Categorization in the **speech to song transformation**

BAS CORNELISSEN[1]
MAKIKO SADAKATA[1,2]
HENKJAN HONING[1]

1 Music Cognition Group, Institute for Logic, Language and Computation, University of Amsterdam, The Netherlands
2 Artificial Intelligence department, Radboud University Nijmegen, The Netherlands

When repeated, speech sometimes behaves so strangely that it starts to sound as song. Using an online experiment, we constructed a large collection of stimuli, enabling further analyses of the transformation.

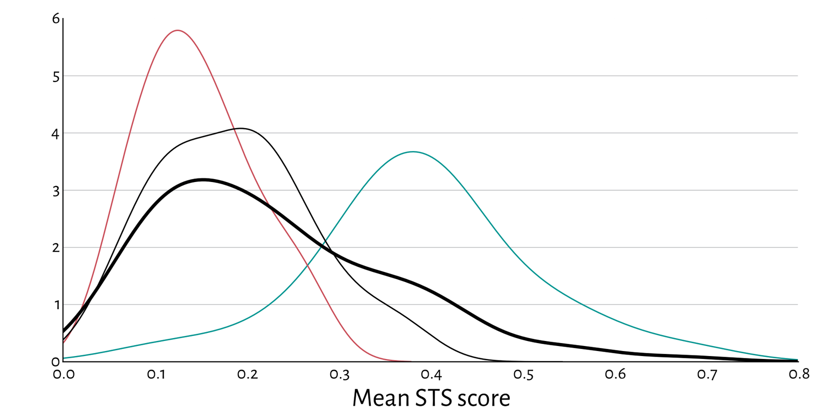

Figure 1 Estimated distribution of the STS-scores (0=speech, 1=song) for different subsets: the transforming and non-transforming stimuli from Tierney et al. (2013) and the randomly sampled stimuli. The STS-scores reproduce the distinction between transforming and non-transforming stimuli.

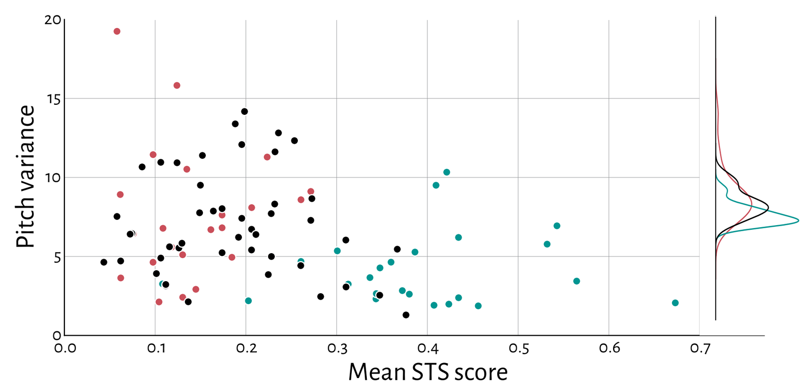

Figure 2 Pitch variance correlates with the STS score (Spearman's r= -0.24, p<0.02). Pitch variance was measured by the average standard deviation of frequencies within syllab.es.

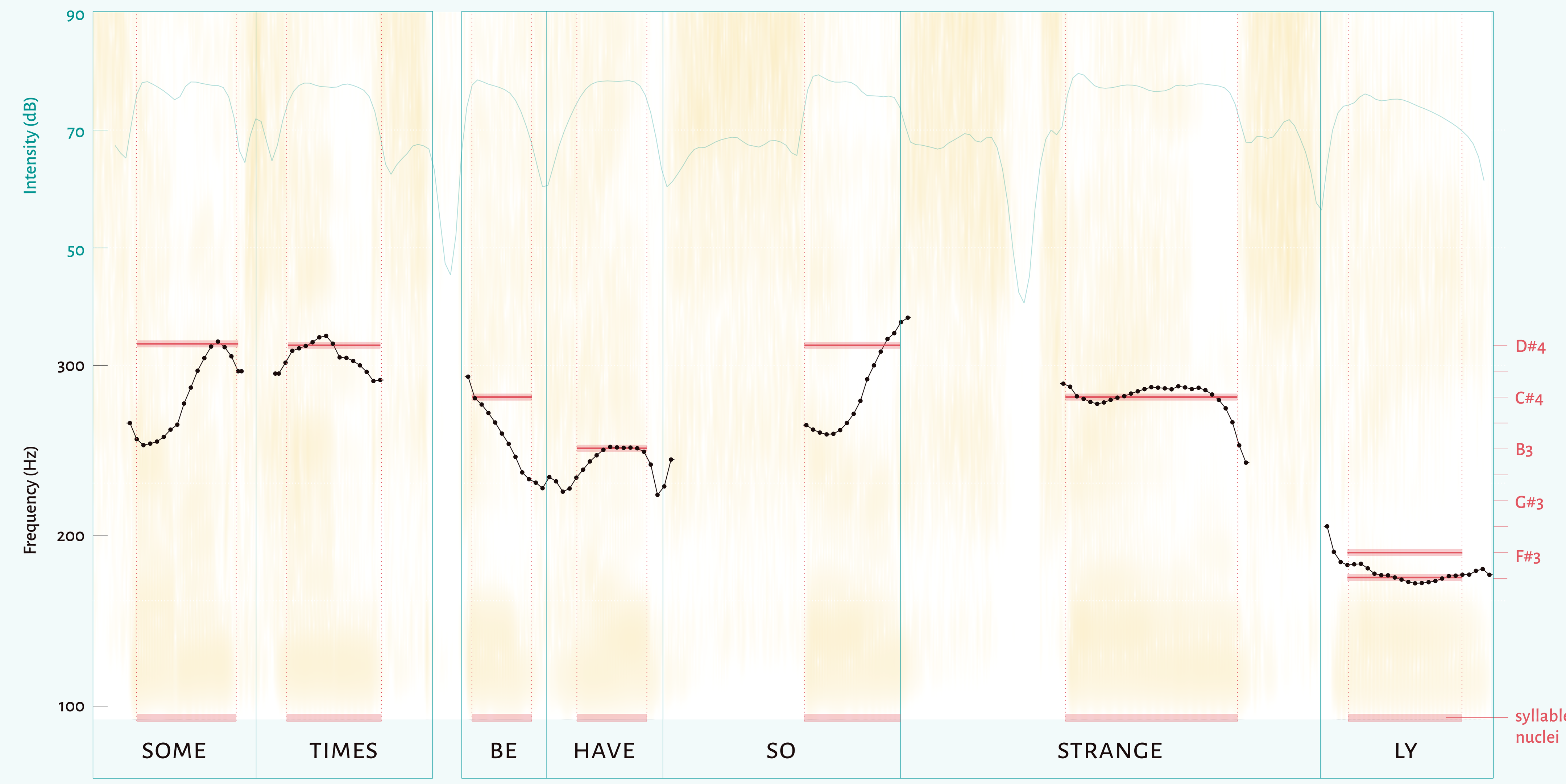

Figure 3 During the STS a melody appears in speech — but which one? This figure shows pitches and intensities for a fragment from Deutsch et al. (2011) together with two transcriptions around the nuclei of the of syllables (note the inconsistency in the last syllable). The pitch contours were extracted using Praat (Boersma, 2002).

What happens in this transformation?

The speech to song transformation (STS)occurs when a speech fragment is repeatedly presented and starts to sound as a song. Several studies seem to converge to an explanation of the STS along the following lines:[1] As with other perceptual transformations in repeated speech,[2] attention shifts to low-level prosodic features of the sound.[3] This may cause a reinterpretation of the sound, when the salience of different prosodic aspects change (notably, pitch, in the STS). In some cases the speech allows for a reinterpretation as song and a melody might emerge.

What influences the transformation?

Previous studies found effects at roughly three levels. First, at the level of the **listener**, native language,[4] suggestive instruction[5] and perhaps musical training[6] influences the STS. Second, in the **repetition** itself, changing the interstimulus pause or the repeated phrase has strong effects.[7] Finally, the prosodic features of the **repeated phrase** seem crucial. Greater pitch stability and perhaps rhythmic regularity, typical in songs, for example stimulate the transformation.[8]

ONLINE EXPERIMENT

288	137	2126
STIMULI	SUBJECTS	TRIALS
6	7.4	7
LANGUAGES	RATINGS/STIM	REPETITIONS

We randomly sampled **259 speech fragments from the audiobook library LibriVox** and added 48 fragments from Tierney et al. (2013) and 1 from Deutsch et al. (2011). The stimuli of Tierney et al. (2013) were divided into transforming and non-transforming stimuli. The fragments come from six languages: **English, Swedish, German, Spanish, Dutch** and **French**. In the online experiment, they were all rated by 161 subjects. After filtering, 137 subjects remained and 288 stimuli were rated at least 5 times, resulting in an average of 7,4 (1.2) ratings per stimulus.

During the repetitions, subjects moved a slider to match their perception. From this we calculated the **STS score** (the difference between final and initial rating) and the **STS onset** (number of repetitions before the slider was moved)

SOUNDS LIKE **SPEECH** — SOUNDS LIKE **SONG**

What have we done?

Why do certain phrases transform while others do not? The literature suggests that transforming sitmuli are in a way more 'song-like,' but sofar mainly in terms of low-level acoustic features. Moreover, most studies used a small set of stimuli with limited variation in e.g. melodic struture.[9] To allow for further analyses, of the cues influencing the categorization as song, a larger set of stimuli is required. We collected many additional stimuli and used an online experiment to rate the strength of the STS in all of them (see red box). Here we report preliminary results for a subset of 95 English stimuli.

Reliability We replicated several findings, indicating reliability of the web-based measurements. First, the transforming from Tierney et al. (2013) stimuli get higher STS-scores and earlier STS-onset than the non-transforming ones (figure 1). Second, stronger transformations happened earlier: the STS-score strongly correlates with the STS-onset.[10] Third, the pitch stability correlates with the STS-score (see figure 2)[11].

STS occurence The results indicate that the STS does not occur often: figure 1 illustrates that the scores were not bimodally distributed. The categorization between speech and song seems to be graded in the STS.

Other observations We did not find an effect of stimulus duration on the STS-score. We analyzed the vocalic structure of the stimuli using various measures (mean duration, standard devation, (relative) number of vocalic intervals, and nPVI), but found no effects.

What has to be done?

The collected dataset allows for many further analyses of the STS. One possiblity is looking at higher-level musical features such as melodic structure (see green box).

FINDING THE SPEECH IN SONG

There are two kinds of categorization in the STS. Besides the categorization of the stimulus as more speech- or songlike, the pitches are also categorized as discrete notes in transforming stimuli (figure 3). The former categorization is not easily captured, but latter is also far from trivial. Human annotator consistency on the 50 most transforming stimuli in our dataset was far from perfect (average difference of 0.4 semitones per syllable). Accordingly, music transcription algorithms (Mauch et al., 2015), although precise, miss many of the notes.

The categorization of pitches might also be used to model the categorization of speech and song. Treating the STS as a noisy channel, one would then aim to identify the hidden structure (melody) that generated the observed sound. Besides a melody, this would give a measure of the song-likeness of the speech signal that might be used to predict the STS. Our preliminary experiments were not yet conclusive on the feasibility of this approach.

Acknowledgements
We would like to thank all those who participated voluntarily in our online experiment. Gerben Groeneveld, Sophie van Weeren and Tiarma Witte played a crucial role in gathering the stimuli — thanks a lot! We are also grateful to Adam Tierney for sharing his material with us. Finally, the first author's visit to the conference would have been impossible without financial support from the Institute for Logic, Language and Computation at the University of Amsterdam.

UNIVERSITY OF AMSTERDAM

1 The clearest outline of these ideas can be found in Margulis (2013) and Falk et al. (2014), but Tierney et al. (2013) draw similar conclusions
2 Deutsch et al., 2011; Tierney et al., 2013. Also note the analogy with the verbal transformation effect, where the reinterpretation amounts to the shifting of word boundaries.
3 E.g. Margulis (2013), but also by Falk et al. (2014),
4 Marguls et al., 2015. Note that the STS occurs accross languages.
5 See e.g. Falk et al., 2014
6 Vanden Bosch der Nederlanden et al., 2015; but see Falk et al., 2014 and Hymers et al., 2015. Deutsch et al., 2011; Vanden Bosch der Nederlanden et al., 2015
7 Deutsch et al., 2011
8 The importance of pitch stability is very clear from Tierney et al. (2013) and Falk et al. (2014). Rhythmic cues were also reported in Falk et al. (2014).
9 The work by Tierney et al. (2013) forms an important exception, in which they used 24 transforming and 24 non-transforming stimuli. However, deciding on a categorization beforehand obscures a more graded categorization we report here and also does not allow any conclusions about the commonality of the STS.
10 Falk et al.2014
11 E.g. Tierney et al. 2013

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10), 341-345.

Deutsch, D., Henthorn, T., & Lapidis, R. (2011). Illusory transformation from speech to songa). *The Journal of the Acoustical Society of America*, 129(4), 2245-2252.

Falk, S., Rathcke, T., & Bella, S. D. (2014). When Speech Sounds Like Music. Journal of Experimental Psychology. *Human Perception and Performance*, 40(4), 1491–1506. http://doi.org/10.1037/a0036858

Hymers, M., Prendergast, G., Liu, C., Schulze, A., Young, M. L., Wastling, S. J., … Millman, R. E. (2015). Neural mechanisms underlying song and speech perception can be differentiated using an illusory percept. *NeuroImage*, 108, 225–33. http://doi.org/10.1016/j.neuroimage.2014.12.010

Margulis, E. H. (2013). Repetition and emotive communication in music versus speech. *Frontiers in Psychology*, 4(APR), 1–4. http://doi.org/10.3389/fpsyg.2013.00167

Margulis, Elizabeth H., Rhimmon Simchy-Gross, and Justin L. Black. "Pronunciation difficulty, temporal regularity, and the speech-to-song illusion." *Frontiers in psychology* 6 (2015).

Tierney, A., Dick, F., Deutsch, D., & Sereno, M. (2013). Speech versus song: multiple pitch-sensitive areas revealed by a naturally occurring musical illusion. *Cerebral Cortex*, 23(2), 249-254.

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., … Dixon, S. (2015). Computer-aided Melody Note Transcription Using the Tony Software : Accuracy and Efficiency. Proceedings of the First International Conference on Technologies for Music Notation and Representation, 8

Vanden Bosch der Nederlanden, C. M., Hannon, E. E., & Snyder, J. S. (2015). Everyday musical experience is sufficient to perceive the speech-to-song illusion. *Journal of Experimental Psychology: General*, 144(2), e43–e49. http://doi.org/10.1037/xge0000056