Singing synthesis and the Vocal Tract Organ

David M Howard
York Centre for Singing Science, Audio Lab,
Department of Electronics, University of York, YO10 5DD, UK
david,howard@york.co.uk

Abstract

Vocal synthesis has been the subject of investigation since the late 18th century when von Kempelen produced his mechanical 'speaking machine'. The advert of electronics has enabled a number of different methods of voice synthesis to be realized in practice. Recently with the advent of 3-D printing and magnetic resonance imaging of human vocal tracts, it has been possible to create synthetic vocal sounds that combine both mechanical (3-D printed tracts) and electronic (synthesized larynx sound source) to enable the effects of various parts of the vocal tract on the acoustic output to be investigated. Given that the 3-D tracts look rather like organ pipes, the author (an organist) has developed a new musical instrument based on this technology, which is called the Vocal Tract Organ. This paper reviews voice synthesis techniques and describes the structure and operation of the Vocal Tract Organ.

Human voice synthesis

The human voice production system consists of three elements [1]: the power source (breathing), the sound source (the vibrating vocal folds in the larynx for pitched sung sounds) and the sound modifiers (the varying resonant acoustic properties of the tubes of the throat, mouth and nose above the larynx). For the purposes of considering voice synthesis, only the sound source and sound modifiers are relevant since the power source for electronic voice synthesis is electrical rather than air flowing from the lungs.

One of the earliest examples of a speech synthesiser is Baron Von Kempelen's "speaking machine" in the 1790s. This was a mechanical model of the human speech production system, which was "played" by a human operator. An original exists in the Deutsches Museum in Munich and the author's modern replica is shown in figure 1. Its power source is the under-arm bellows, its sound source is a vibrating reed and its sound modifiers is the leather tube controlled by hand representing the mouth. In addition, it has additional outputs for the production of the consonants in 'sea' and 'she'.

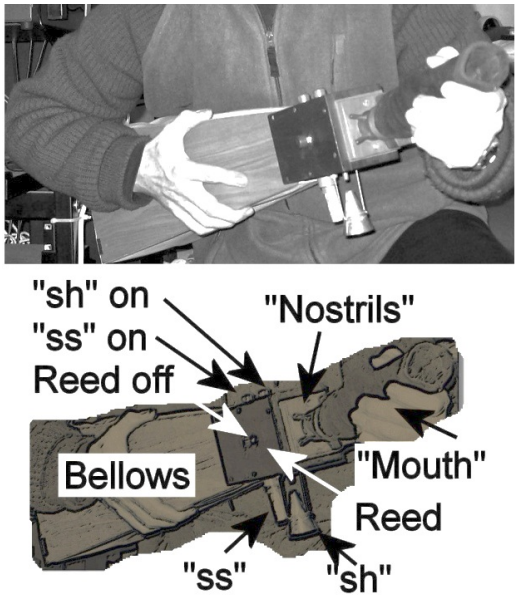

Figure 1: The author's replica Von Kempelen machine (top) and labelled version (bottom).

With the advent of electronics, a number of different approaches have been adopted for electronic speech synthesis ranging from modifying natural speech recordings to direct modelling of the physiological processes involved in speaking. Styger and Keller [2] provide a useful summary of speech synthesis methods under four categories shown in table 1. These approaches are analysed on the basis of two scales using a star rating in the table:
(A) *flow of control parameters* (the required sampling rate), and
(B) *model complexity* (speech production knowledge required).

For the manipulation of natural speech waveforms (1), the required *control parameters* are at a maximum of once per fundamental period, but the *model complexity* is at a minimum. For vocal tract articulatory physical modelling (4) the required *control parameters* are at a minimum of once per articulatory gesture, but the *model complexity* is at a maximum requiring detailed knowledge of vocal tract articulation.

#	Category	control parameters	model complexity
1	**Manipulation of natural speech waveforms** *no knowledge of speech production mechanism*	****	*
2	**Linear predictive synthesis** *all-pole acoustic model*	***	**

	of the vocal tract		
3	**Formant synthesis** *formant parameters are varied*	**	***
4	**Vocal tract articulatory physical modelling** *control of articulation itself*	*	****

Table 1: Organisation of speech synthesis methods by from Styger and Keller [2] in terms of the *flow of control parameters* and *model complexity* to illustrate key differences between available methods.

Typical experiences with today's electronic voice synthesis are that they can produce a highly *intelligible* speech output, but that it is rare if ever that an electronically synthesized voice based on methods 2-4 is mistaken as having emerged from a human vocal tract. Few if any of today's synthesizers are able to produce a *natural* sounding speech output. Formant synthesis (3) is the most commonly used method and it has been used for a long time [2-4], which is based directly on the source/filter model [5] of speech production.

The "Festival" system [6-7], in which recorded speech waveforms are manipulated to create connected speech, is a popular basis for electronic speech synthesis. The key issue is keeping the joins transparent to the listeners

.

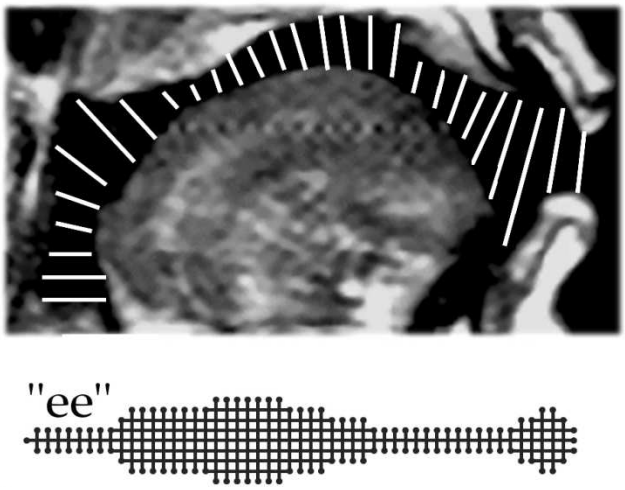

Figure 2: Magnetic resonance image of the vowel "ee" and its 2-D waveguide mesh representation (the larynx is on the left and the lips on the right).

Synthesis based on vocal tract articulatory physical modelling (4 in table 1) recreates vocal tract articulation behaviour to simulate directly the process of speaking itself, rather than the sounds it generates. Sound is created via physical modelling of acoustic pressure in the vocal tract with an appropriately placed sound source input for pitched and non-pitched sounds. Physical modelling of acoustic pressure in the vocal tract was first based on the 1-D digital waveguide [8], and it can be extended to 2-D or 3-D [9-10], but the computation load increases rapidly, making the running of a 3-D model impossible now in real-time on today's PCs (up to an 1 hour's processing can be required for 1 second of speech output). A magnetic resonance image (MRI) for the vowel "ee" along with its waveguide mesh layout is shown in figure 2; note that the larynx is on the left and the lips on the right.

The Vocal Tract Organ

The vocal tract is a tube and a visual link with the pipes of a pipe organ is readily made, despite the bend in the vocal tract. In particular, it has the potential to challenge the *Vox Humana* stop that is found in a number of large pipe organs, but which typically sounds most unlike the human voice! The complete 3-D vocal tract dimensions can be measured from an MRI session based on a set of MRI pictures taken across the vocal tract. From these a 3-D print can be created of the vocal tract, which will be an accurate representation of whatever sound was being articulated (and held steady for around 16 seconds) in the MRI scanner. If the larynx end of the 3-D print is set up to couple with a suitable loudspeaker, it can be made to sound if an appropriate larynx sound source excitation is provided to the loudspeaker. An example 3-D print coupled to a loudspeaker drive unit (Adastra 952.210) is shown in figure 3.

The prototype Vocal Tract Organ consists of six 3-D printed Vocal Tracts for the same vowel, but with slightly different lengths (each differs from its neighbour by 2.5 mm to ensure that the outputs are not absolutely identical) with their loudspeaker drivers. Each printed vocal tract can be made to sound if a suitable voice source signal is applied at the position of the larynx in the neck [1]. The larynx sound source waveform is a close approximation to that observed from the vibrating vocal folds, and its practical implementation is based on the Liljencrants/Fant (LF) glottal source model [11], which is synthesised in practice using Pure Data, or *PD*, [12]. PD is well suited to this because it enables a wavetable synthesiser to be implemented that is based on either one cycle that is either (a) calculated from a set of harmonic amplitudes (a pulse and a sawtooth waveform is available in the system that is based on these; the user can switch between them), or (b) drawn by hand using the mouse (this is how the LF model is implemented enabling changes to its shape to be easily tested). The implementation of this glottal source for the Vocal Tract Organ for multi-part, or *polyphonic*, synthesis is described in [13]. In order that the result is perceived as being close to a natural output, each channel has a separate setting for vibrato rate, vibrato depth and volume. An

overall volume control is also included which can be set using the mouse and a slider or externally manipulated via a MIDI control parameter. These can be set independently either via an on-screen slider with the mouse or over MIDI (Musical Instrument Digital Interface) via any programmable MIDI controller device. The organ is played via a MIDI keyboard.

Figure 3: A 3-D print for the vowel “ah” sitting atop its loudspeaker drive unit as used in the Vocal Tract Organ.

Four-part chorale style music can be played on the Vocal Tract Organ (quasi soprano, alto, tenor, bass, or SATB), an implementing 6-channels means that there are two spare channels to ‘catch’ additional notes during *legato* playing between chords. Clearly this could be changed as appropriate to performance

needs. The six audio outputs are routed to the loudspeaker drivers via an RME Fireface 400 multi-channel digital to analogue converter, via six amplifiers.

The author composed two pieces to demonstrate the Vocal Tract Organ, specifically to enable its output to be compared directly with live singers. In the first, a barbershop-style vocalize called *Vocal Vision II*, two male singers sing two parts and two other parts are played on the Vocal Tract Organ. A performance of *Vocal Vision II* can be seen and heard on YouTube [10]. In the second, the Vocal Tract Organ accompanies a solo soprano singing an opera aria. The original stimulus for this was a black-tie after-dinner event in the presence of a member of a member of the British Royal Family for which a short *flash mob* opera aria was required highlighting an advance in engineering in a musical context. *O mio babbino caro* (Puccini) was sung by a mezzo-soprano, for which a new chorale-like keyboard accompaniment was created. Filming was not allowed at the first performance, but it was at a second a few weeks later and this can be viewed on YouTube [11] (from 2m50s following a brief presentation about the organ).

Conclusions

Vocal synthesis is highly intelligible but typically non-natural and likely to be improved by moving towards computed models of articulation inhuman voice production. Such models are based on magnetic resonance imaging of the vocal tract, which provides images that can be 3-D printed. These 3-D prints look like organ pipes, and when placed on suitable loudspeakers drivers, they can be made to sound like the vowels of the original speaker. A Vocal Tract Organ has been developed that uses a set of such 3-D prints; a modern-day *Vox Humana* organ stop that offers new possibilities for music performance.

Acknowledgements

The author thanks the staff at the York Neuroimaging Centre for their help in capturing the images, and Pete Turner for help with the 3-D print creation and electronic implementation of the Vocal Tract Organ.

References

[1] Howard, D.M., and Murphy, (2008). D.T.: Voice science, acoustics and recording, Plural Press, San Diego.
[2] Styger, T., and Keller, E. (1994). Formant synthesis, In: *Fundamentals of speech synthesis and speech recognition*, Keller, E. (Ed.), Chichester: John Wiley and Sons, 109-128.
[3] Holmes, W.J., and Holmes, J.N., (2002). *Speech synthesis and recognition*, London: CRC Press.

[4] Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer, *Journal of the Acoustical Society of America*, **67**, (3), 971-995.
[5] Fant, G. (1960). *The acoustic theory of speech production*, The Hague: Mouton.
[6] Taylor, P.A., Black, A., and Caley, R. (1998). The architecture of the festival speech synthesis system, in *Proc. 3rd ESCA Workshop in Speech Synthesis*, Jenolan Caves, Australia, 147–151.
[7] WWW-1: http://www.cstr.ed.ac.uk/projects/festival/ (last accessed 11th March 2014)
[8] J. L. Kelly and C. C. Lochbaum, (1962). Speech synthesis, in *Proceedings of the 4th International Congress on Acoustics*, Copenhagen, Denmark, 1–4.
[9] Mullen, J., Howard, D.M., and Murphy, D.T. (2006). Waveguide Physical Modeling of Vocal Tract Acoustics: Improved Formant Bandwidth Control from Increased Model Dimensionality, *IEEE Transactions on Speech and Audio Processing*, **14**, (3), 964-971.
[10] Mullen, J. Murphy, D.T. and Howard D.M. (2007). Real-time dynamic articulations in the 2D waveguide mesh vocal tract model, *IEEE Transactions on Speech and Audio Processing*, **15**, 2, 577-585.
[11] Fant. G., Liljencrants, J., and Lin, Q. G.: A four-parameter model of glottal flow, STL-QPSR, **2,** (3), 119-156, (1985)
[12] WWW-1: http://www.puredata.info (last accessed 11th March 2014)
[13] Howard, D.M., Daffern, H., and Brereton, J.: Four-part choral synthesis system for investigating intonation in a cappella choral singing, Logopedics Phoniatrics Vocology, **38**, (3), 135-142, (2013)
[14] WWW-3: http://www.youtube.com/watch?v=pUryWk-s9Ig (last accessed 11th March 2014)
[15] WWW-4: http://www.youtube.com/watch?v=SX-f1oU0_Kk (last accessed 11th March 2014)