ON THE SIMILARITY OF TONES OF THE ORGAN STOP *VOX HUMANA* TO HUMAN VOWELS[1]

Fabian Brackhane[2] and Jürgen Trouvain[3]
[2]Institut für Deutsche Sprache (IDS), Mannheim, Germany, [3]Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany
e-mail: [2]brackhane@ids-mannheim.de, [3]trouvain@coli.uni-saarland.de

Abstract

In mechanical speech synthesis from the 18th up to the 20th century, reed pipes were mainly used for the generation of the voice and the organ stop vox humana was central in this process. This has been described in different historical documents which report that the vox humana in some organs sounded like human vowels. In this study, tones of four different voces humanae were recorded to investigate their similarity to human vowels. The acoustical and perceptual analysis revealed that some, though not all, tones show a high similarity to selected vowels.

1 Introduction

Many authors of the 18th and 19th century consider the organ stop *vox humana* as the prototype for a mechanical speech synthesiser or, more specifically, as the prototype for a vowel synthesiser. In this view, the task would be to develop the vowel-like features of the *vox humana* to a "speech organ" as Euler (1773: 246) suggested.

However, evidence for a real similarity to vowels is either missing or does not hold up under today's standards. Based on personal experience, the resemblance of the sound of modern and historical *voces humanae* and human vowels does not seem to be very close. For this reason, we performed a study including an acoustic analysis, as well as perception tests, to verify the historical descriptions of the *vox humana* and its similarity to human vowels.

2 The mechanism and use of the organ stop vox humana

The organ stop *vox humana*, consisting of reed pipes, has been described since the middle of the 16th century (Eberlein, 2007: 817). An organ stop is a set of organ pipes with different pitches but constructed in the same way. It can be switched "on", i.e., admitting the pressurised air to the pipes of this stop, or "off", i.e., stopping the air. Organs usually have multiple stops (often between 25 and 30 and not all of their pipes are visible from the outside). The majority of the stops are flue

[1] A shorter version of this article was published under the title “The organ stop 'vox humana' as a model for a vowel synthesizer” in the proceedings of the 14th Interspeech (Lyon) 2013, pp. 3172-3176.

pipes (see Fig. 1 bottom), although reed pipes are also common (see Fig. 1 top). A characteristic feature of the reed pipes used in a *vox humana* is the *resonator* that is of a relatively constant size independent of the pitch of the pipe. This means that there are possibly slight differences with respect to the size of the *resonators* because each pipe of a given *vox humana* stop is hand-made. For almost every other organ stop consisting of reed pipes, the length of the *resonator* decreases successively with the increasing pitch of the pipes. In case of the *vox humana*, the *resonators* act as a filter in such a way that formants can be observed that are similar to those found in human vowels (Lottermoser, 1936: 48; Lottermoser, 1983: 135).

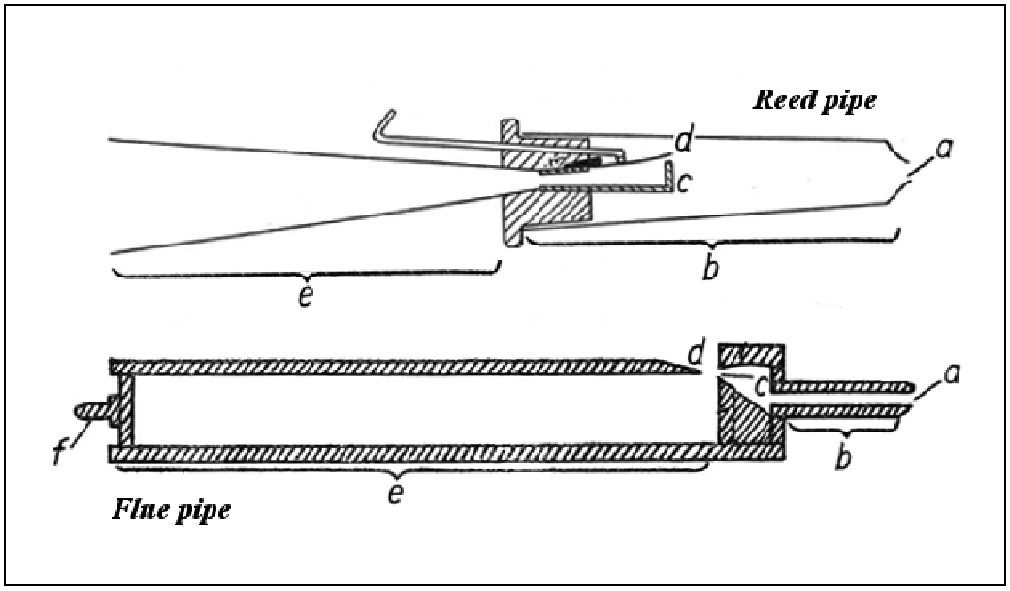

Figure 1. Schematic drawing of a reed pipe (top, re-drawn after Lottermoser 1936: 15) and a flue pipe of the type stopped diapason (bottom, redrawn after Adelung 1982: 43). The air flows into the pipes (a) passing the socket or boot (b). The air in the reed pipe (top) will be excited by the reed tongue (d) that lies on the shallot (c). The excitation of the air in the flue pipe (bottom) is possible by an increased air pressure at the windway (c) and the continuation towards the upper lip (d). The resonator (top e) and the body (bottom e) act as acoustic filters. (f) represents the cap needed for stopped flue pipes.

The term *vox humana* originates from the use of an organ reed stop with proportionally short resonators which substitute for the human singing voice. For this reason, it was never used solo but was usually played together with the so-called *tremulant* and the stopped flue stop *bourdon* (also called *stopped diapason*) of the same pitch. The *tremulant* changes the pressure of the air streaming to the pipes in brief intervals. The resulting sound, which resembles the vibrato of a human singing voice, has been named *vox humana*. Thus, the organ stop *vox humana* has been used

as a substitute for the human singing voice, however, it was not considered to be an *imitation* of the human voice.

However, knowledge about the original meaning of the term *vox humana* has been lost over time and was considered as an *imitation* rather than as a *subsititution*. For these historical reasons, there is not only one construction type but various ones. Nearly every organ builder of the 18th century intended to invent a really natural sounding *vox humana*. Thus, the name *vox humana* can be considered as a programmatic title rather than as a technical term. Numerous historical documents attested that these pipes clearly sounded like vowels (e.g. Greß, 2007: 27).

This new usage made organ builders (e.g. Joseph Gabler), as well as researchers such as Leonhard Euler (1707-1783) and Christian Gottlieb Kratzenstein (1723-1795), to consider the *vox humana* as the prototype of speech synthesis.

3 Recordings and acoustic analysis of various *voces humanae*

3.1 Data

It was our aim to test the historical statements concerning the similarity of the *vox humana* sound to those of human vowels. This required recordings of those organs where the stops are historically authentic (and not re-constructed). The research question was whether pipes of a *vox humana* really displayed formant structures similar to those of human vowels. More specifically, we were interested in determining whether certain vowel qualities could be recognised reliably by human listeners.

The first author recorded selected tones from the originally preserved *vox humana* stops of four different organs from the middle of the 18th century. Three of these were located in churches in the southwest Germany: Abteikirche Amorbach, Schlosskirche Meisenheim and Stadtkirche Simmern (AMO, MEI, SIM henceforth). These organs were built between 1767 and 1782 by craftsmen from the same family of organ builders (Stumm), and all three organ stops had the same construction style and sizes (see Fig. 2).

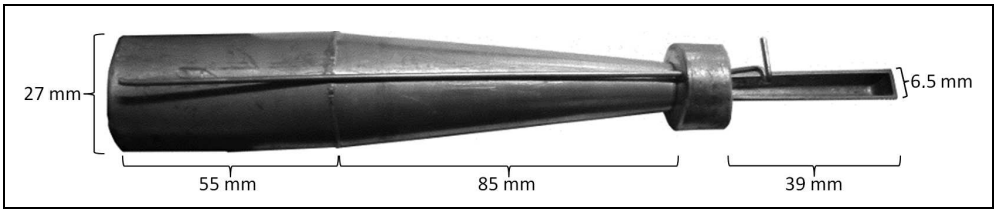

Figure 2. Reed pipe from the *vox humana* (Tone g0) from Amorbach (1782), without the boot (cp. (b) Fig. 1 top) and without the reed tongue (cp. (d) Fig. 1 top).

In addition, *the vox humana* of the organ of the Stadtkirche in Waltershausen (Thuringia, Eastern Germany) was recorded (WAL henceforth) at a later time. This organ stop is a copy of the *vox humana* from the great organ at the monastery in Weingarten (1750) which is famous because of its constructor, Joseph Gabler, who

attempted to build pipes with a sound that resembled human singing voices in a particular way. The *resonators* of these pipes were adapted to human larynges.

The tones C, G, c0, g0, c1, g1, c2, g2 and c3 (in an alternative notation C, G, c, g, c', g', c'', g'', c''') were recorded from the *voces humanae* of all four organs (9 tones * 4 organs = 36 recordings in total). In SIM, we also recorded the historical (i.e., not reconstructed) reed pipe stops *trumpet* and *crumhorn* (for C and g0). These two stops substantially differ from the *vox humana* in their construction styles and they were recorded for comparison with the *vox humana* of the same organ and the measurements found in Lottermoser (1983). Only two tones were selected: C as the lowest one and g0 because it has been described as particularly vowel-like (see e.g., Frotscher 1927: 54). Thus, the total number of recorded tones increased to 40.

The tones in AMO, MEI and SIM were played solo for the recordings, i.e., as pure tones and for this reason without the additional stops *stopped diapason* and *tremulant,* which are typically used in musical tradition. The *vox humana* in WAL could not be played solo for technical reasons; consequently the tones here were played in combination with the flue pipes of the *stopped diapason*, but without the *tremulant*.[2]

The microphone was placed at a distance of about half a metre above the resonators to produce comparable recordings in the acoustically different churches and to reduce the echo and filter effects of the rooms as much as possible (although the influence of the acoustic conditions of the churches can never be completely excluded). All recorded tones were about 5 seconds in duration. This length is due to the fact that the reed pipes need a relatively long time to reach the stationary phase.

The acoustic analysis of the data included the measurement of F_0 and the first three formants. For each 5-sec tone, the first and last 5% of the duration were ignored and from the remainder of the tone, 10 equidistant values were taken. The analysis was performed with the phonetic standard freeware Praat (version 5.3.19).

3.2 Results

The values for the fundamental frequency show that all four organs differed in their F_0 for virtually all tones (see Table 1). For example, the tone G, comparable to a bass voice, ranged from 98 Hz in AMO to 105 Hz in MEI. In the following sections, only the results for SIM and WAL are reported due to a high level of comparability of the stops in AMO, MEI and SIM.

The spectra of all *voces humanae* tones showed clear formant structures. This is also true for the additional stops; *trumpet* and *crumhorn* (see Fig. 3). However, the formant shapes of the *voces humanae* showed more similarity to the formants of typical human speech. Interestingly, in all four organs, the values for F_0 and F_1

[2] Unfortunately, the recordings of the three Stumm organs in AMO, SIM and MEI were already finished when we surprisingly had the opportunity to record the organ in WAL. Thus, the recordings from WAL are not fully comparable to the recordings of the other three *voces humanae.*

converged or even merged for the two and sometimes three highest tones, which made a visible distinction nearly impossible.

Table 1. F_0 values in Hz of all tones of all voces humanae.

Tone	AMO	MEI	SIM	WAL
C	66	70	69	69
G	98	105	102	103
c0	132	141	136	139
g0	197	210	205	208
c1	263	281	274	277
g1	395	411	408	415
c2	527	562	548	554
g2	790	844	818	832
c3	1054	1124	1093	1108

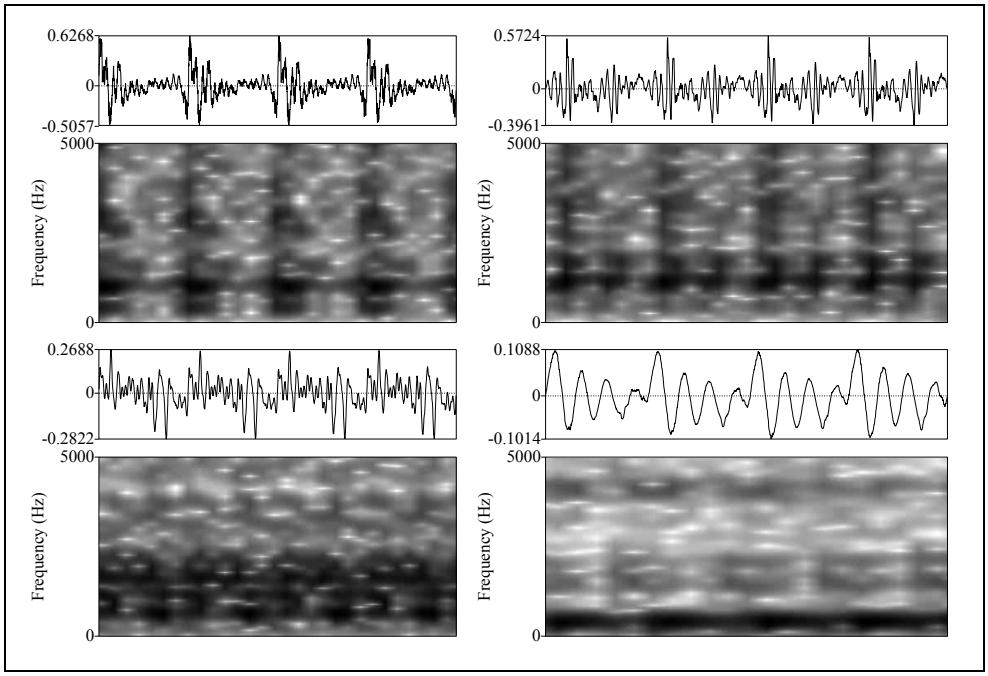

Figure 3. Waveforms and spectrograms of sections with four periods taken from the tone C of the stops in SIM: *vox humana* (top left), crumhorn (top right) and trumpet (down left) (duration: 60 ms) and from the vowel /ø/ of a male German speaker (down right; duration: 42 ms, F_0: 97 Hz).[3]

[3] Recordings of the tone G which would be more comparable to the human voice were not available for all stops.

In Figure 4 (a-c), the spectral distribution of the *vox humana* from SIM (from Fig. 3) and WAL are compared with the spectrum of the human vowel that is also shown in Figure 3. The decline of the spectral slope was much more intense for the human vowel than for the organ-generated tone.

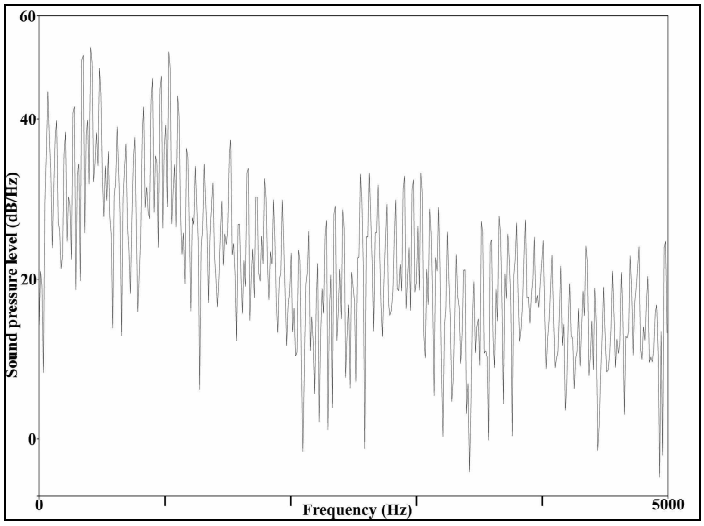

Figure 4a. Spectrum for the middle part of tone C of the stop vox humana from SIM from 0 to 5 kHz.

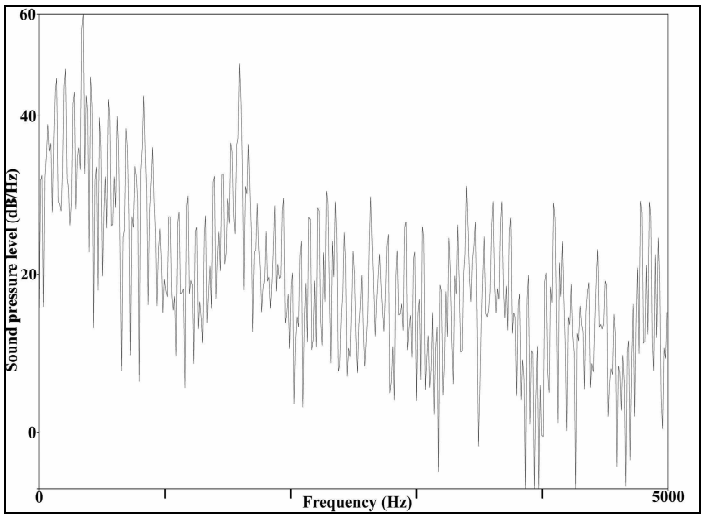

Figure 4b. Spectrum for the middle part of tone C of the stop vox humana from WAL from 0 to 5 kHz.

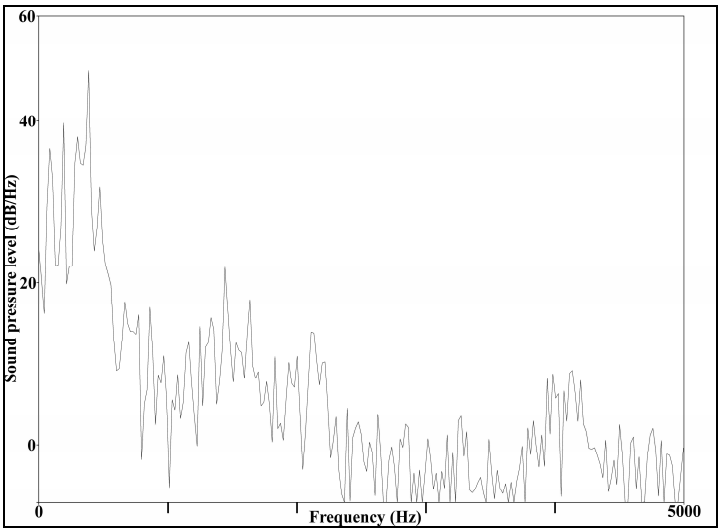

Figure 4c. Spectrum for the middle part of vowel /ø/ of a human male voice from 0 to 5 kHz.

But there are also differences in the spectral distributions of the *voces humanae*. Figure 4 (a and b) displays the harmonic distribution of energy for the C tones of the *voces humanae* in SIM and WAL, with the latter having a much higher level of intensity.

Figure 5 (a and b) displays the locations of F_1, F_2 and F_3 for the *voces humanae* of SIM and WAL. One can see that the formant distribution of the WAL tones mainly reflected changes in F_1 (from 400 to 1300 Hz), whereas the tones from the SIM organ showed a larger variation in F_2. Compared to the formant space of human (male) voices (German speakers producing long, tense vowels, taken from Simpson, 1998), both organs generated a smaller vowel space. In addition, the organs' vowel spaces had higher average formant values than the human vowel space. This formant shift is illustrated in the very small overlap of the spaces for the SIM *vox humana* and the human voice.

Inspection of the F_3 values revealed a much wider formant range for the organs compared to a male voice. For instance, F_3 of the SIM organ ranged between 1900 and 2800 Hz, WAL between 2000 and 3000 Hz, whereas the F_3 of the human voice ranges between 2200 and 2500 Hz.

For two tones, c1 from MEI and SIM, respectively, maximal energy was found on the 7^{th} harmonic (at around 1970 Hz). This is in line with a previous study by Lottermoser (1983: 135) on the acoustics of reed pipes for the tone C. However, the maximal energy of all other tones from AMO, MEI and SIM were irregularly distributed on other harmonics. The tones for WAL could not be considered because the additional labial pipes changed the energy distribution in a substantial way (cp. the differences in the harmonic distribution in Fig. 4b).

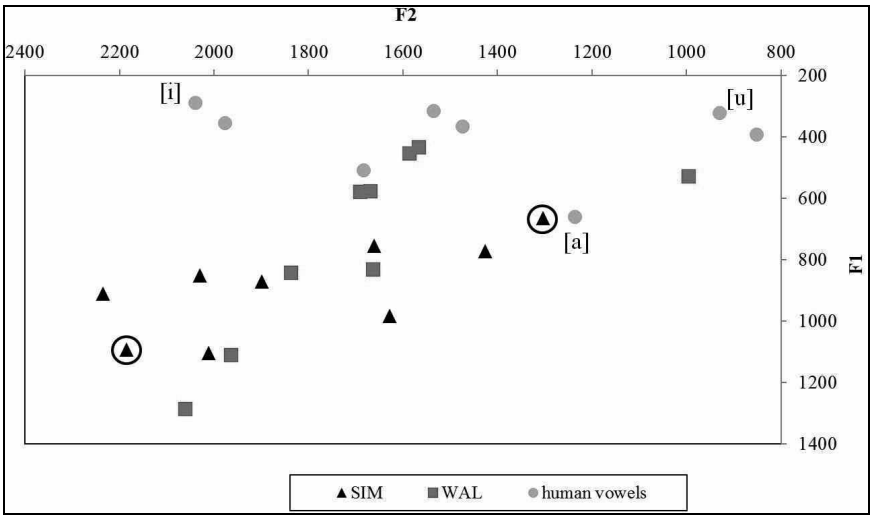

Figure 5a: Values for F_1 and F_2 of the tones of the *vox humana* in SIM (black triangles) and WAL (red squares) as well as standard values for the German long vowels of male voices (Simpson, 1998) (green dots). The encircled triangles indicate well recognised vowel qualities: [i] on the left, [ø] on the right.

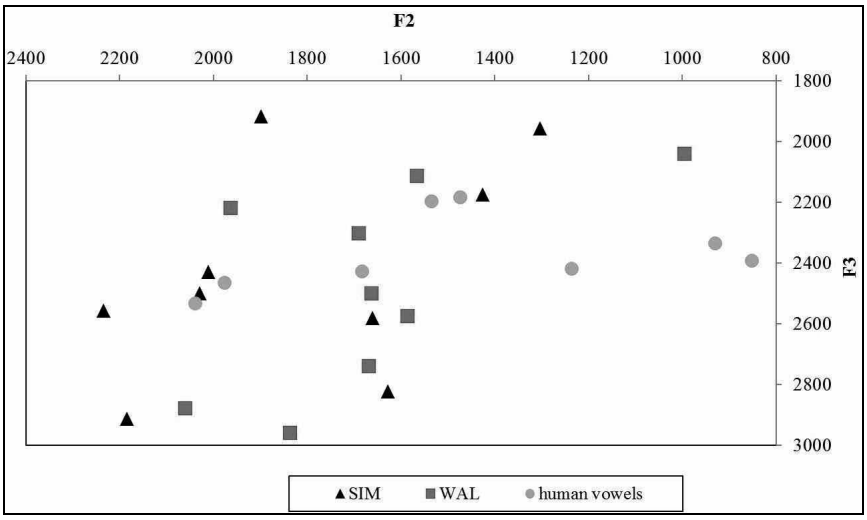

Figure 5b. Values for F_2 and F_3 of the tones of the *vox humana* in SIM (black triangles) and WAL (red squares) as well as standard values for the German long vowels of male voices (green dots) after Simpson (1998).

3.3 Discussion

The differences in fundamental frequencies for the same tone across organs can be explained by the fact that in the 18th century, the fundamental frequency had not yet been standardised with a fixed value (in contrast to today). Thus, the tuning of the tones could vary according to the region and to the size of the organ.

These data suggest that a central feature of the reed pipes from different *voces humanae* is a spectral distribution with a clear ″formant-like″ structure, illustrated by differentiated frequency bands of higher intensity. Formants could also be found for the reed stops *crumhorn* and *trumpet* (cp. Fig. 3), however, the distribution of these formants seemed to be less similar to those of the human voice. It is unclear whether further stops, especially flue stops, which account for two thirds of all pipes in a typical organ, also show formants[4]. Moreover, it is unclear which reed pipes show the largest similarity to the formants of human vowels.

The formant values of the *voces humanae* produced a vowel space that was smaller in size and with more upshifted formant values in comparison to a human speaking voice (cp. Fig. 5). This possibly could be explained by the smaller "vocal tract" of the investigated *voces humanae* in comparison to a human vocal tract which is given in text books with 17.0 cm x 4.5 cm (e.g., Pompino-Marschall, 2009: 160). This size is also in contrast to measures of the *voces humanae* from organs built by Stumm with pipe sizes of 14.0 cm x 2.7 cm for the tone g0. Fitch and Giedd (1999) reported average values of vocal tract length (based on MRI data) for young male adults (aged 19 to 25) of only 15.0 cm, whereas the values for 13-16 year old male children were 14.0 cm. The latter correlates with the length of the resonator of a *vox humana* pipe. This agreement is interesting when considering that the use of the *vox humana* stop once was a substitution for church boys' choirs.

The one-directional variation of the vowel space in the formant plane can also be explained with the resonator of the *vox humana* as a single-opened conic tube without any constrictions. It is usually assumed that vowels produced in a human vocal tract need two cavities, a back cavity and a front cavity. For a single cavity, as in the *vox humana* pipe, the higher formants should just be multiples of the first formant. However, this is not exactly the case when we compare the formant values in Table 2.

Future experiments with formant synthesis could show whether the measured formants from the organs can generate acoustic patterns which sound like humanoid vowels to the listener. Formant synthesis could also be used for experimentation with spectral tilt, which is reduced in the *vox humana* compared to a human voice. One explanation for this reduction in spectral tilt is the absorbing characteristics of the human oral cavity, which are not found in metal organ pipes.

The spectral distribution of the *vox humana* is partially different from that of human vowels. As already described in Lottermoser (1983: 135), the maximal energy of the highest tones can be found in the 7th harmonic of the solo played voces humanae. On the other hand, the fundamental frequency was hardly ever found to be the strongest harmonic (2 exceptions out of 27 tokens from AMO, MEI and SIM). The *vox humana* in WAL was played in combination with another stop which, in this case, caused the fundamental frequency to be the strongest harmonic.

[4] It is planned for future studies to record flue pipes as well. This would allow a comparison to reed pipes with respect to formant structures.

Table 2: Values of all tones for SIM and WAL for (measured) F_1, doubled F_1, (measured) F_2, three times F_1 and (measured) F_3.

	SIM					WAL				
Tone	**F_1**	**F_1*2**	**F_2**	**F_1*3**	**F_3**	**F_1**	**F_1*2**	**F_2**	**F_1*3**	**F_3**
C	982	1964	1628	2946	2823	453	906	1587	1359	2573
G	664	1328	1304	1992	1957	529	1058	996	1587	2040
c0	773	1546	1426	2319	2174	576	1152	1670	1728	2739
g0	754	1508	1661	2262	2580	434	868	1567	1302	2114
c1	871	1742	1899	2613	1919	843	1686	1838	2529	2959
g1	852	1704	2029	2556	2500	1286	1572	2061	3858	2878
c2	1104	2208	2011	3312	2430	578	1156	1690	1734	2301
g2	911	1822	2234	2733	2557	831	1662	1663	2493	2499
c3	1092	2184	2185	3276	2912	1110	2220	1964	3330	2218

4 Perception tests

The aim of the perception tests was to find out whether listeners could reliably associate the recorded tones to vowel categories. If so, it would be interesting to know more about the underlying nature of these perceptual impressions.

Two listening tests were performed. The first test could be seen as a pilot test, whereas the second test was a repetition of the first one, with substantial improvements. Since both tests were very similar, they are presented together.

4.1 Method

There were 40 stimuli for the first test consisting of the 36 *vox humana* tones plus the four tones from the stops *crumhorn* and *trumpet*. Each stimulus had a duration of 5 seconds. Twenty German linguists served as participants. The stimuli were presented via headphones in a randomised order and could be played as often as the participant wished. The participants were asked to indicate the vowel quality of each stimulus, if possible in terms of IPA cardinal vowels. There was also the option to say "no vowel". The answers were given in spoken form directly to the experimenter.

The second test was similar to the first one but with some modifications. This time, the experiment was performed using a web-based platform for the perception tests (with the help of Draxler, 2011) in order to test more participants (with German as their first language). In total, there were 29 participants, including linguists and non-linguists. The number of stimuli was reduced to 18 (using the *voces humanae* in SIM and WAL), plus the four tones from the stops *trumpet* and *crumhorn*. Each stimulus occurred three times, resulting in 66 stimuli presented in randomised order. Since the *voces humanae* from SIM and WAL showed the most contrasting results in the first test, these were selected for the second test. Each stimulus was shortened to 400 ms (taken from the middle part) in order to make it comparable to a long vowel in German. The vowel categories in the second test were the letters representing all long, tense vowels in German: I, Ü, E, Ö, Ä, A, O, U, which represent the vowels /i, y, e, ø, ɛ, a, o, u/. The first test revealed that only three out of twenty participants were able to use the IPA system, so letters were used to permit more consistent answers. The answer "no vowel" was not possible this time. For

technical reasons, one stimulus was not correctly played (c0 from WAL). Consequently, the corresponding results will not be presented.

4.2 Results

The results (see Table 3) for both *voces humanae* clearly indicated the correlation between the fundamental frequency and the vowel category. In other words, the higher the F_0, the more /i/-like the selected vowel and the lower the F_0, the more /o/-like the vowel.

The tones at the periphery (in terms of F_0, as well as F_1, F_2 and F_3) were assessed more consistently than those in the middle region. This is obvious for instance for the SIM tones in the second experiment, which revealed a stable c3 for /i/ (84%), but far less consistency for the next lower tone, g2 (between /i/ and /e/, with a tendency to /i/). The tone c1 is more or less equally distributed between the qualities of /e/, /ɛ/, /ø/ and /a/. For the corresponding tone of WAL, the listeners largely preferred /a, ɑ/ (test 1) or even /u/ (test 2).

The tones from the comparative stops *crumhorn* and *trumpet* showed less consistent answers than those of the *voces humanae,* especially for the tone g0. Comparing the results of the organs of SIM and WAL, it was evident that the tone-vowel correspondences of SIM showed a higher level of consistency than those of WAL (except for C and the maverick answer for c1).

The general tendencies of the first perception test were confirmed by the second, but on a more reliable basis. The results were sometimes clearer (e.g., for c0 and g1 in SIM) and often led to a higher level of consistency for SIM as well as for WAL.

Table 3. Percentages of answers for the stimulus tones of SIM and WAL for both perception experiments. The values for F_0 and the formants are in Hz. The stops were voces humanae (VH), crumhorn (CR) and trumpet (TR). Vowel categories in experiment 1 were clustered according to the German vowel letters. The most frequent answer for each tone is given in bold. Grey-shading of cells according to numbers: 100-80% (darkest grey), 79-60%, 59-40%, 39-20% (lightest grey), 19-0% (no shading).

Experiment 1																	Experiment 2								
no V	i	y	e	ɛ	ø/œ	a/ɑ	o/ɔ	u	Σ	stop	tone	F_0	F_1	F_2	F_3	i	ɑ	e	ɛ	ɔ	a	o	u	Σ
40	**0**	**0**	**0**	**0**	25	10	25	**0**	**100**	VH Simmern	C	**69**	**982**	**1628**	**2823**	**0**	**2**	**2**	**8**	31	**7**	38	**11**	**100**
5	**0**	**0**	**0**	**10**	**80**	**0**	**0**	**5**	**100**		G	**102**	**664**	**1304**	**1957**	**0**	1	**2**	**8**	**85**	**2**	**1**	**0**	**100**
5	**0**	**0**	**5**	35	35	**10**	**10**	**0**	**100**		c^0	**136**	**773**	**1426**	**2174**	**0**	1	14	15	64	**3**	**2**	**0**	**100**
20	**0**	**0**	20	35	**10**	**10**	**5**	**0**	**100**		g^0	**205**	**754**	**1661**	**2580**	**0**	**0**	15	**36**	41	**7**	**1**	**0**	**100**
15	**0**	**0**	**15**	25	**15**	20	**5**	**5**	**100**		c^1	**274**	**871**	**1899**	**1919**	**2**	**2**	20	21	29	20	**3**	**3**	**100**
20	**0**	**0**	35	**20**	**0**	20	**5**	**0**	**100**		g^1	**408**	**852**	**2029**	**2500**	**5**	**6**	**59**	**16**	**6**	**6**	**0**	**3**	**100**
25	**5**	**5**	20	**5**	**0**	35	**0**	**5**	**100**		c^2	**548**	**1104**	**2011**	**2430**	**15**	**13**	37	**2**	**8**	23	**1**	**1**	**100**
30	50	**0**	**10**	**0**	**0**	**5**	**0**	**5**	**100**		g^2	**818**	**911**	**2234**	**2557**	51	**7**	25	**2**	**3**	**7**	**1**	**3**	**100**
15	**80**	**0**	**0**	**0**	**0**	**5**	**0**	**0**	**100**		c^3	**1093**	**1092**	**2185**	**2912**	**84**	**9**	**3**	**0**	**0**	**3**	**0**	**0**	**100**
15	**0**	**0**	**0**	**5**	40	**15**	**20**	**5**	**100**	VH Walsenhausen	C	**69**	**453**	**1587**	**2573**	**0**	1	**1**	**2**	33	1	40	**21**	**100**
10	**0**	**0**	**0**	**0**	20	30	35	**5**	**100**		G	**103**	**529**	**996**	**2040**	**1**	1	**6**	**3**	40	**9**	32	**7**	**100**
15	**0**	**5**	**10**	**0**	65	**0**	**0**	**5**	**100**		g^0	**208**	**434**	**1567**	**2114**	**1**	26	**15**	1	32	**3**	**9**	**11**	**100**
10	**0**	**5**	**10**	**5**	**10**	35	**5**	**20**	**100**		c^1	**277**	**843**	**1838**	**2959**	**2**	**9**	**10**	**2**	**10**	**5**	**5**	56	**100**
25	**10**	**15**	20	**5**	**5**	20	**0**	**0**	**100**		g^1	**415**	**1286**	**2061**	**2879**	**9**	29	28	**2**	**13**	**6**	**5**	**9**	**100**
35	20	35	**5**	**0**	**0**	**5**	**0**	**0**	**100**		c^2	**554**	**578**	**1690**	**2301**	33	21	28	1	**7**	**2**	**0**	**8**	**100**
35	**15**	35	**0**	**0**	**0**	**5**	**0**	**10**	**100**		g^2	**832**	**831**	**1663**	**2499**	39	22	**3**	1	1	**8**	**3**	22	**100**
30	35	**5**	**0**	**5**	**5**	15	**0**	**5**	**100**		c^3	**1109**	**1110**	**1964**	**2218**	**75**	11	**2**	1	**0**	**9**	**0**	**1**	**100**
30	**0**	**0**	**0**	**5**	**5**	40	15	**5**	**100**	CR	C	**69**	**1097**	**1688**	**3014**	**0**	**0**	**3**	13	28	30	**17**	**9**	**100**
10	**0**	**0**	**0**	15	35	40	**0**	**0**	**100**		g^0	**205**	**941**	**1434**	**1989**	**1**	**0**	28	20	34	14	**3**	**0**	**100**
35	**0**	**5**	**0**	**0**	**10**	15	30	**5**	**100**	TR	C	**69**	**695**	**1494**	**2001**	**0**	1	**1**	**6**	20	14	33	25	**100**
35	**5**	**10**	30	**0**	**0**	**10**	**0**	**10**	**100**		g^0	**205**	**1576**	**1747**	**2631**	**6**	**16**	22	**2**	**13**	**10**	**8**	23	**100**

4.3 Discussion

The perception experiments demonstrated that some, though not all, tones were reliably associated with vowels. This was definitively the case for the tones c3 as /i/ and G as /ø/ in SIM. The association rates for these two tones as human vowels were similar to the recognition rates of human vowels produced in CV and VC English syllables (Weber and Smits, 2003), where some vowels reached recognition rates as low as 45%. This is particularly evident for vowels that are not at the periphery of the vowel system. This finding can be compared with some of our results, for instance that of c1 for SIM. In both listening tests, this particular tone was associated with an even distribution between /e/, /ɛ/, /ø/ and /a/ as an area covering non-high and non-back vowels. Interestingly, all but one of about twenty visitors to the poster presentation at the Interspeech conference (with various language backgrounds) associated c1 of SIM with /ɛ/ when listening to it via headphones.

The tones from the *vox humana* in WAL produced less consistent association rates than the SIM tones and those from the other two churches (not reported here). In WAL, the tones were recorded in combination with the flue pipes from the *stopped diapason* leading to a different spectral distribution: the lower harmonics and the fundamental frequency were quite strong compared to the other organs. The two different stops merge to a new synthetic tone colour and were not perceivable separately.

There was a very strong relationship between the F_0 of the tones and their perceived vowel quality, which can be traced back to sound symbolism (Ohala, 1994). However, F_0 alone cannot explain these results. Obviously, the formant structure also plays a role. For instance, in SIM, the tone c3, reliably associated with /i/, showed very high values for F_2 and F_3; whereas G, heard as /ø/, possessed the lowest values for these formants.

It is striking to see that other stops with reed pipes, in our case *crumhorn* and *trumpet,* did *not* show as consistent results as the *voces humanae*, although F_0 and formant structure were also present there. Obviously, a *vox humana* was able to produce more human vowel-like sounds than *crumhorn* and *trumpett* the other organ stops that also use a reed pipe.

5 Conclusion

We could partially replicate the historically documented enthusiastic impression of the *vox humana* as an instrument with which it is possible to play human-like vowels. Although it is not clear how to explain this effect, we could show that *voces humanae* differ from other organ stops with reed pipes in terms of similarity to the human voice. This is interesting because von Kempelen (1791) used an excitation mechanism similar to a reed pipe in his famous speaking machine (see Kempelen (1791) for the original text and e.g. Brackhane (2011) for a historical reception).

Since we focused on isolated tones in this project, we cannot say anything about the influence of temporal and intensity dynamics, which can possibly explain the *vox humana* as a vowel synthesiser to a certain degree. The desire to generate

isolated vowels with the help of separate *vox humana*–like pipes was also required in the second part of the prize question of the St. Petersburg academy in 1780: "1) Qualis sit natura et character litterarum vocalium a, e, i, o, u tam insigniter inter se diversorum. 2) Annon construe queant instrumenta ordini toborum organicorum, sub termin vocis humanae noto, similia, quae litteratum vocalium a, e, i, o, u, sonos exprimant. (What is the nature and character of the vowels a, e, i, o, u, which are so different from each other? Is it possible to construct an instrument like the organ pipes called *vox humana* that can produce the vowels a, e, i, o, u?)" [translation of the authors from Kratzenstein (1781)]. Kratzenstein won the prize by producing /a, e, o, u/ according to the principles of the *vox humana* with a small organ consisting of four reed pipes, but for /i/ he used a flue pipe. Our study shows that more vowels than those can be convincingly produced with a *vox humana* in an organ, including an /i/.

The *vox humana* is definitively a fascinating musical instrument, which is partially able to generate human speech. However, the *vox humana* is not a genuine mechanical vowel synthesiser as hoped in historical times.

6 Acknowledgements

The authors thank Christoph Draxler for his support with the second perception test as well as Bernd Möbius, Eva Lasarcyk, Peter Birkholz, Coriandre Vilain and John Ohala for feedback on this research. Our thanks go also to the visitors of the poster presentation at Interspeech 2013 in Lyon. We are also grateful to the anonymous reviewer whose comments helped to improve this paper as well as to Ruth Huntley-Bahr.

References

Adelung, W. 1982. *Einführung in den Orgelbau*. Wiesbaden: Breitkopf.

Brackhane, F. 2011. Die Sprechmaschine Wolfgang von Kempelens – von den Originalen bis zu den Nachbauten. *Phonus 16* (Reports in Phonetics, Saarland University), pp. 49-148.

Draxler, Chr. 2012. Percy - An HTML5 framework for media rich web experiments on mobile devices. *Proc. 12th Interspeech*, Florence, pp. 3339-3340.

Eberlein, R. 2007. Vox humana. In: H. Busch and M. Geutig, (eds) *Lexikon der Orgel*. Laaber: Laber-Verlag.

Euler, L. 1773. *Briefe an eine deutsche Prinzessin über verschiedene Gegenstände aus der Physik und Philosophie: Aus dem Französischen übersetzt. Band 2*. Leipzig: Junius.

Fitch, W.T. and J. Giedd 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America 106*(3), pp. 1511-1522.

Frotscher, G. 1927. *Die Orgel*. Leipzig: Weber.

Greß, H. 2007. *Die Orgeln Gottfried Silbermanns*. Dresden: Sandstein.

Kempelen, W. v. 1791. *Wolfgangs von Kempelen Mechanismus der menschlichen Sprache nebst Beschreibung seiner sprechenden Maschine*. Wien: Degen.

Kratzenstein, Chr.G. 1781. *Tentamen resolvendi problema ab academia scientiarium imperiali petropolitana ad annum 1780 propositum*. St. Petersburg: Academia Scientiarum.

Lottermoser, W. 1936. *Klanganalytische Untersuchungen an Orgelpfeifen*. Berlin: Junker &

Dünnhaupt.
Lottermoser, W. 1983. *Orgeln, Kirchen und Akustik. Bd. 1*. Frankfurt/ Main: Bochinsky.
Ohala, J.J. 1994. The frequency code underlies the sound symbolic use of voice pitch. In: L. Hinton, J. Nichols and J. J. Ohala (eds): *Sound Symbolism*. Cambridge: Cambridge University Press, pp. 325-347.
Pompino-Marschall, B. 2009. *Einführung in die Phonetik* (3rd edition). Berlin: de Gruyter.
Simpson, A. 1998. *Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonetischen Theoriebildung*. (Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK) 33). Kiel.
Weber, A., and R. Smits 2003. Consonant and vowel confusion patterns by American English listeners. *Proc. 15th International Congress of Phonetic Sciences*, Barcelona, pp. 1437-1440.