Software voices and how we perceive them: The communicative potential of timbre transferred stimuli

1. Introduction

Emergent technologies in audio synthesis and manipulation have historically enabled the pursuit of new research questions within fields such as linguistics, acoustics, cognitive science, and psychoacoustics. According to Agus et al. this development can be traced from the use of magnetic tape technologies in the 1960s, through digital synthesis and vocoders in the later part of the 20th century (Agus et al., 2019). The present paper can be seen as a two-fold excursion into a new chapter of audio manipulation techniques, at once approached as a tool to revisit long-standing research-questions within the tradition of linguistics and psychoacoustics, as well as something which demands the formulation of novel research questions entirely.

A major question within linguistics and other fields of research interested in speech and audio perception is the semantic contributions of non-linguistic features such as dynamic patterns, pitch contour, and timbres. Linguist John J. Ohala identifies research into such form-function relationships as somewhat risky, potentially jeopardizing Saussure's famous principle declaring the arbitrariness of the sign in relation to its meaning. Ohala's own work draws on zoologist Eugene S. Morton's examination of form-function relationships in animal vocalizations to suggest, that human vocalizations carry social messages via their intonation (Morton, 1977; Ohala, 1995). Ohala dubs this theory 'the frequency code', which in short states that vocalizations with lower fundamental frequencies (F0) or pitch-falls are used to express domination, whereas higher F0's and ascending pitch contours express submissiveness. Gussenhoven and Chen have used the frequency code theory to explain the ability of subjects being able to recognize manipulated speech stimuli in a pseudo language as either questions or statements, the questions being perceived as submissive request compared to the more assertive statements. (Gussenhoven & Chen 2000). Similar effects have been identified in instrumental music, where lower and harsher sounds are considered more dominant than their high-pitched counterparts (Huron et al., 2006).

Research into the semantic information carried via non-linguistic features have often relied on technical stimuli designed to isolate certain of these features to increase experimental control: Remez et al. have illustrated how artificial sinusoidal sounds synthesized to mimic the frequency curve and dynamic temporal patterns of a spoken sentence could be perceived as language and that certain subjects could extract non-present phonological information from pitch and dynamic patterns alone, presenting a challenge to the theories of language perception at the time (Remez et al., 1981). Ohala used a similar kind of stimuli to study the assertiveness of high and low-pitched utterances 'stripped' of words and devoid of any spectral information (Ohala, 1982), whereas Shannon

and colleagues extracted dynamic temporal patterns such as onset envelopes from speech recordings and applied them as time-varying filters to noise signals with different numbers of frequency bands, finding that subjects were able to perceive words from as few as three modulated noise bands (Shannon et al., 1995). More recently, researchers such as Suied et al. and Isnard et al. have constructed stimuli intended to isolate the effects of timbre in sound recognition (Suied et al., 2014; Isnard et al., 2016). In both cases, very short audio-clips with similar pitch and loudness but differing timbral features were used as stimuli in a sound recognition task. Isnard et al. used 'auditory sketches' first developed by Suied et al. 2013, utilizing algorithmic spectral reduction to create simplified sound-tokens of speech as well as instrumental and environmental sounds (Suied et al., 2013).

Whereas most of the experiments mentioned above relied on stimuli obtained either via subtractive processes removing the undesired features from the signal through audio manipulations (as in Ohala, 1982 and Isnard et al., 2016) or by constructing artificial stimuli from the ground-up using tools such as synthesis to create deliberately abstract simplified representations of speech (as in Remez et al., 1981 and Shannon et al., 1995), recent technologies allow for stimuli-creation using a combination of subtractive[1] and synthetic approaches. In the present paper, such a stimuli-set is constructed and tested in a pilot study designed to examine its ability to carry communicative messages via non-linguistic features. Specifically, the google-developed technology DDSP (Differential Digital Signal Processing) is utilized to create a test battery of vocal utterances manipulated via timbre transfer techniques, which enables the timbral profile from one sound (a) to be mapped onto the structure of another sound (b), while retaining the dynamic pattern and frequency contour of sound b (Engel et al., 2020). Following the question-intonation phenomena illustrated by Gussenhoven and Chen among others, a collection of recorded speech samples with and without a final pitch rise are transformed into various non-speech timbres including both instrumental and environmental sounds.

The initial goal of the project is to test whether participants can recognize the transformed speech samples as communicative linguistic entities functioning as either questions or statements, despite their lack of speech-like timbres and lexical information. To accommodate this goal, the following hypothesis is tested experimentally:

> *H_0: There is no significant difference in the pattern of accurately recognizing sound samples as either statements or questions between an experimental group presented with timbre transformed stimuli and a control group presented with unprocessed utterances.*

[1] The term subtractive here denotes the process of arriving at a reduced stimuli by isolating certain vocal features in a sample of natural speech, as opposed to a constructive method based on synthesis, whether additive or subtractive.

Since the stimuli-design is based on what Christine Bartels refers to as a "*simplified notion of question intonation*" (Bartels, 1999), a few comments on the discussion regarding intonation and sentence function is due. One point from Bartels' dissertation is that far from all questions in the English language are characterized by a final pitch rise. She points to studies suggesting that the opposite phenomena is statistically more common for yes-no questions (Bartels, 1999). Gussenhoven & Chen similarly report that languages such as Chickasaw display opposite form-function relationships where intonational rises signify statements and falls are used for questions (Gussenhoven & Chen, 2000). Finally, Christine Gunlogson argues that intonation alone cannot transform an utterance from statement into question on its own; the utterance is embedded in a context assisting the listener in decoding the proper communicative function of the given vocalization, hence she encourages a multidimensional understanding including more factors besides the final pitch variation (Gunlogson, 2003).

Despite the controversy surrounding question intonation, the assumption that pitch contours by themselves and isolated from any linguistic context can denote utterances as either statements or questions seems highly beneficial when seeking to establish perceptual effects of timbral manipulations. By comparing the recognition accuracy-rate of the manipulated sounds with the performance of a control group exposed to the unmanipulated samples, the assumption that a simplified question intonation can enable successful discrimination between questions and statements is simultaneously put to the test.

2. Experiment

2.1 Task

To test whether the timbre transferred stimuli can carry communicative messages, sets of manipulated speech samples were designed and presented to participants via the online survey tool *SoSci Survey* (v.3.4.06). The experiment URL was shared through various communicative channels such as social media and e-mail among both linguistically naïve and linguistically knowledgeable potential participants. The survey was online for 7 days and gathered data from 16 voluntary participants. No personal data besides native language(s) was registered from participants, ensuring an anonymous design. Participants clicking the link were automatically and discretely distributed between the four experimental conditions via a random generator set to draw without replacement. After being presented with an introduction screen and indicating their native language via a multiple-choice selection, participants were presented with 12 randomly shuffled sounds of 2 sec. length, manipulated according to the participant's assigned condition. After each sound, the participant was asked to evaluate and indicate (by clicking one of two boxes) whether the sample sounded more like a question or more a statement. Repeated playback

of the sound was allowed. The experiment would automatically terminate after the participant had evaluated all sounds.

One potential draw-back of the online survey format is that the experimental situation takes place in a less controlled environment. A sound-test was included on the second page of the questionnaire to attempt to ensure, that participants had audible playback before the stimuli was presented, but no effort to standardize audio playback devices was attempted.

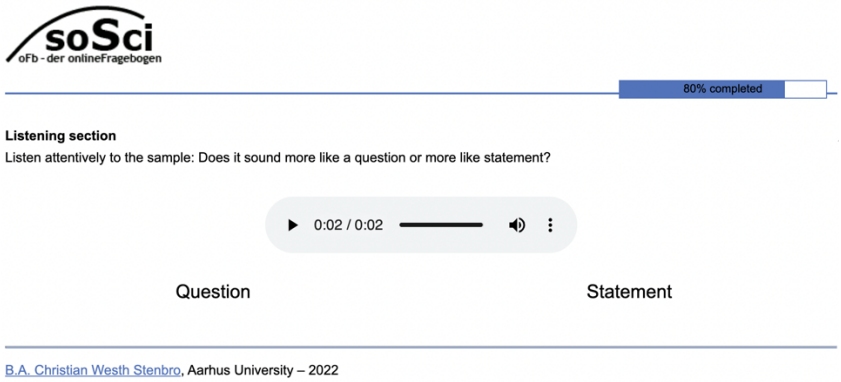

Figure 1: Reproduced screenshot from the online experiment sound-presentation page

2.2 Stimuli

Four different stimuli test-batteries were designed, recorded, and finally manipulated via timbre transfer tools as implemented in three publicly available Google-Colab notebooks provided by Engel et al. and Alonso & Erkut respectively (Engel et al., 2020; Alonso & Erkut, 2021). The process of stimuli creation is broken down into four segments to clarify the different steps.

2.2.a Designing Speech Utterances

Six pairs of sentences were borrowed or constructed to present a range of questions and statements with varying lengths and levels of complexity. The concept of these pairs is inspired by Gunlogson's dissertation, where she presents pairs of declaratives with rising and falling intonations at the end (Gunlogson, 2003). Functionally, the rising intonation changes the declarative from a statement to a question although grammatically it maintains the form of a statement. The advantage of using declaratives such as these where pitch is the only marker, is that it rejects the possibility of participants basing their perception on an analysis of grammatical features in the control condition. Three declarative pairs with varying lengths are borrowed directly from Gunlogson. The simplest of these consists of a pronoun-verb contraction and a participle (the rain-example presented in figure 2). Three additional pairs were constructed, one of them consisting of

a single word [“Oranges.” | ”Oranges?”] and two of them utilizing a pseudo language to cover both different sentence-lengths and to examine whether the intended function of the utterance is perceived by the control group independent from lexical meaning.

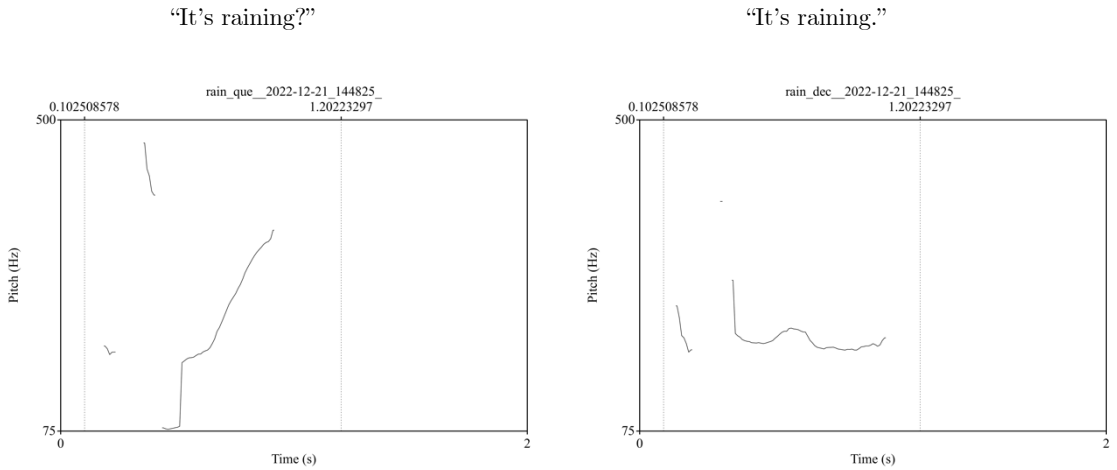

Figure 2: Top: Extract from manuscript used to record question/statement pairs, this one originally presented in Gunlogson, 2003. Bottom: Pitch contours for the recorded stimuli pair computed by the author via Praat.

2.2.b Recording and Editing

Multiple versions of each sentence were recorded in two settings varying between an empathized final pitch rise for the question declaratives and no pronounced final rise for the statement declaratives. The speaker was instructed to emphasize the final pitch rise for the question-variants as a too subtle variations could possibly lead to too ambiguous pitch contours after the timbre transfer process. Here, an exaggerated approach was deemed more constructive keeping the goal of the experiment in mind. From these recordings, the most representative pairs were selected based on repeated listening and visual assessment of pitch contours computed using Praat (Boersma and Weenink, 2022). Generally, the samples with the least jittering pitch contours displaying the final pitch rise clearly (in the case of the question-variant) were preferred as stimuli. The audio was recorded to Ableton Live (v.10) in .wav format at 44.1 kHz using a RODE NT5 microphone and a Focusrite Scarlett 6i6 audio interface. Additional audio processing such as cutting away excess silences and standardizing the sample lengths were attended to within the Ableton Live session.

2.2.c Training Watery Sound Model

Transforming timbres requires the construction of timbral models. These are digital entities containing data to reconstruct one sound sample’s timbral features within the structural features of another sound. While two of the employed models (violin and flute)

are native to the Google demo, a third model was trained specifically for this project to test whether a less tone-like timbre designed to resemble environmental sounds such as streaming water would yield a different recognition pattern. The model should cover a range of variations in dynamics and pitch to accommodate the loudness and intonation patterns in natural speech samples, since a model trained on data without a given acoustic feature would not know how to respond to such a feature in the sound it is supposed to replicate. However, even when designing the data to accommodate such shortcomings, the timbre transfer technology is reported to have issues with acoustic features such as silences, generating low frequencies in their place (Alonso & Erkut, 2021).

Physical modelling tools were utilized to synthesize watery sounds, such as those of a flowing river or a liquid being poured from a cup into a flooded sink, via the Sound Design Toolkit (Baldan et al., 2017) as implemented in the visual programming language Max (v.8). During the synthesis, real-time control over parameters such as density and pitch ratio enabled the construction of a varied sound dataset covering a wide range of amplitudes and pitch-ranges, in turns making a flexible timbre model. Approximately 20 minutes of unedited synthesized water sounds were used to train a timbral model via a modified version of the Google colab script *01_ train.ipynb* developed by Alonso & Erkut and available from Alonso's GitHub repository (Alonso & Erkut, 2021) [1]. The training-process followed the procedure specified in the README.mp file of this repository. The version of DDSP was modified from 1.6.2 to 1.9.0 by the author to solve dependency conflicts.

2.2.d Timbre Transfer

Each intonational variant of each of the six speech samples were transformed into three timbres. Specifically, the Colab notebook *DDSP Timbre Transfer Demo* by Engels et al. were used to transfer the flute and violin timbres, whereas Alonso & Erkut's *02 run* notebook were used to transfer the watery timbre. During the process of transformation, sensitivity to quiet audio parts were adjusted to obtain samples with a more accurate representation of the speech recordings pitch contour. According to Engel et al. the DDSP timbre transfer extracts both F0 (the fundamental frequency) and loudness features from the sound to be resynthesized via the timbral model. While this suggests a non-destructive process, many of the processed samples contain artefacts, some of which effect the pitch contour of the sound, which might create more ambiguous samples. Some of these artefacts, such as sudden short pitch-jumps occurring immediately after the utterance, were manually removed, although artefacts occurring within sentences between words remained in the final samples.

Ultimately, a stimuli-set of 48 samples each of 2 sec. length were constructed. All sounds, including those for the control-group, were down-sampled automatically to 16.000 Hz

during the timbre-transfer process and were subsequently converted to MP3 at 128 kbps via FFMPEG run from the command line (Tomar, 2006).

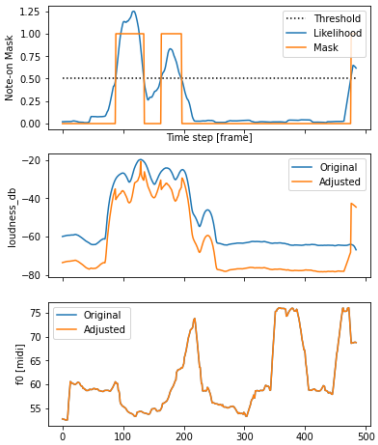

Figure 3: Reproduced screenshot of loudness and pitch information extracted from a speech sample (blue) and readjusted to guide the re-synthesis of the timbre transferred sound. The plot is automatically generated by the Google Colab *DDSP Timbre Transfer Demo* during the transfer process.

3. Analysis

A few words on experimental design and the choice of statistical test are due before moving on to the analysis results. For the data analysis a logistic regression model was fitted to accommodate the binomial nature of the response data obtained via the alternative forced choice test. The experiment is a between-subjects design where each subject is assigned to a single condition group based on the random generator implemented in the online questionnaire. This means that no subject is exposed to stimuli from more than one condition. While this design ensures independence of data-points between conditions, dependencies within the conditions could appear due to the test-battery stimuli presentation method, where each subject responds to 12 sound items. This could be considered a violation of the independence assumption of the logistic regression model, although the degree to which subjects would vary individually in this specific recognition test is unclear. Ultimately, a naïve logistic regression model not taking possible random variations for individual subjects into account was deemed the more feasible choice due to the small sample size of the study. 16 participants contributed data to the experiment, with as few as 3 participants in some conditions. A more complex model would require more participants to accurately estimate the random effects, whereas using the naïve model taking each answer as an independent sample, effectively boosts the sample-size to $n =$

162, increasing the statistical power of the model.[2] Implications of this choice will be treated in greater detail in the discussion section.

3.1 Logistic Regression Model as Hypothesis Testing

To test the hypothesis stating that the recognition pattern is unaffected by subject condition, a logistic regression model using timbre condition as predictor was fitted in R version 4.2.1:

$$\text{p(Recognition)} \sim \text{Timbre Condition}$$

A chi-squared test revealed that including the timbre condition as predictor significantly improved the fit of the model, $\chi^2(3) = 22.53$, $p < .000$. Each timbre condition's effect on recognition probability can be accessed via the model's estimated coefficients: The effect on recognition accuracy for the unmanipulated *voice* timbre condition is given in the model intercept, which is found to significantly increase the probability of recognition ($\beta = 3.37$, SE = 0.72, z = 4.68, $p < .000$) compared to a model using a constant as predictor. The *flute* timbre condition is estimated to significantly decrease the probability of accurate recognition compared to the baseline condition ($\beta = -2.86$, SE = 0.78, z = -3.67, $p < .000$). A similar pattern appears in the remaining timbre transfer conditions: The model estimates a significant decreasing effect for the *violin* timbre condition ($\beta = -1.76$, SE = 0.85, z = -2.08, $p < .05$), and the *water* timbre condition ($\beta = -1.90$, SE = 0.81, z = -2.35, $p < .05$), although the beta coefficients are less steep compared to the *flute* condition, meaning that the decreasing effect is less pronounced for these conditions.

A prediction table with 95% confidence intervals is computed to ease the interpretation of the model's predicted accuracy rate for each condition. The prediction table is converted into a plot visualizing the same information:

	Condition	Rounded_LB	Probability	Rounded_UB
1	water	0.68	0.81	0.9
2	violin	0.68	0.83	0.92
3	flute	0.48	0.62	0.75
4	voice	0.88	0.97	0.99

Figure 4: Probability Table

[2] Since each of the 16 participants were subjected to 12 individual sound stimuli or trials, the sample size is calculated as: *n = 16 * 12 = 162*

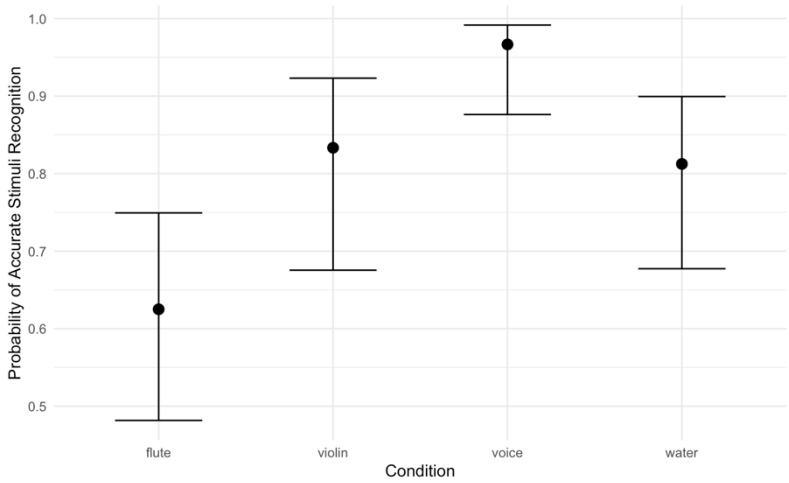

Figure 5: Plot visualizing the predicted probabilities for accurate recognition of stimuli with 95% confidence intervals for each of the four conditions.

4. Discussion

The hypothesis testing revealed that the state of timbre condition effected the predicted probability of accurately recognizing the communicative intend of the stimuli. In other words, the null hypothesis formulated in the introduction stating that there is no difference in the recognition pattern between the experimental groups and the control group is rejected. Even if the null hypothesis is rejected this does not mean, that the timbre transferred stimuli failed in carrying communicative messages; on the contrary, the prediction table computed via the coefficient estimates reveals that the predicted probabilities for all conditions but the flute timbre range well above 50% accuracy, meaning that more than half of the time subjects would be able to recognize the intended communicative message embedded within the manipulated sound samples. The highest accuracy-rate is found in the violin condition where the predicted probability of accurate recognition is estimated at 83% within a confidence interval spanning from 68% to 92%. In comparison, the predicted probability for the water condition is 81% with lower and upper confidence intervals of 68% and 90% percent. Perhaps surprisingly, the flute condition yields the lowest predicted probability of accurate prediction at 62%, with confidence intervals spanning from 48% and 75%.

The high recognition rate for two out of three of the timbre transferred stimuli categories would suggest that subjects assigned to these conditions could successfully decode the communicative messages without access to the specific timbral or lexical information that normally characterizes speech, in turn suggesting that the sounds were considered as linguistic or at least communicative entities. These results would seem to support findings from the classic studies by Remez et al. as well as Shannon et al. referenced in the

introduction, suggesting that subjects can percieve sinusoidal tones and noise signals as speech and even extract non-present linguistic information from such signals insofar as they are manipulated to resemble either the dynamic patterns or pitch contours found in speech (Remez et al., 1981; Shannon et al., 1995). It is reasonable to speculate that these same temporal patterns in pitch and dynamics are responsible for the relatively high accuracy rate found in the recognition of utterances manipulated to sound as violin- and water sounds, since both the pitch contour and loudness curves of the original utterances are preserved in the manipulated samples (see figure 3).

The lower prediction for accurate stimuli recognition in the flute condition warrants further discussion. One interpretation is that certain timbres – in this case those created via the violin and the watery model – are better at communicating in gestures borrowed from the speaking voice, although the theoretical foundation for such a claim is lacking. As the experiment does not provide data to address these differences, any attempts at such remain speculative until further research has been conducted. Another proposed explanation for the deviation in estimated accuracy-rates for the flute condition relates less to inherent qualities of the timbre somehow obstructing recognition of the communicative intend, than it does to errors and glitches in the timbre transfer process. While the violin and water timbre models perform reasonably well, the poorer performance of participants in the flute condition could suggest difficulties specificality in mapping the flute timbre model onto speech samples, creating artifacts such as pitch glissandos not derived from pitch contour extracted from the utterance. Further experiments would benefit from a spectral analysis of the pitch and dynamic information found in the transformed stimuli to ensure a more tightly controlled test-battery, ruling out the possibility of non-timbre variations creating different recognition patterns between conditions.

Finally, due to decisions during the experimental design-phase as well as the small amounts of participants drafted, the statistical power of the results are worth discussing. The logistic regression model fails to incorporate possible random effects rising from individual differences, ignoring potential dependencies of data points. Taking these methodological issues into consideration, one should be careful not to generalize based on the results of the pilot study or overstate the importance of its findings. Despite these shortcomings, the experiment displays the potential of using DDSP timbre transfer to create novel types of test stimuli for research-use, combining communicative features traditionally associated with speech and suggested to be perceivable in this study with timbral profiles extracted from disparate sources such as instrumental and environmental sounds.

Further work could explore the possibility of merging communicative messages from the world of language with semantic features residing within the timbres themselves. Can people perceive sounds as both carriers of linguistic messages *and* as something signifying

a sound source, whether thunder or crashing glass or barking dogs at the same time? Demonstrating such abilities could enable new methods for sound-design in games and virtual worlds to create singing rivers, whispering winds, or questioning insects. Similarly, one could take inspiration from Yee and Bailenson's proposed 'proteus effect', stating that the appearance of a user's avatar in a VR space can affect the user's behavior (Yee & Bailenson, 2007). As showcased by artist Holly Herndon, timbre transfer can be implemented to perform real-time voice transformations, effectively allowing people to talk or sing in voices modelled on other people's voices (Herndon, 2022). Future research could investigate whether speaking in voices with different timbral profiles could alter not just speech perception on the end of the listener but also the communicative choices of the speaker in terms of phrasing, intonation, or vocabulary. Would we talk differently if we spoke in different voices? Via emergent technologies the pursuit of such strange, yet fundamental questions of voices and their connection to human identity are within reach.

References

Agus, T. R., Suied, C., & Pressnitzer, D. (2019). Timbre Recognition and Sound Source Identification. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 59–85). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_3

Alonso, J., & Erkut, C. (2021). *Explorations of Singing Voice Synthesis using DDSP*. https://doi.org/10.5281/ZENODO.5043851

Baldan, S., Delle Monache, S., & Rocchesso, D. (2017). The Sound Design Toolkit. *SoftwareX*, *6*, 255–260. https://doi.org/10.1016/j.softx.2017.06.003

Bartels, C. (1999). *The intonation of English statements and questions: A compositional interpretation*. Garland Pub.

Boersma, Paul & Weenink, David (2022). Praat: doing phonetics by computer [Computer program]. Available from http://www.praat.org/

Engel, J., Hantrakul, L., Gu, C., & Roberts, A. (2020). *DDSP: Differentiable Digital Signal Processing* (arXiv:2001.04643). arXiv. http://arxiv.org/abs/2001.04643

FFmpeg Developers. (2022). ffmpeg tool (Version 5.1) [Software]. Available from http://ffmpeg.org/

Gunlogson, C. (2003). *True to Form: Rising and Falling Declaratives As Questions in English*. Taylor & Francis Group. http://ebookcentral.proquest.com/lib/asb/detail.action?docID=182951

Gussenhoven, C., & Chen, A. (2000). Universal and language-specific effects in the perception of question intonation. *6th International Conference on Spoken*

Language Processing (ICSLP 2000), vols 2, 91-94–0. https://doi.org/10.21437/ICSLP.2000-216

Herndon, H. (2022). What if you could sing in your favorite musician's voice? [Video]. TED Conferences. Retrieved from: https://www.ted.com/talks/holly_herndon_what_if_you_could_sing_in_your_favorite_musician_s_voice on the 4th of January 2023

Huron, D., Kinney, D., & Precoda, K. (2006). Influence of Pitch Height on the Perception of Submissiveness and Threat in Musical Passages. *Empirical Musicology Review*, *1*(3), 170–177. https://doi.org/10.18061/1811/24068

Isnard, V., Taffou, M., Viaud-Delmon, I., & Suied, C. (2016). Auditory Sketches: Very Sparse Representations of Sounds Are Still Recognizable. *PLOS ONE*, *11*(3), e0150313. https://doi.org/10.1371/journal.pone.0150313

Ohala, J. J. (1982). The voice of dominance. *The Journal of the Acoustical Society of America*, *72*(S1), S66–S66. https://doi.org/10.1121/1.2020007

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science (New York, N.Y.)*, *212*(4497), 947–949. https://doi.org/10.1126/science.7233191

RStudio Team (2022). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Suied, C., Agus, T. R., Thorpe, S. J., Mesgarani, N., & Pressnitzer, D. (2014). Auditory gist: Recognition of very short sounds from timbre cues. *The Journal of the Acoustical Society of America*, *135*(3), 1380–1391. https://doi.org/10.1121/1.4863659

Suied, C., Drémeau, A., Pressnitzer, D., & Daudet, L. (2013). Auditory Sketches: Sparse Representations of Sounds Based on Perceptual Models. In M. Aramaki, M. Barthet, R. Kronland-Martinet, & S. Ystad (Eds.), *From Sounds to Music and Emotions* (Vol. 7900, pp. 154–170). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41248-6_9

Yee, N., & Bailenson, J. (2007). The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Communication Research*, *33*(3), 271–290. https://doi.org/10.1111/j.1468-2958.2007.00299.x

Additional resources

[1] Alonso, J. GitHub repository containing scripts for model training and timbre transfer: https://github.com/juanalonso/DDSP-singing-experiments

[2] Link to the author's GitHub repository containing examples of the audio stimuli used in the online experiment: https://github.com/christianstenbro/Did-that-flute-just-say-something.git