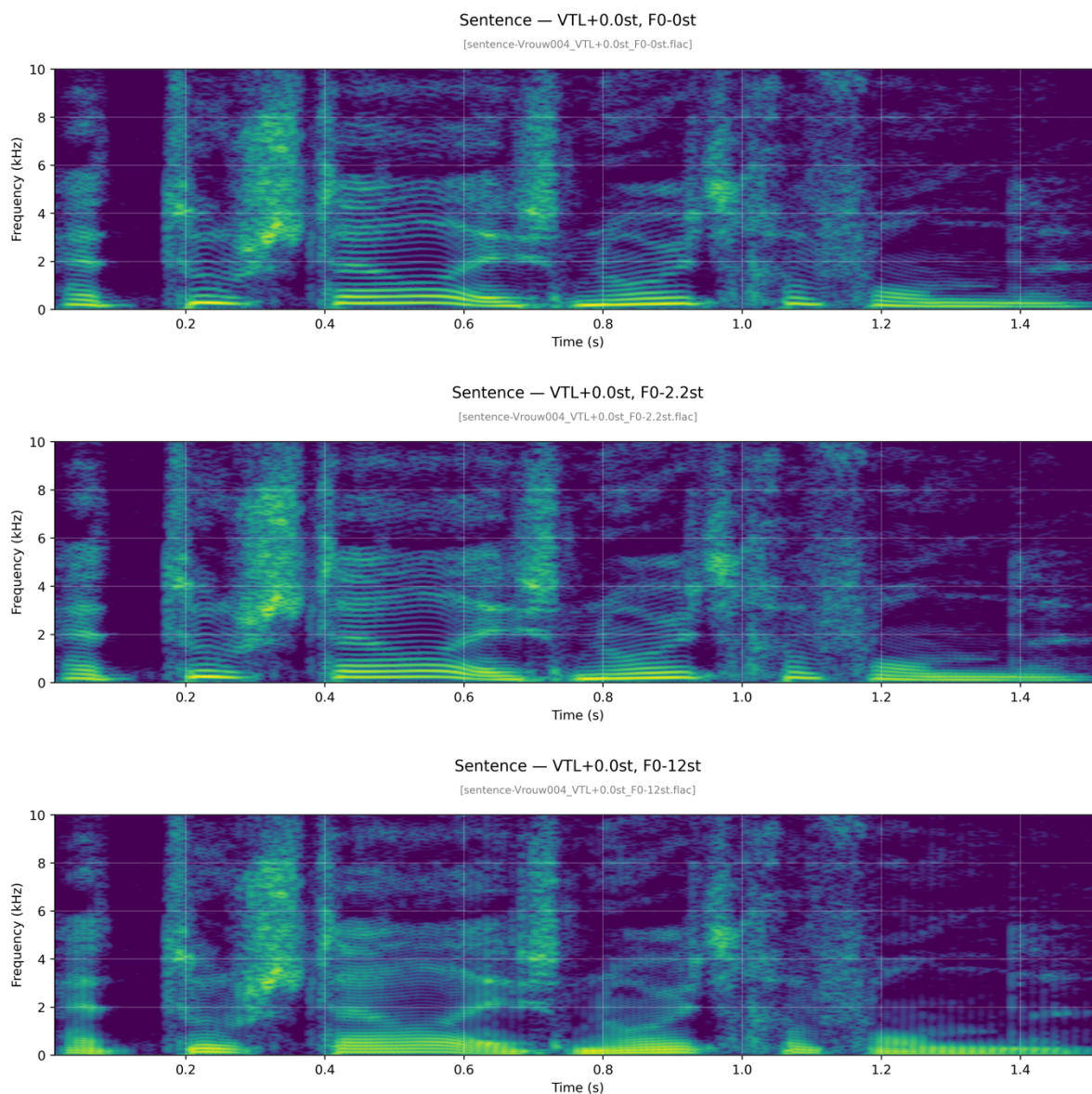
Sentences



Manipulation of VTL — the average VTL JND for sentences was 1.0 st:



Manipulation of f_0 — the average f_0 JND for sentences was 2.2 st:

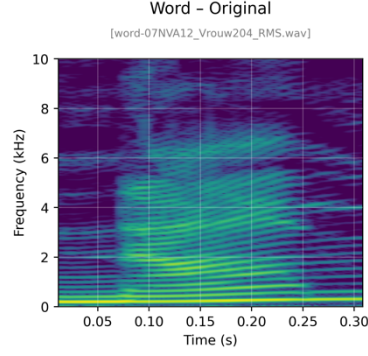


Note that the spectrogram's window length is fixed, which means that when f_0 becomes small, it appears as temporal modulation rather than harmonic structure.

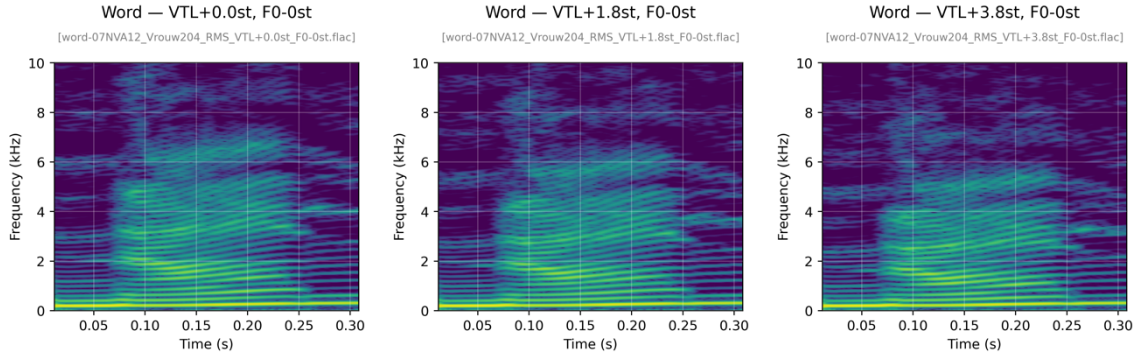
The animations corresponding to the VTL and f_0 manipulations are `sentence-Vrouw004_VTL.gif` and `sentence-Vrouw004_F0.gif`, respectively.

Words

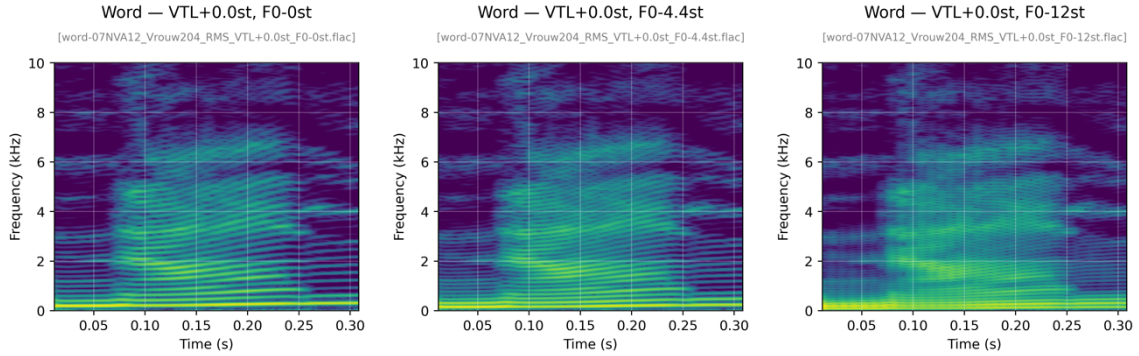
The Dutch word “naam” (“name” in English) was chosen as example for the isolated word condition. The words were excised from the VU sentences.



Manipulation of VTL — the average VTL JND for isolated words was 1.8 st:



Manipulation of f_0 — the average f_0 JND for isolated words was 4.4 st:



The animations corresponding to the VTL and f_0 manipulations are `word-07NVA12_Vrouw204_RMS_VTL.gif` and `word-07NVA12_Vrouw204_RMS_F0.gif`, respectively.

Note that, for all examples, we also included an audio sample with a VTL shift of +3.8 st and an f_0 shift of –12 st, which corresponds to a convincing female to male voice transformation.

Licenses and consent

The VariaNTS corpus is distributed under a Creative Commons Attribution 4.0 International license (CC-BY 4.0). Informed consent was obtained from all the participating speakers.

The VU sentences corpus can be obtained by contacting the authors of Versfeld et al. (2000).

References

- Morise, M. (2015). CheapTrick, a spectral envelope estimator for high-quality speech synthesis. *Speech Communication*, 67, 1–7. <https://doi.org/10.1016/j.specom.2014.09.003>
- Morise, M. (2016). D4C, a band-aperiodicity estimator for high-quality speech synthesis. *Speech Communication*, 84, 57–65. <https://doi.org/10.1016/j.specom.2016.09.001>
- Morise, M., Miyashita, G., & Ozawa, K. (2017). Low-Dimensional Representation of Spectral Envelope Without Deterioration for Full-Band Speech Analysis/Synthesis System. *Interspeech 2017*, 409–413. <https://doi.org/10.21437/Interspeech.2017-67>
- Morise, M., & Watanabe, Y. (2018). Sound quality comparison among high-quality vocoders by using re-synthesized speech. *Acoustical Science and Technology*, 39(3), 263–265. <https://doi.org/10.1250/ast.39.263>
- Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D(7), 1877–1884. <https://doi.org/10.1587/transinf.2015EDP7457>
- Versfeld, N. J., Daalder, L., Festen, J. M., & Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *The Journal of the Acoustical Society of America*, 107(3), 1671–1684. <https://doi.org/10.1121/1.428451>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>