Sine-wave Synthesis, or Sine-wave Speech. A Short Introduction by Manfred Bartmann

On the CD, three tracks make use of sine-wave syntheses: No. 3 *April '84 - fieldworking in East Frisia,* No. 4 *An East Frisian farm-hand song (Deenstenleed) revisited,* and No. 5 *Rökeldoab Dada, a grooving Low German mouth music.* For further information on these tracks please refer to the corresponding booklet entries. Please also refer to the sound files embedded in this document. Their playback requires an applicable PDF-viewer. © Adobe Reader XI will do.

The pure-tone whistles within these tracks may remind one of warbling sounds for which R2-D2, an astromech droid popularised by the Star Wars movies, has become famous.

... please note that we did not hire R2-D2 as a studio musician !

R2-D2's iconic sounds were created by sound designer Ben Burtt, and I always had thought that he had used a software developed by Philip Rubin way back then for the purpose of what was to be called sine-wave synthesis (Rubin 1982). However, in an interview still to be found on YouTube (search keyword "Ben Burtt R2D2"), Burtt pointed out that he just imitated the sounds that an infant would make. Nevertheless by combining his voice with electronic sounds, Burtt managed to establish the impression that a robot was able to convey emotion via its electro-acoustic system, even without using words. Nonetheless R2-D2's somehow whistled utterings have always reminded me of speech, the more so after having studied Bregman's and Ahad's demonstrations of auditory scene analysis (CD Bregman/Ahad 1996, Demonstration 23: Sine-wave speech). From Bregman and Ahad I learnt that the perception of speech is very resilient. As a result, speech signals can be reduced dramatically without affecting their intelligibility.

This doesn't mean that any chosen sound signal should not bear at least some very special characteristics to be interpreted as speech. As a rule, they are a little bit more abstract than those Remez calls "traditional speech cues" (Remez et al. 1981), because these characteristics can be traced back to one's language acquisition. This goes with the fact that speech signals come with mayor energy bands due to acoustic resonances of the human vocal tract. They are called formants, most of them produced by tube and chamber resonance. Extensive research has proved evidence that formants not only make up the timbre of vowels, but also the timbres of musical instruments (Mertens 1975, Reuter 1997; Reuter/Auhagen 2014). The story goes like this: if an auditory scene were a wild and open ocean, hard to cope with and even harder to analyse, formants would play the role of wooden flotsam to cling to for the sake of a proper auditory scene analysis (Bregman 1990). Formants show as peaks in the amplitude/frequency spectrum. As such they are stable, reliable and even independent from fundamental pitches. Formants bear the information that humans require to distinguish between speech sounds or even between musical instruments.

No. 3 *April '84 - fieldworking in East Frisia*

Using the recording of me singing this song, we decided to call attention to some of the sung phrases by adding traces of the first 3 formants. They appear as certain harmonics that have been isolated from the vocal sound. **Fig. 1** shows a spectrogram, i.e. a sound image that shows how the spectral density varies with time (horizontal). Degrees of amplitude are shown at various frequencies on a vertical axis. In **Fig. 1** pink stands for little energy, green followed by yellow for more and more energy. The spectrograms are plotted with a logarithmic vertical frequency-axis. Formants are traces of acoustic resonances, but as they have been measured here as amplitude peaks in the frequency spectrum of the sound, they appear here as mere sine-waves. I extracted them from the audio signal using the Dutch software 'PRAAT - doing phonetics by computer' (APP Boersma/ Weenink 2015), after having studied Rubin (1982).

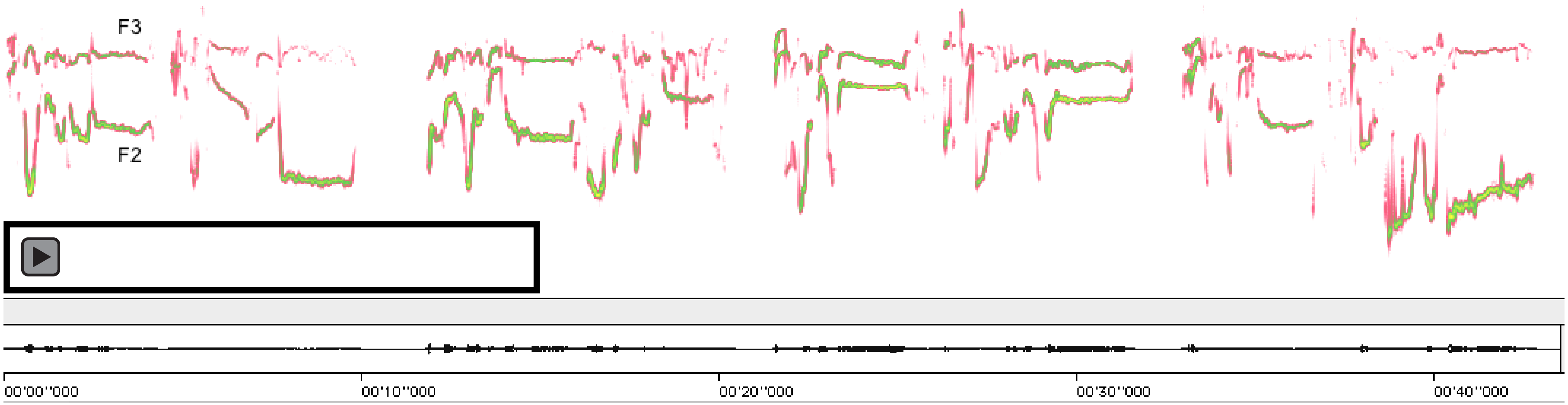
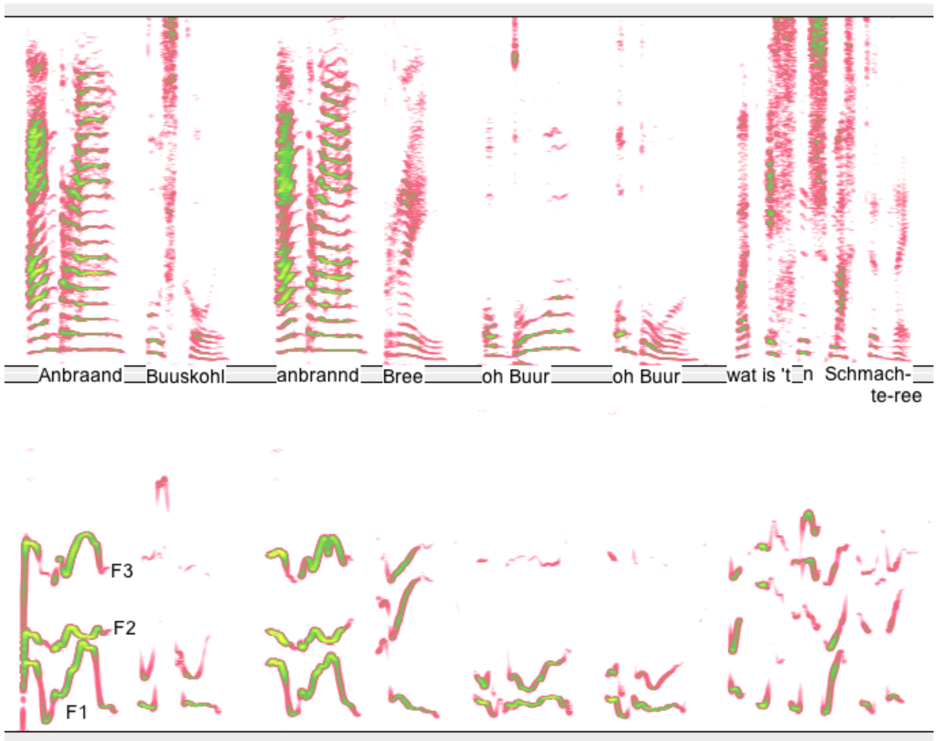


Fig. 1: Traces of the second (F2) and the third (F3) formant extracted from the recorded vocal track. While listening to the embedded sound file, note that the pure whistle tones are in harmony. This is because they stem from the very same vocal track and the human voice organ produces a series of harmonic overtones. As their amplitudes decrease monotonically with frequency, F3 shows mostly weaker spectral peaks than F2. © Manfred Bartmann 2017.

No. 4 ***An East Frisian farm-hand song (Deenstenleed) revisited***



For CD-track No. 4 we recorded a saying in Low German. It is linked to a farm-hand's protest song. The saying could be translated as

Burnt white cabbage, burnt buttermilk pudding - oh farmer, oh farmer, what a starvation ! In Low German it says : *Anbrannd Busskohl, anbrannd Bree, oh Buur, oh Buur, wat is't 'n Schmachteree.*

Fig. 2: **T**wo spectrograms both with a linear vertical frequency-axis. *Above:* Spectrogram of Manfred Bartmann's recitation of the Low German saying "Anbrannd Buuskohl". See the text for further details. *Below:* The speech signal reduced to traces of the first three formants. © Manfred Bartmann 2017.

The first 32 seconds of track No. 4 (not shown as a spectrogram here) contain what in psychoacoustics are called A/B testings: At the beginning the traces of the first formant F1 are made audible, immediately giving the impression that this signal has to do with speech. When the traces of the formants F1 and F2 are played back simultaneously one gets the impression that the signal stems from a human voice. After that, the unencoded phrase can be heard. Finally, the three formants F1, F2, and F3 are played back simultaneously, mirroring the unencoded phrase in an even sharper way. **Fig. 2** is meant to illustrate how this idea evolved.

After all, sine-wave replicas of audio recorded utterances provide evidence that an estimate of the vocal tract resonances is just enough to evoke a fairly good speech intelligibility, the more so if the listener has been told beforehand what to listen for. Having learnt this and then going back to the beginning of track No. 4, the speech will immediately pop out when listening to the first signal - which has nothing but the traces of F1. All in all, a so-called schema-driven grouping will always overrule any primitive auditory scene analysis (Bregman 1990). It simply depends on the schema that has been suggested.

No. 5 ***Rökeldoab Dada, a grooving Low German mouth music***

Fig. 3 contains a series of spectrograms which represents an early stage of exploration of how to make music of speech signals and corresponding drum signals (Yoder/Bartmann 2017). Please note that different spectrograms may represent very similar sounds, as well as similar spectrograms may refer to different sounds. **Fig. 3** displays explorations of sounds for the purpose of comparing what one sees with what one hears. They are presented as A/B testings, that being a method to test a listener's response to version A against version B, while version B is always modified in some respect. In other words: a perceptual set is put forward as a basis or as a predisposition for speech perception, the more so to be draw attention to any modification. **Fig. 3** can be seen as a first attempt to exploit psychoacoustic A/B testings in an artistic manner.

See Bailes/Dean 2009; Best/Morrongiello/Robson 1981; APP Boersma/ Weenink 2015; Bregman 1990; CD Bregman/Darwin 2008, Demonstration 23: Sine-wave speech; Mertens 1975; Remez 1981, 1994, 2005; Remez/Fellowes/Rubin/Talker 1997; Remez/Pardo/Piorkowski/Rubin 2001; Reuter 1997; Reuter/Auhagen 2014; Rubin 1982; Pardo/Remez 2006; Sundberg 1987; Yoder/Bartmann 2017.

Some search keywords for Wikipedia: A/B testing, Ben Burtt, Formant, Human voice, Music Cognition, Psychoacoustics, Sinewave synthesis.

Some search keywords for YouTube: Ben Burtt R2D2, Human voice, Sine Wave Speech, sinespeech.wmv, Sine-Wave Speech Demonstration - Disinformation 2014.

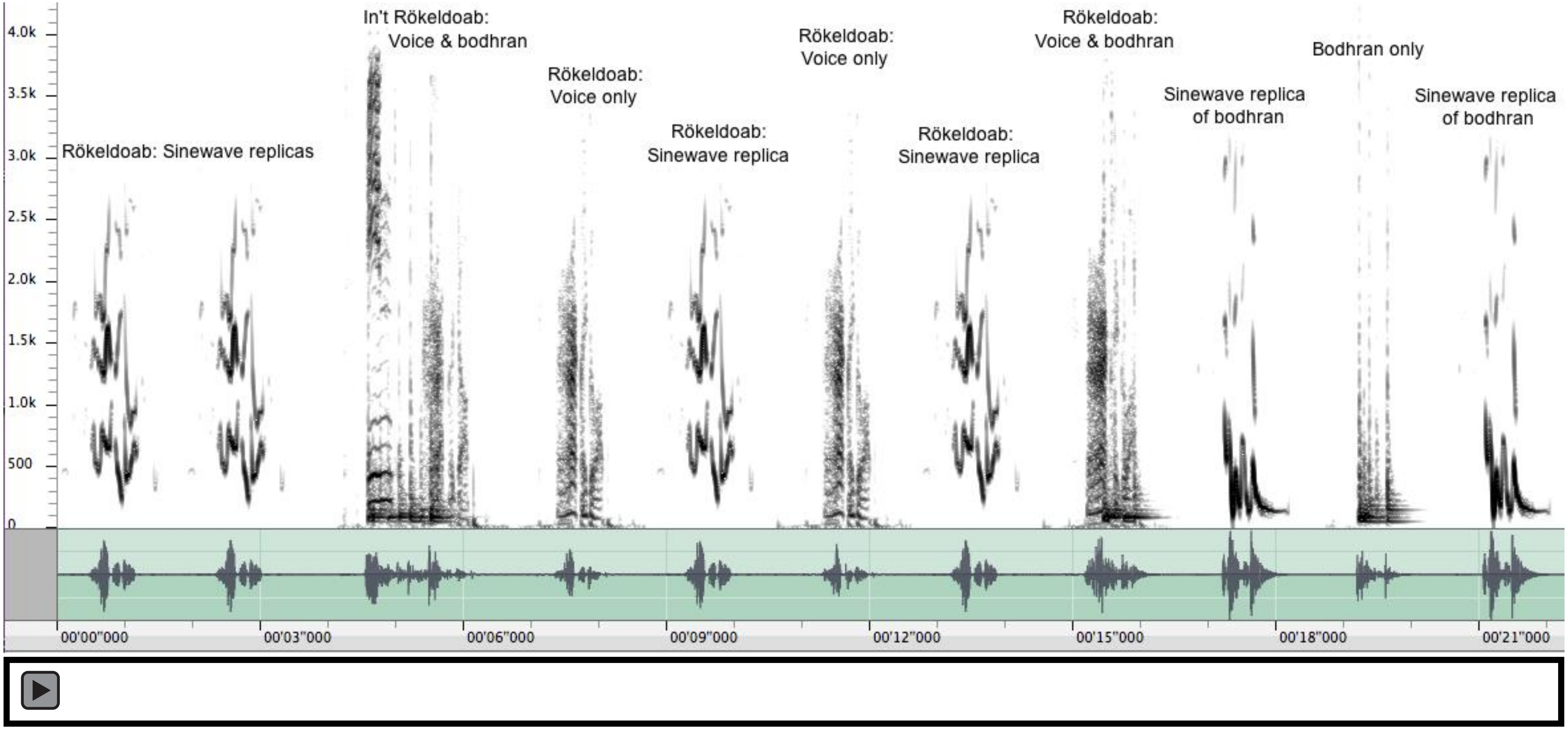


Fig. 3: Rökeldoab. Making music of a speech signal and a drum signal, first having extracted traces of formants as sine waves, then using A/B testings. For further information please see Yoder/Bartmann (2017). © Manfred Bartmann 2017.