PERCEPTION OF FORMANT DISTORTION IN GERMAN WORDS AND NON-WORDS

Uliana Eliseeva[1], *Ivan Yuen*[2], *Bernd Möbius*[3]

[1,2,3]*Department of Language Science and Technology, Saarland University, Germany*
[1]*123elmark@gmail.com,* [2]*ivyuen@lst.uni-saarland.de,* [3]*moebius@lst.uni-saarland.de*

Abstract: Concatenative text-to-speech (TTS) systems remain a widely used cheaper alternative to neural TTS systems. Yet concatenation of prerecorded units entails some drawbacks, such as spectral distortion, the perceptual consequences of which remain unclear. In an attempt to bridge this gap, our study focused on the effect of spectral distortion in vowel formants on perceived speech quality in naturally-read manipulated German words as well as non-words. More specifically, we explored the distortion effect on a varying number of affected formants, at different magnitude and directionality in two corner vowels /a:/ and /i:/. The results indicate that single formant manipulations have a less pronounced effect on the listeners' perception compared to multiple formant perturbations. The threshold at which the distortion became generally audible was estimated to lie between 0.4 and 1.0 bandwidth. The directionality of the distortion was not found to be significant.

1 Introduction

With neural text-to-speech (TTS) being the state-of-the-art approach for speech synthesis, more traditional methods such as concatenative TTS are still widely used as a cheaper alternative to synthesise voices. Concatenative systems use a corpus of natural speech which is cut into smaller units, typically diphones, where speech sounds that are subject to coarticulation are split in the middle and concatenated with other units to produce the desired speech output. Although the unit candidates for the output are selected so as to minimise acoustic discrepancies at their junction, concatenation of prerecorded units almost always entails some distortion, such as spectral one, the perceptual consequence of which is not well understood.

Although there is considerable variation in formant frequencies in naturally produced vowels [1], formant discontinuities of the type observed at the concatenation point do not occur in natural speech. Olive et al. [2] mentioned that two recordings of a phrase produced by one speaker with varying vowel frequencies of F2 were reported to sound identical in a comparison task. But they also found that after concatenating those segments, the produced 50 Hz formant mismatch became audible. This indicates that audibility of spectral change at concatenation point is affected by some specific factors.

One of these factors could be the lexical effect which was shown to alter formant discrimination behaviour. In the experiments by Kewley-Port and Zheng [3], vowel formants F1, F2, and F3 were gradually shifted up to 10%. The vowels were presented in isolation, in syllables, phrases and sentences. Each stimulus included recordings of a modified and an unmodified speech segment. In each experiment, listeners were exposed to a target vowel and asked to select the recording that contained the same vowel. The results indicated a significant drop in manipulation threshold of F2 for syllables compared to phrases and sentences.

A manifold nature of speech perception is perhaps one of the main limiting factors to finding a reliable relation between acoustics and perceptual listening scores. As of now, the studies

 DOI: 10.35096/othr/pub-7079, URN: urn:nbn:de:bvb:898-opus4-70793

investigating effects of acoustic feature mismatches on speech perception, especially spectral features, often report inconsistent results. Slight changes in methodology such as varying discourse, synthesised voice gender, and speech material length from two to multiple sentences [4] or different formulations of listening tasks [5, 6] yielded different correlations between spectral distance measures and perceptual scores, ranging drastically from 0.17 to 0.78 (Pearson correlation coefficient).

To address this issue, our study explored the connection between spectral distortion and speech quality perception in a more systematic way. We investigated the effect of spectral distortion in vowel formants on perceived speech quality in words and non-words, exploring its magnitude, directionality, and the number of affected formants. Word-length stimuli were assumed to be enough to produce lexical effects without masking the effects under investigation. Speech-like non-word stimuli were utilised to observe to what extent lexical processing affects the threshold of formant discrepancy.

2 Material and Methods

2.1 Research Questions and Hypotheses

In this work, we posed three main research questions:

RQ1 When distorted, do vowel formants differ in their effect on the perceived quality of word-length speech?

RQ2 Is there an effect of magnitude of the distortion?

RQ3 Is there an effect of directionality of the distortion?

Based on the findings from formant discrimination studies [7, 8], we expected that the formants would not induce a significant difference (H1). Furthermore, increasing the number of distorted formants was hypothesised to increase listeners' sensitivity to the manipulation because combinations of formants quantitatively is a more severe vowel degradation that affects vowel identity and we might observe an additive effect (H2).

As to the magnitude, we hypothesised that the distortion has to be of at least 0.5 bandwidth to be perceptually detectable (H3) as the reported 50 Hz for F2 in Olive et al. [2] correspond to this bandwidth value (see Material section for bandwidth explanation).

With respect to the directionality of the discrepancy, it was assumed that there might be an interaction between the directionality and vowel identity as well as directionality and formant (H4). The formant space of the vowels under analysis permits the formant range to be wider only in one direction (up or down) without overlapping with other vowels in their proximity. For instance, F1 can be extended up in /aː/ and down in /iː/ as there are no other vowel phonemes in that area.

2.2 Material

The stimuli for our perceptual experiment included pairs of monosyllabic or disyllabic German words and non-words with target vowels /aː/ and /iː/ in the stressed position where the three lower formants were manipulated. The speech material was recorded by a 28 year old native male speaker of Standard German and his acoustic production was verified by another native German speaker. The words and non-words (referred to as items) are listed in Table 1. They were selected so that the target vowels were in the least coarticulated environment between

Sound	Words	Non-words
/aː/	Tat, Staat, Daten, Zitat, Diktat, Mandat, Dativ	Ftaat, Gdaat, Taate, Sotaat, Letaat, Podaat, Sadaat
/iː/	Titel, Titer, Titus, Bandit, Kredit, Tektit, Dieter	Tit, Tita, Rotit, Wodit, Adit, Klatit, Ditum

Table 1 – Word and non-word items for stimuli

plosive alveolar consonants [2]. The average item duration was 0.708 seconds (SD=0.147 seconds, range=0.445–1.186 seconds) and the average target vowel duration was 0.259 seconds (SD=0.072 seconds, range=0.123–0.420 seconds).

Imitation of the formant discrepancy occurring in concatenative TTS would imply introducing a discontinuity in the spectral envelope of these vowels. To this end, we first confined the discrepancies to two types, an upward shift and a downward shift in formant values, wherein the trajectory of the formant stayed unchanged. The shift was defined for individual formants as well as for their combinations with one constraint – they must move in the same direction in order to simulate the inter-speaker acoustic variability (e.g., due to vocal tract size). Within-speaker variation was simulated with single-formant shifts. We chose formant bandwidth as a scale for our shifts motivated by the need to make different formant shifts more comparable among each other and relate them to human perception of vowels to some degree [9, 10]. Bandwidth values were fixed at 60 Hz, 100 Hz, 180 Hz for F1, F2, F3, respectively, regardless of the actually produced frequency of the vowels.

We manipulated the target vowels' spectrum in our stimuli with legacy-STRAIGHT [11], a tool for analysis, modification and synthesis of natural speech that was utilised in a number of studies on auditory perception where a spectrum was modified in some way [8, 7, 12]. The method to perform manipulations was adopted from the studies by Liu and Kewley-Port [8, 7]. In short, the spectrum obtained in STRAIGHT was changed so that the formant peaks were shifted by some specified factor and the gaps were filled with the value of the adjacent dip. Vowels manipulated with this method were pretested for naturalness. Except for the extreme values of 2.0 BW, the perceived naturalness of the resynthesised vowels was not significantly different from the original ones. In our main experiment, we used three equidistant shift values of 0.4 BW, 1.0 BW and 1.6 BW.

The process of creating the final stimuli included cutting out the target vowel together with its immediate context. Next, the spectrum for the segment was extracted with STRAIGHT and modified according to the algorithm described above. After that, the segment was resynthesised with STRAIGHT. The control item was resynthesised as a whole without manipulations. Next, a copy of the control item and the modified segment were cut for concatenation. The manipulated vowel segment was manually selected for each item and started approximately at the temporal mid point of the vowel after it has reached the articulatory target. The manipulated vowel part was concatenated with the preceding and following parts from the control item using Praat concatenation function with an overlap of 0.01 seconds. Figure 1 demonstrates the final test and control items' formant trajectories.

In the experiment, test stimuli were presented as pairs of items, namely *control item* + *test item*. In control stimuli, both items were control items, in other words, neither of them contained a manipulation. The total number of stimuli was 196, which consisted of 168 test stimuli, 14 control stimuli and 14 fillers.

2.3 Method

2.3.1 Participants

The data were collected from 69 native German participants either the via online research platform Prolific or in person. Data from one participant were discarded as their mean response

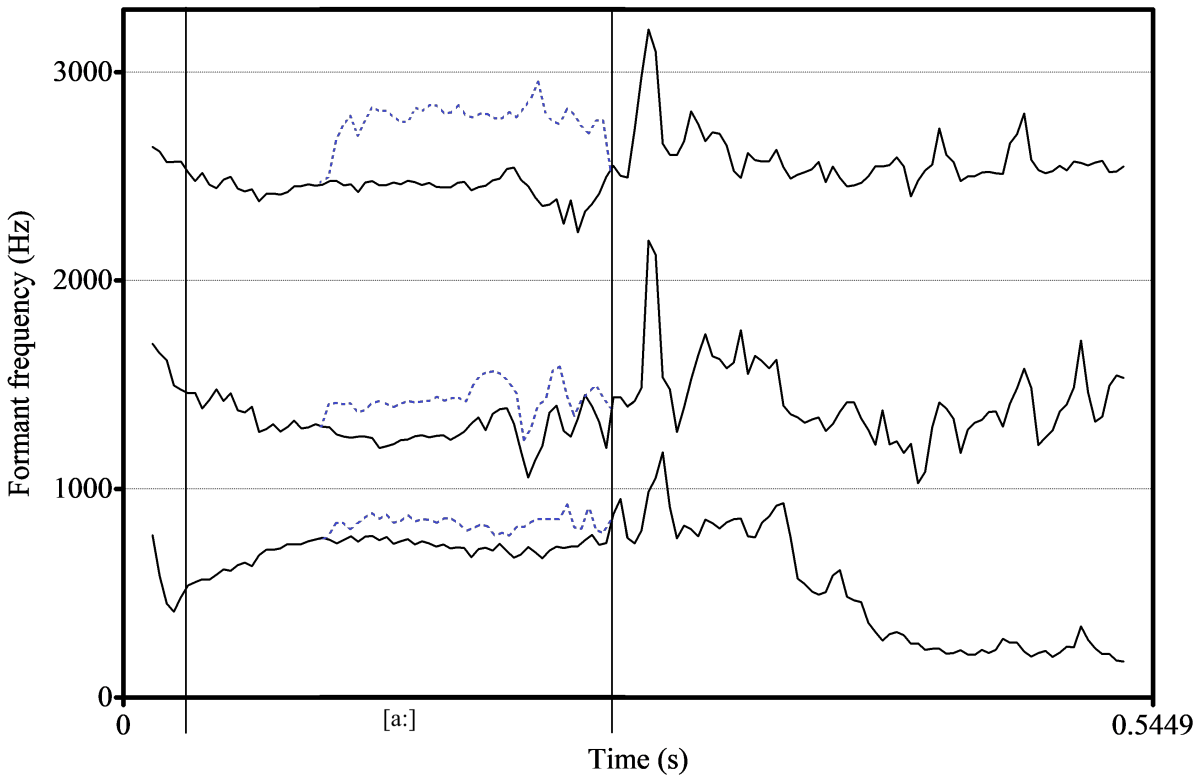

Figure 1 – Formant tracks of word *Daten*: control item (solid) and test item with 1.6 BW upward shift on F1, F2, F3 (dashed)

time was 0.79 seconds (SD=0.37, median=0.72 sec) which is 3.4 times faster than the mean of the rest of the participants (mean=2.68 sec, SD=2.01, median=2.00 sec) after high outlier replacement.

The average age of the participants was 26 years (SD=5.9 years, range=18-41 years), 44% of the participants identified as male and 56% as female. The majority (64.7%) reported to speak one foreign language, while 25% were monolingual and seven participants stated to speak more than one foreign language (two and three). Five participants indicated to have hearing difficulties, however this factor turned out to be insignificant in the process of model selection and their data was included in the analysis.

2.3.2 Procedure

The experiment was programmed in PsychoPy [13], version 2022.2.5. For the online launch, the code was integrated with the Pavlovia server designed by the PsychoPy developers.

On-site participants were tested individually in a quiet room at Saarland University. We used a laptop with loudspeakers and the preset volume of 48 dB(A). The experimenter left the room after introducing participants to the task. Online participants were allowed to use any type of PC and were reminded to set up their volume to at least 50%[1].

At the start of the experiment, the participants read instructions in German which mentioned that the study was about perception but did not reveal the purpose of the experiment. We used a forced-choice binary preference task and asked the participants to listen to audio recordings of words and non-words in pairs and indicate which of the two items sounded more pleasant.

The instructions were followed by two test blocks with words and non-words separately. Both blocks started with a 5-trial practice round. Within the blocks, the stimuli were presented in random order.

[1]This corresponds to 40-50 dB in most laptops and PCs.

A trial screen depicted two sound icons corresponding to two items. The orthographic form of the items was not shown to the participants. Each item in a trial could be played for a maximum of three times and had to be played at least once before the preference keys were activated. The experiment was self-paced and the participants could change their preference as many times as they wished before proceeding to the next trial.

In order to reduce the risk of fatigue, the Prolific participants were exposed to half of the stimuli in each block. For that, the stimuli were split into two sessions with a balanced design for all conditions in each session. The average completion time for the online participants was 12 minutes including the instruction and survey blocks. The on-site participants were exposed to all the stimuli and took 21 minutes on average. Per trial, participants spent 4.23 seconds on average (SD=2.03 seconds, range=1.08-29.80 seconds).

2.4 Data Analysis

The final dataset contained 50 observations per test stimuli. The statistical analyses were conducted in RStudio with R version 4.1.1 [14].

The response variable was encoded as 1 (a participant preferred the item **without** manipulation) or 0 (a participant preferred the item **with** manipulation). We will call the ratio of preference of items without manipulation over the sum of the two types of preference a *ratio score*. The ratio score was used as a dependent variable in the analysis.

The control condition was dummy coded as having 0 magnitude and its ratio score was set to chance (0.5) within and across all grouping factors.

A generalised linear mixed effects model (gLME) specified with the binomial family from the lme4 package (version 1.1-31) was used for data analysis. We performed a backward model selection starting with all the random factors with intercepts and slopes for participants' Age, Number of Languages, Hearing, Gender, and Item. Our final random effect structure for every model included Age and Item random effects with random intercepts only. It also should be noted that the variation in the participation mode (online vs. in person) did not result in a significant difference between the participants. Therefore, data were collapsed in the analysis.

As to the fixed structure, we observed a constant influence of Vowel Duration (VD) and Word Duration (WD) on the manipulation audibility and included both VD and WD without interaction in the fixed structure of every model. More precise model specifications depended on the research question and hypothesis.

3 Results

The gLME models' results were analysed with the car::Anova function which uses a Chi-Square Test of Independence under the hood.

With respect to RQ1, the Formant factor was significant overall ($\chi^2(6) = 34.1$, $p<.001$). To address the question on whether spectral distortion at multiple formants would worsen perceived speech quality, we compared the manipulations based on the number of formants involved. The stimuli were grouped as single formant manipulation (SFM), i.e. F1, F2, or F3, or multiple formant manipulation (MFM), i.e. F1F2, F1F3, F2F3, F1F2F3. This grouping turned out to have an effect on listeners' preference, $\chi^2(1) = 29.3$, $p<.001$.

Within single and multiple levels, the Formant factor did not show any significance, indicating that formants did not differ from one another in single and multiple manipulation. This result is consistent with the findings of Liu and Kewley-Port [7] and supports H1. On the other hand, H2 is only partially supported as there were no differences between two- and three-formant manipulation groups.

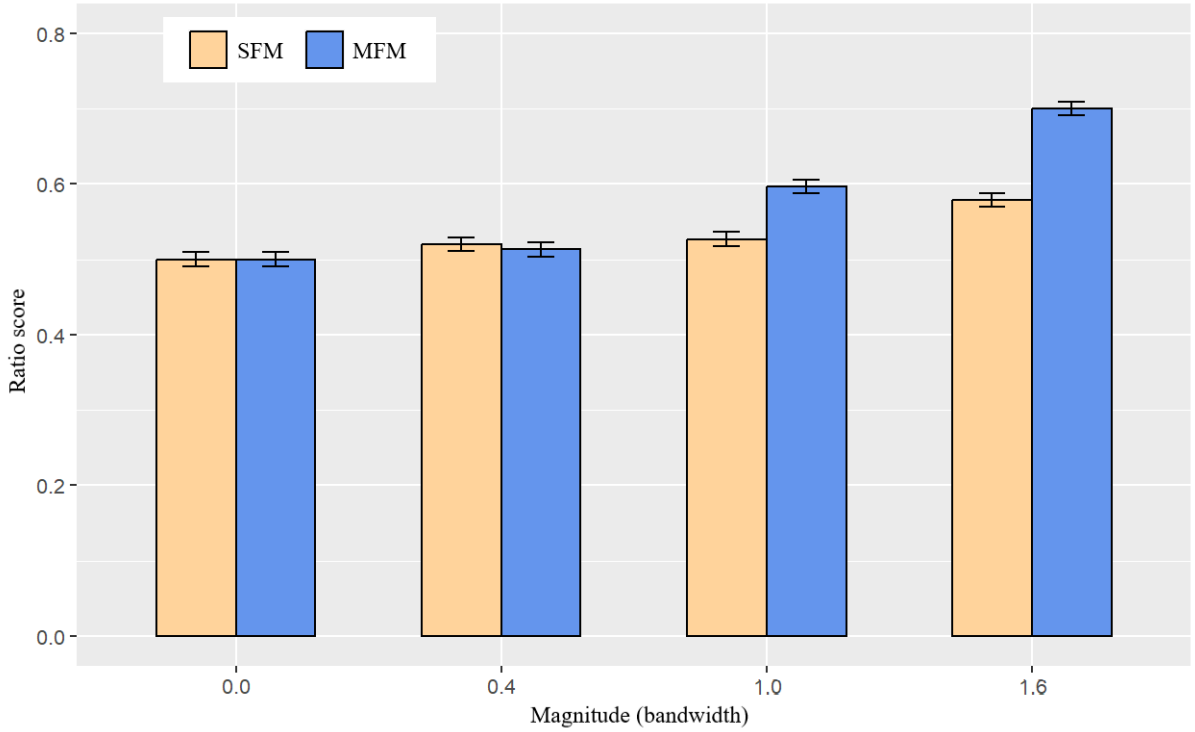

Figure 2 – Mean ratio score by magnitude and formant type

Regarding the magnitude of the manipulation, we saw that Magnitude ($\chi^2(3) = 149.5$, p<.001) played a role in human preferences overall, confirming RQ2. The magnitude effect gained significance starting at 1.0 BW. The manipulation of 0.4 BW was not different from control, providing evidence for H3.

In the general model, Formant and Magnitude interaction almost reached statistical significance ($\chi^2(18) = 28.7$, p=.052) while this interaction reached actual significance when the formant factor was grouped under two levels: SFM and MFM ($\chi^2(3) = 17.9$, p<.001). As evident in Figure 2, the difference between SFM and MFM becomes larger when the magnitude of the manipulation increases and with a larger magnitude of formant distortion the MFM appears to be more detrimental to perception than SFM.

Notably, the Magnitude in SFM was found to interact with Word Type ($\chi^2(3) = 9.7$, p<.05). Non-words had a stronger Magnitude pattern while words seemed to be less sensitive to it. We investigated this interaction further and found that Magnitude did not appear to be significant in words where either F1, F2, or F3 was manipulated ($\chi^2(3) = 5.8$, p=.12). Vowel, on the other hand, reached significance in this model, $\chi^2(1) = 5.8$, p<.05. The results indicate that Vowel /aː/ seemed to be the only source of variance here and /iː/ was less affected by the manipulations.

As to the directionality of the formant distortion, the model did not attribute any significant variance to this factor ($\chi^2(1) = 0.79$, p=.37). We could not assess its interaction with other factors as it led to non-convergence, presumably due to the lack of data. Thus, H4 can be neither supported nor rejected.

4 Discussion and Conclusion

In this study, we investigated to what extent acoustic formant discontinuities in concatenative TTS affect human perception in subjective quality evaluation. To this end, we conducted a perception experiment to determine listeners' sensitivity to different types of discrepancies in vowel formants in German words and non-words. The manipulations and different contexts for the vowels aimed to elicit effects that would either confirm or disprove our expectations of

which factors are relevant for speech quality perception.

It was hypothesised that distortions of formant combinations was more detrimental than that of single formants as this would imply an additive and therefore quantitatively larger distortion. This was found to be partly true as we observed an increase in listeners' sensitivity to multiple formant manipulations compared to single formants, a consistent effect regardless of word type and vowel and intensified by larger magnitudes of the manipulation. But we did not observe this effect for two-formant compared to three-formant conditions. This disagrees with the assumption of the additive effect of formant distortion on perception. As for the individual formants, no effect was found, suggesting that listeners were equally sensitive to F1, F2 and F3 discrepancies expressed on the bandwidth scale, indicating that vowel identities were either not affected strongly enough or that listeners did not perceive any ambiguity in the manipulated vowels. This showed that the acoustic characteristics alone could not predict listeners' behaviour in a straightforward manner.

On the same note, the vowel effect was generally not present. Hypothetically, a more densely populated phonemic space around /iː/ should promote less tolerance to the allophones of this vowel which we expected to observe as an increase in sensitivity to any manipulation on this vowel. On the contrary, the /iː/ vowel did not differ from the /aː/ vowel in all conditions except SFM in words. Interestingly, the listeners did not hear the manipulation of any magnitude in this specific condition. Potentially, some of the words with target /iː/ could have affected our results because *Titel, Bandit* and *Kredit* have two acceptable pronunciations with long /iː/ and short /ɪ/ which makes the formant range for these words larger. Another explanation for the lack of a vowel effect is that it was not possible to estimate the interaction between Vowel and Direction which could reveal the differences between the vowels. For example, along the frequency axis an increase in F1 of /iː/ and decrease in F1 of /aː/ make the vowels acoustically closer to other vowels as F1 approaches the centre of the vowel space.

Nevertheless, the lack of a vowel effect shows that our data does not support the assumption that the proximity to other vowels can decrease tolerance to the manipulation. This also indicates the need for a review of how crucial vowel prototypicality, one of the components of the TTS objective function, is for overall speech quality perception.

Another interesting observation arises from the absence of the word type effect everywhere except in SFM in interaction with the magnitude factor. This only difference between words and non-words provides some evidence for lexical processing taking place. Words with single formant shifts showed no magnitude effect, presumably inhibited by lexical retrieval. However, it is unclear why the lexical effect was not elicited in words with multiple formant distortions.

Lastly, the hypothesis that the minimum audible shift should be of 0.5 bandwidth seems to be not far from the truth. Our findings show that the audibility threshold for formant distortions lies between 0.4 BW and 1.0 BW. Compared to Liu and Kewley-Port [7] who report 0.37 Barks as their general audibility threshold in formant discrimination studies, our audibility threshold for formant distortions lies approximately between 0.16 Barks (0.4 BW) and 0.42 Barks (1.0 BW) when averaged across vowels and formants. Given that, we would assume that the threshold rather leans towards a full bandwidth.

One also has to keep in mind that this range is estimated for the word-length items and shorter or longer speech segments could facilitate or inhibit the effect. Provided that lexical items do require larger shifts to overcome lexical processing, higher order speech chunks, where retrieval is followed by integration, will probably have an audibility threshold much higher than the range we could look into in this study.

References

[1] PETERSON, G. E. and H. L. BARNEY: *Control methods used in a study of the vowels*. *Journal of the Acoustical Society of America*, 24, pp. 175–184, 1952.

[2] OLIVE, J., J. VAN SANTEN, B. MOEBIUS, and C. SHIH: *Synthesis*. In R. SPROAT (ed.), *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach*, pp. 213–218. Kluwer Academic Publishers, 1998.

[3] KEWLEY-PORT, D. and Y. ZHENG: *Vowel formant discrimination: Towards more ordinary listening conditions*. *The Journal of the Acoustical Society of America*, 106, pp. 2945–58, 1999. doi:10.1121/1.428134.

[4] MÖLLER, S., F. HINTERLEITNER, T. H. FALK, and T. POLZEHL: *Comparison of approaches for instrumentally predicting the quality of text-to-speech systems*. In *Proc. Interspeech 2010*, pp. 1325–28. 2010. doi:10.21437/Interspeech.2010-413.

[5] HUANG, D.-Y.: *Prediction of perceived sound quality of synthetic speech*. *APSIPA ASC 2011*, pp. 977–982, 2011.

[6] HINTERLEITNER, F., S. ZABEL, S. MÖLLER, L. LEUTELT, and C. NORRENBROCK: *Predicting the quality of synthesized speech using reference-based prediction measures*. In *Proc. ESSV*, pp. 99–106. 2011.

[7] LIU, C. and D. KEWLEY-PORT: *Vowel formant discrimination for high-fidelity speech*. *The Journal of the Acoustical Society of America*, 116, pp. 1224–33, 2004b. doi:10.1121/1.1768958.

[8] LIU, C. and D. KEWLEY-PORT: *Straight: A new speech synthesizer for vowel formant discrimination*. *Acoustics Research Letters Online*, 5, pp. 31–36, 2004a. doi:10.1121/1.1635431.

[9] DE CHEVEIGNÉ, A.: *Formant bandwidth affects the identification of competing vowels*. In *ICPhS*, pp. 2093–96. 1999.

[10] ISHIKAWA, K., M. MEYER, J. MACAULSAN, and S. BOYCE: *Predicting intelligibility of dysphonic speech with automatic measurement of vowel related parameters*. *The Journal of the Acoustical Society of America*, 141, pp. 3838–38, 2017. doi:10.1121/1.4988541.

[11] KAWAHARA, H., I. MASUDA-KATSUSE, and A. CHEVEIGNÉ: *Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds*. *Speech Communication*, 27, pp. 187–207, 1999. doi:10.1016/S0167-6393(98)00085-5.

[12] ASSMANN, P. and W. KATZ: *Synthesis fidelity and time-varying spectral change in vowels*. *The Journal of the Acoustical Society of America*, 117, pp. 886–895, 2005. doi:10.1121/1.1852549.

[13] PEIRCE, J., J. R. GRAY, S. SIMPSON, M. MACASKILL, R. HÖCHENBERGER, H. SOGO, E. KASTMAN, and J. K. LINDELØV: *Psychopy2: experiments in behavior made easy*. *Behavior Research Methods*, 51, pp. 195–203, 2019. doi:10.3758/s13428-018-01193-y.

[14] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.