The organ stop "vox humana" as a model for a vowel synthesiser

Fabian Brackhane[1], *Jürgen Trouvain*[2]

[1] Institut für Deutsche Sprache (IDS), Mannheim, Germany
[2]Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany

brackhane@ids-mannheim.de, trouvain@coli.uni-saarland.de

Abstract

In mechanical speech synthesis reed pipes were mainly used for the generation of the voice. The organ stop "vox humana" played a central role for this concept. Historical documents report that the "vox humana" sounded like human vowels. In this study tones of four different "voces humanae" were recorded to investigate the similarity to human vowels. The acoustical and perceptual analysis revealed that some though not all tones show a high similarity to selected vowels.

Index Terms: vowel synthesis, historical instruments

1. Introduction

Many authors of the 18th and 19th century consider the organ stop "vox humana"[1] 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" [1: 246].

However, evidence for a real similarity to vowels is either missing or does not stand standards of today. Based on own 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 known since the middle of the 16th century [2: 817] (see Fig. 1 top). An organ stop is a set of organ pipes with different pitches but of the same construction type. 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, not all of their pipes are visible from outside). The majority of the stops are flue pipes (see Fig. 1 bottom) although reed pipes are also common parts of organs (see Fig. 1 top). A characteristic feature of the reed pipes used in a "vox humana" is the "resonator" that is of a constant size independent of the pitch of the pipe. The "resonators" act as a filter in such a way that formants can be observed similar to those found in human vowels [3: 48, 4: 135].

The term "vox humana" originates from the use of the organ stop substituting the human singing voice. It is usually played not alone but together with the tremulant and the flue stop "bourdon" or also called "stopped diapason" (see Fig. 1 bottom) of the same pitch. The tremulant changes the pressure of the air streaming to the pipes in brief intervals. The so produced sound which resembles the vibrato of a human singing voice has been named "vox humana". Later the name has been transported to the special construction of the reed pipes, however, the knowledge about the etymology has been lost.

Since the 18th century the name "vox humana" for the organ stop was used as a programmatic title rather than as a technical term. This new usage caused organ builders as well as researchers such as Leonhard Euler or Christian Gottlieb Kratzenstein to consider the "vox humana" as the prototype of speech synthesis. There are numerous historical documents in which it is attested that these pipes clearly sound like vowels (e.g. [5: 27]).

3. Recordings and acoustic analysis of various "voces humanae"

3.1. Data

It is our aim to test the historical statements concerning the similarity of the "vox humana" sounds to those of human vowels. This requires recordings of those organs where the stops are historically authentic (and not re-constructed). The research question is whether pipes of a "vox humana" really show formant structures similar to those of human vowels. More specifically we are interested in the question whether certain vowel qualities can reliably be recognised by human listeners.

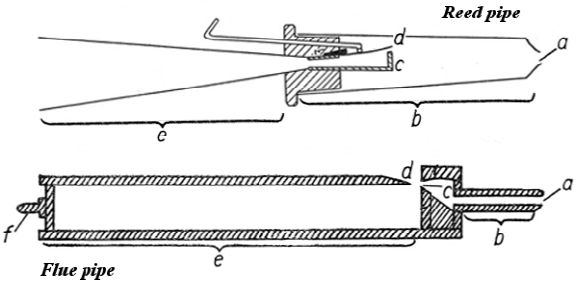

Figure 1: *Schematic drawing of a reed pipe (top, redrawn after [3: 15]) and a flue pipe of the type "stopped diapason" (bottom, redrawn after [6: 43]). The air flows into the pipes (a) passing the socket or boot (b). The air in the reed pipe (top) will be excited by the reed tongue (d) that lies on 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 pipes.*

[1] The terminology in organ building in English, French and German can evoke confusions with the one used in the speech sciences. An "organ stop" corresponds to "jeu d'orgue" in French and to "Register" in German. "Reed and flue pipe" corresponds to "jeu d'anche and "jeu à bouche" in French and to "Zungen-" and "Labialpfeife" in German.

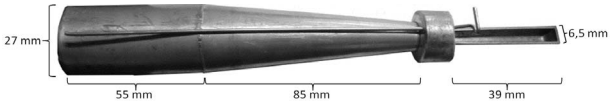

Figure 2: *Reed pipe from the "vox humana" (Tone g^0) from Amorbach (1782), without the boot (cp. Fig 1 top) and without the reed tongue (cp. (d) Fig 1 top).*

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

In addition, the "vox humana" of the organ of the Stadt-kirche in Waltershausen (Thuringia, Eastern Germany) was recorded (WAL henceforth). This organ stop is a copy of a "vox humana" from a monastery in Weingarten (1750) which is famous for its constructor, Joseph Gabler, who attempted to build pipes with a sound that resembles human singing voices in a particular way. The resonators of these pipes were adopted to human larynges.

The tones C, G, c^0, g^0, c^1, g^1, c^2, g^2 and c^3 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 re-constructed) reed pipe stops "trumpet" and "cromorne" (for C and g^0). These two stops substantially differ from the "vox humana" in their construction styles and they were recorded for reasons of comparison with the "vox humana" of the same organ and the measurements found in [4]. Only two tones were selected: C as the lowest one and g^0 because it has been described as particularly vowel-like (see e.g. [7: 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 is 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". The microphone was placed in a distance of about half a metre above the resonators to have comparable recordings in the acoustically different churches and to reduce the echo effect 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 show durations of about 5 seconds. This extensive duration is due to the fact that the reed pipes need a rather long time before reaching 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 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 differ in their F_0 for virtually all tones (see Table 1). As an example the tone G, comparable to a bass voice, ranges from 98 Hz in AMO to 105 Hz in MEI.

All tones of all "voces humanae" show clear formant structures. This is also true for the additional stops "trumpet" and "cromorne" (see Fig. 3). However, the formant shapes of the "voces humanae" show more similarity to formants typical of formants of human speech. Interestingly, in all four organs the values for F_0 and F_1 converge or even merge for the two and sometimes three highest tones which makes a visible distinction nearly impossible.

Figure 4 displays the location of F_1 and F_2 for the "voces humanae" of SIM and WAL. It is visible that the formant distribution of the tones from WAL mainly reflects changes of F_1 (from 400 to 1300 Hz) whereas the tones from the SIM organ show a larger variation of F_2. Compared to the formant space of human (male) voices (German speakers for the long, tense vowels [8]) both organs show a smaller space. In addition, the organs' vowel spaces have higher average formant values than the human vowel space. This formant shift is illustrated in the very small overlap of the spaces of the SIM "vox humana" and the human voice.

The inspection of the values of F_3 reveals a much wider formant range for the organs compared to a male voice. For instance F_3 of the SIM organ ranges between 1900 and 2800 Hz, WAL between 3200 and 2000 Hz whereas the F_3 of the human voice ranges between 2200 and 2500 Hz.

For two tones, c^1 from MEI and SIM, respectively, the maximal energy was found on the 7th harmonic (at around 1970 Hz). This is in line with a previous study [4: 135] (on the acoustics of reed pipes) for the tone C. However, the maximal energy of all other tones from AMO, MEI und SIM was 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.

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
c^0	132	141	136	139
g^0	197	210	205	208
c^1	263	281	274	277
g^1	395	411	408	415
c^2	527	562	548	554
g^2	790	844	818	832
c^3	1054	1124	1093	1108

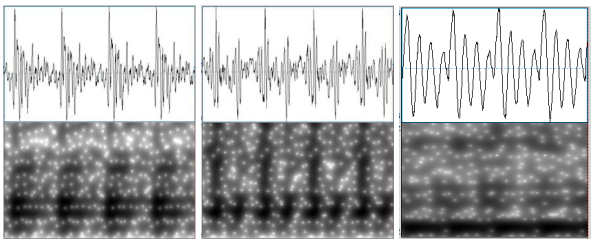

Figure 3: *Waveforms and spectrograms of sections with four periods taken from the tone C of the stops "vox humana" (left) and "cromorne" (middle) (duration: 60 ms) and from the vowel [ø:] of a male German speaker (right; duration: 42 ms, F_0: 97 Hz).*

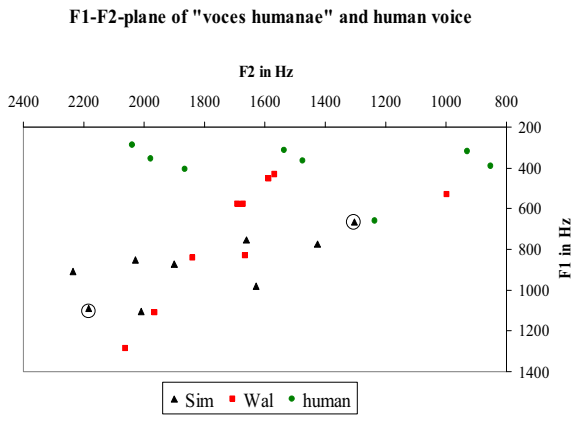

Figure 4: *Values for F_1 and F_2 of the tones of the "vox humana" in Simmern (black triangles) and Walters-hausen (red squares) as well as standard values for the German long vowels of male voices [8] (green dots). The encircled triangles indicate well recognised vowel qualities: [i] on the left, [ø] on the right.*

3.3. Discussion

The differences of the organs concerning their fundamental frequencies for the same tone can be explained by the fact that in the 18th century the fundamental 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.

It could be shown that a central feature of the reed pipes from different "voces humanae" is their formant-like structure. But formants were also found for the reed stops "cromorne" and "trumpet. It is unclear whether further stops, especially flue stops, which account of normally ⅔ of all pipes in an typical organ, also show formants. Moreover it is unclear which reed pipes show the largest similarity to the formants of human vowels.

The formant values of the "voces humanae" shaped a vowel space which is smaller in size and with more upshifted formant values in comparison to a human speaking voice. Future experiments with formant synthesis could show whether the measured formants from the organs can generate acoustic forms which sound like humanoid vowels for the perceiver.

The spectral distribution is partially different to the one of human vowels. As already described in [4: 135] the maximal energy of the highest tones can be found on the 7th harmonic in the solo played "voces humanae" whereas the fundamental frequency was nearly never found to be the strongest harmonic (2 exceptions out of 27 tokens). The "vox humana" in WAL was played in combination with another stop which caused the fundamental to be the strongest harmonic.

4. Perception tests

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

4.1. Method

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

The second test was similar to the first one but with some changes. This time the test was web-based (with the help of [9]) in order to recruit more subjects (with German as their first language). In total there were 29 subjects, including linguists and non-linguists. The number of stimuli was reduced to 18 (from two "voces humanae") plus the four "others" with the possibility to repeat each stimulus 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 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. Letters allow more consistent answers. The answer "no vowel" was not possible this time. For technical reasons one stimulus was not correctly played (c^0 from WAL). Consequently the corresponding results will not be presented.

4.2. Results

The results (see Table 2) for both "voces humanae" clearly indicate the correlation between the fundamental frequency and the vowel category. The higher the F_0 the more [i]-like the selected vowel, 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_2) were assessed more consistently than those in the middle region. This is obvious for instance for the SIM tones in the second experiment with a very stable tendency of c^3 for [i] (84%) but a far less consistency for the next lower tone g^2 (between [i] and [e], with a tendency to [i]). The tone c^1 is more or less equally distributed between qualities of [e], [ɛ], [ø] and [a]. For the corresponding tone of WAL the listeners mostly preferred [a/ɑ] (test 1) or even [u] (test 2).

The tones from the comparative stops "cromorne" and "trumpet" show less consistent answers than those of the "voces humanae", especially for the tone g^0. Comparing the results of the organs of SIM and WAL it becomes evident that the tone-vowel correspondences of SIM show a higher level of consistency than those of WAL (except for C and the maverick answer for c^1).

The general tendencies of the first perception test were confirmed by the second but now on a more reliable basis. The results were sometimes clearer (e.g. for c^0 and g^1 in SIM) and often led to a higher level of consistency. However, the differences between both experiments also reveal that the test was not considered an easy task (which was informally expressed by some subjects).

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**	**ü**	**e**	**ä**	**ö**	**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 Waltershausen	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

Table 2. *Percentages of answers for each stimulus tone for both perception experiments. The values for* F_0 *and the formants are in Hz. The stops were "voces humanae" (VH), "cromorne" (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). The stimuli of the second experiment can be found under "additional files".*

4.3. Discussion

The perception experiments show that *some* though not all tones were reliably associated with vowels. This is definitively the case for the tones c^3 as [i] and G as [ø] in SIM. The recognition rates for these two tones/vowels are similar to those of human vowels produced in CV and VC English syllables [10] where some vowels reached recognition rates as low as 45%.

The tones from the "vox humana" in WAL reached less consistent recognition 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.

There is a very tight relationship between the F_0 of the tones and their perceived vowel quality which can be traced back to sound symbolism [11]. However, F_0 alone cannot explain the results, obviously the formant structure also plays a role. For instance, in SIM, the tone c^3, reliably associated with [i], shows very high values for F_2 and F_3 whereas G, heard as [ø], features the lowest values for these formants.

It is striking to see that other stops with reed pipes, in our case "cromorne" and "trumpet" did *not* show as consistent results as the "voces humanae", although F_0 and formant structure is also present there.

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 not all details are clear of 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 insofar interesting in that von Kempelen used an excitation mechanism similar to a reed pipe in his famous speaking machine [12, 13].

Since we focused on isolated tones here we cannot say anything about the influence of temporal and intensity dynamics which possibly explains to a certain degree the "vox humana" as a vowel synthesiser. Isolated tones from a "vox humana" were also required in the second part of the prize question of the St. Petersburg academy in 1780: "Is it possible to construct an instrument like the organ pipes called "vox humana" that can produce the vowels a, e, i, o, u?" (own translation from [14]). Kratzenstein won the prize by producing [a, e, o, u] according to the principles of the "vox humana" with four reed pipes but for [i] he used a flue pipe. Our study shows that more vowels than those can convincingly be produced with a "vox humana". The "vox humana" is definitively a fascinating musical instrument which is partially able to generate human speech production. However, the "vox humana" is not a genuine 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 and Eva Lasarcyk for feedback on earlier versions of this paper.

7. References

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

[2] Eberlein, R. "Vox humana", in Busch, H. and Geutig, M. Lexikon der Orgel. Laaber: Laber-Verlag, 2007.

[3] Lottermoser, W. "Klanganalytische Untersuchungen an Orgelpfeifen" Berlin: Junker & Dünnhaupt, 1936.

[4] Lottermoser, W. "Orgeln, Kirchen und Akustik" Bd. 1. Frankfurt/ Main: Bochinsky, 1983.

[5] Greß, H. "Die Orgeln Gottfried Silbermanns" Dresden: Sandstein, 2007.

[6] Adelung, W. "Einführung in den Orgelbau" Wiesbaden: Breitkopf, 1982.

[7] Frotscher, G. "Die Orgel" Leipzig: Weber, 1927.

[8] Simpson, A. "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. 1998.

[9] Draxler, Chr. "Percy - An HTML5 framework for media rich web experiments on mobile devices", in Proc. 12th Interspeech, Florence, 3339-3340, 2011.

[10] Weber, A. and Smits, R. "Consonant and vowel confusion patterns by American English Listeners" in Proc. 15th International Congress of Phonetic Sciences (ICPhS 2003). 1437-1440, 2003.

[11] Ohala, J. J. "The frequency code underlies the sound symbolic use of voice pitch", in L. Hinton, J. Nichols & J. J. Ohala [Eds], Sound symbolism. Cambridge: Cambridge University Press. 325-347, 1994.

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

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

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