DEVELOPMENT OF SOUND SOURCE COMPONENTS FOR A NEW ELECTROLARYNX SPEECH PROSTHESIS

Kenneth M. Houston [1], Robert E. Hillman [2], James B. Kobler [2], Geoffrey S. Meltzner [2]

[1] Draper Laboratory MS-53, 555 Technology Square, Cambridge MA 02139, USA

[2] Massachusetts Eye and Ear Infirmary, 243 Charles Street 11th Floor, Boston MA 02114, USA

ABSTRACT

For many individuals who lose their voices due to laryngeal cancer or trauma, the only option for speech is to use an electrolarynx (EL), which is a battery-powered vibrator that is held to the throat. Current devices produce speech that is very machine-like in sound, with low levels of loudness and intelligibility, that also draws undesired attention to the user. A project at Draper Laboratory, the Mass. Eye and Ear Infirmary and MIT aims to develop a much improved EL called the Electrolarynx Communication System (ELCS), which is a DSP-based device consisting of sound source, control, and speech enhancement subsystems or modules. This paper introduces the ELCS and discusses developments to date in the sound source module. Specific topics include the design of a new linear EL transducer and investigations into glottal waveform synthesis which should result in a much more natural speech output.

1. INTRODUCTION

Every year in the United States alone, thousands of people lose the ability to produce voice and speech because of laryngeal cancer or trauma. For many of these individuals, the only option for speech is to use an electrolarynx (EL), which is a battery-operated vibrator that is held against the throat. Unfortunately, current devices produce speech that is very machine-like in sound, with low levels of loudness and reduced intelligibility, that also draws undesired attention to the user. Draper Laboratory is involved in a collaborative effort with the Massachusetts Eye and Ear Infirmary (MEEI) and MIT called the Voice Project of the W. M. Keck Neural Prosthesis Research Center. The aim is to design a new DSP-based EL called the Electrolarynx Communication System (ELCS) which should offer many improvements in sound quality over previous models.

As Figure 1 indicates, the ELCS has three subsystems or modules: 1) The *Sound Source Module* consists of a waveform generator, power amplifier and a linear shaker transducer, and represents the complete functionality of current ELs. 2) The *Sound Source Control Module* provides pitch and amplitude control to the sound source based upon neural inputs. We envision that when the larynx is removed in the future, the severed laryngeal nerve will be transposed (implanted) into a strap muscle near the skin surface. Once the nerve regenerates, the muscle will act as an amplifier of neural signals. Electromyographic (EMG) electrodes on the skin surface will detect the redirected laryngeal nerve activity and control signals will be derived, hopefully much in the same way that control signals for prosthetic limbs are obtained today. 3) The *Speech Enhancement Module* is a real-time enhancement system to further improve the output quality. At a minimum, it will perform low-frequency emphasis and amplification to a loudspeaker when used in noisy environments. Many other improvements are possible, such as the correction for speech distortions caused by alterations to the vocal tract by the laryngectomy operation.

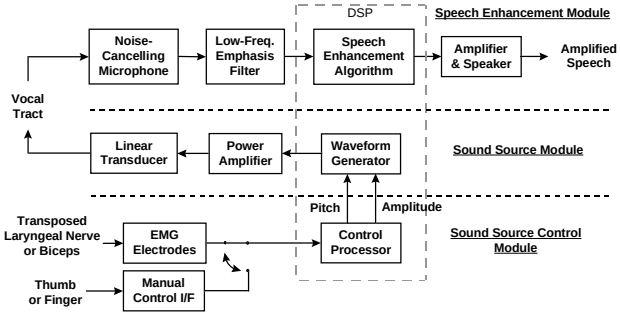

Figure 1. Block diagram of the new Electrolarynx Communication System.

2. LINEAR TRANSDUCER DESIGN

Figure 2 shows a non-linear transducer which is representative of current EL designs. An armature pulsating at the pitch frequency is made to strike a coupler which is held to the throat. The coupler conducts the impulses into the pharynx. The coupler's mechanical characteristics control many aspects of the resulting speech spectrum. Non-linear transducers inherently limit EL designs in the following ways: 1) There is generally a low-frequency deficit below approximately 500 Hz which makes certain vowels hard to distinguish, 2) the spectral envelope is difficult to control, 3) there is a very high level of self-noise, which represents a constant interference to the desired signal, filling in spectral and temporal "valleys" where sound should be absent, and 4) there is a lack of variation in the harmonic structure, giving the sound a metallic and machine-like quality. Developing a linear transducer for an EL is critical because this allows use of arbitrary driving waveforms. Purely electronic waveform synthesis allows for rapid responses to control inputs, permits adjustment of the spectrum as desired, and enables inclusion of features which improve the naturalness of the resulting sound.

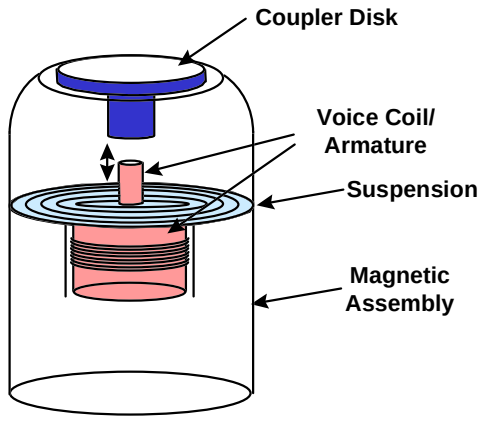

Figure 2. Representative non-linear transducer used in current electrolarynx designs. An armature pulsating at the fundamental pitch frequency is caused to strike a coupler disk which is held to the neck. The spectral characteristics output speech are determined by the mechanical characteristics of the coupler assembly.

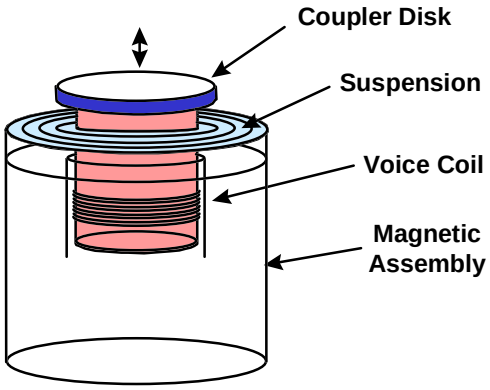

Figure 3. Notional view of linear EL transducer under development, which draws heavily upon loudspeaker technology. The coupler disk which is held to the neck is attached to the voice coil cylinder. The linear nature of the device allows use of electronic waveform synthesis which should result in substantially improved sound quality.

Figure 3 diagrams the new linear transducer. Owing to the similarity to moving-coil loudspeaker technology, a loudspeaker manufacturer is fabricating initial prototypes at the time of this writing. Figure 4 shows equivalent circuits for the transducer [1]. Like a loudspeaker, an electromechanical model defines a motor constant ϕ_M=BL that transforms between electrical to mechanical domains using Force-Voltage/Velocity-Current analogies. Unlike a loudspeaker, the neck mechanical impedance represents the load rather than acoustic radiation alone. (Ideally, acoustic radiation results only when the vibrating pharynx wall interacts with air inside the throat to set up a sound wave - the resulting volume velocity should replicate a normal glottal source. The additional loading due to acoustic radiation is small relative to the neck impedance and can be ignored.)

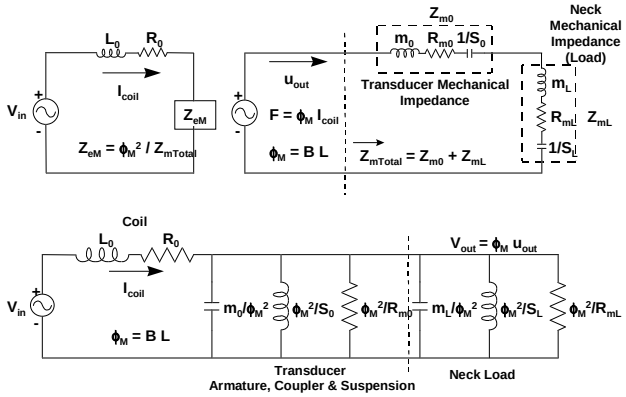

Figure 4. (top) Linear transducer electro-mechanical equivalent circuit diagram. The mechanical impedance of the neck serves as the load to the device. (bottom) Purely electrical equivalent circuit.

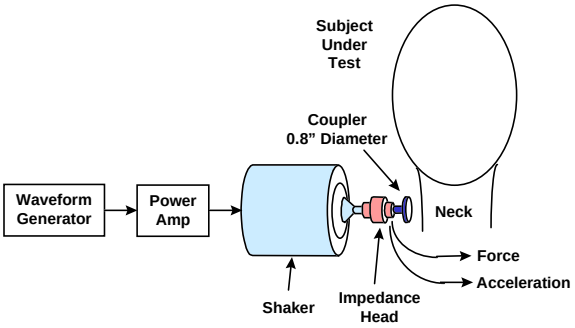

Figure 5. System for the measurement of the neck mechanical impedance. The impedance head is a device which simultaneously measures force and acceleration.

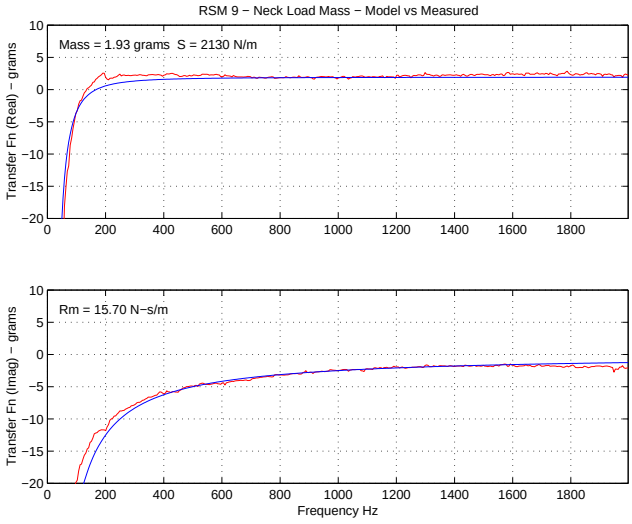

Figure 6. (top) Representative plot of real part of Force/Acceleration ratio, measured and best-fit model (m_L - S_L/ω^2). (bottom) Imaginary part of Force /Acceleration ratio, measured and model (- j R_{mL}/ω).

Parameter	Observed Range Min	Max	Design Value	Units
Load Mass m_L	1.1	1.9	1.8	grams
Mechanical Resistance R_{mL}	8	19	16	N-s/m
Spring Constant S_L	1.5	8	3.0	N/mm

Table 1. Summary of estimated neck mechanical parameters for limited test of 7 subjects (including 4 laryngectomees). The "design value" column represents nominal values used in the transducer design.

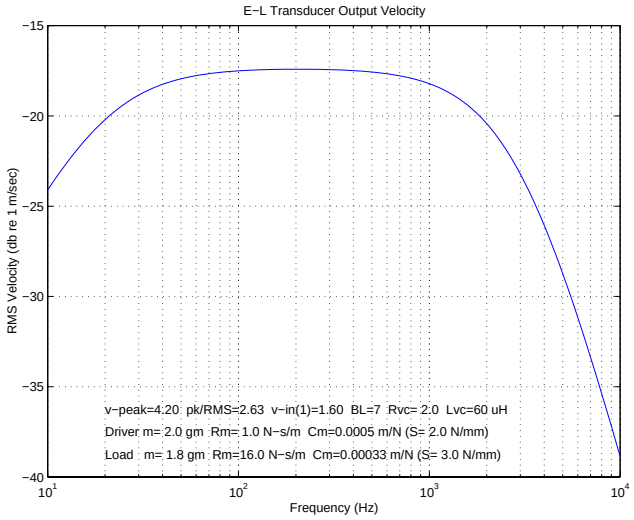

Figure 7. Expected transducer frequency response for a 2.63 Vrms swept sinusoid under nominal load conditions.

In order to properly specify the load so that the transducer could be designed, measurements were conducted on a very limited set of subjects - 3 male larygnectomees, 1 female laryngectomee, and 3 male non-larygnectomees. Figure 5 shows the test setup. An electrodynamic shaker was driven with white noise into a coupler which was placed against the subject's throat. An impedance head sensor measured axial force and acceleration. The transfer function gives the *apparent mass* $M_L(j\omega)$, which for a series Mass-Resistance-Spring combination equals

$$M_L(j\omega) = \text{Force/Acceleration} = m_L - S_L/\omega^2 - j\, R_{mL}/\omega\ ,$$

where m_L = mass in kg, R_{mL} = mechanical resistance in N-s/m (equivalent to kg/sec or "mechanical ohms"), and S_L = spring constant in N/m (sometimes specified as the compliance $C_{mL}=1/S_L$). The *mechanical impedance* $Z_{mL}(j\omega)$ is the ratio of force to velocity, so

$$Z_{mL}(j\omega) = \text{Force/Velocity} = j\omega M_L(j\omega) = R_{mL} + j(\omega m_L - S/\omega) \text{ N-s/m.}$$

Figure 6 shows a representative plot of the real and imaginary parts of the measured transfer function for $M_L(j\omega)$, and best-fit curves. It may be seen that the first-order series mass-spring-resistance model provides a reasonable fit. Table 1 summarizes the measured parameters, which should be valid over approximately 50-2000 Hz. It is interesting to note that the moving mass in the neck is in the 1-2 gram range, or less than the mass of a US penny (2.6 grams). No significant differences between laryngectomees and non-laryngectomees were noted in our limited sample. The expected nominal velocity frequency response for a 2.63 Vrms swept sinusoid excitation is shown in Figure 7. The device should have a flat response over 20-2000 Hz, so the full audio band can be realized (subject to power budget limitations at low frequencies and appropriate equalization at high frequencies). With the indicated velocity (-17.3 dB re 1 m/sec or 0.14 m/sec rms), speech outputs of approximately 85 dBA are expected.

3. WAVEFORM GENERATOR

As mentioned above, it is desired to set up a sound wave within the pharynx which closely matches a normal glottal excitation. The user simultaneously manipulates the vocal tract in the same way as in normal speech to produce a speech output at the lips. The waveform generator should therefore produce some approximation of a glottal source waveform, appropriately compensated for distortions introduced by the transduction process.

The literature is replete with glottal waveform models [2]. An early model is the Rosenberg model [3], shown in Figure 8. When such a waveform is played through a linear EL transducer, the resulting sound is somewhat better than a conventional EL, but is still highly objectionable: the sound is metallic and machine-like. This is primarily because the waveform is defined over a single cycle and repeated, and as a consequence, all harmonics are in lock-step with the fundamental. Other glottal models with more sophisticated parameterization of a single cycle suffer from the same problem, even if noise is added or the "arrival times" of the impulses are dithered.

To obtain a rich, natural sound (whether synthesizing voice or musical instruments), a proper harmonic structure is required where the overtones drift in frequency relative to the fundamental [4]. A simple way to capture the harmonic structure is record a voice and inverse filter, as shown in Figure 9. This is analogous to waveform sampling in musical instrument synthesis (known to produce high quality results), except in the case of voice, the effect of the vocal tract must be removed. A held vowel sound (such as /e/ in "bet") is recorded for several seconds, and is subsequently LPC-analyzed using a high order filter (N=41) and inverse filtered to obtain a whitened residual. Pitch variations are then smoothed through interpolation and a low pass filter (-12 dB/octave) is applied. An example of the result is shown in Figure 10. As can be seen, while similar to Figure 8, there is considerable irregularity from cycle to cycle. When the inverse filtered waveform is applied to the waveform generator in Figure 11 and a linear transducer, the metallic quality completely disappears, and the speech in fact retains many of the qualities of the original speaker. Note that the table of glottal samples must be of a certain minimum length (>2 seconds), or else the periodicity associated with the table length is quite noticeable.

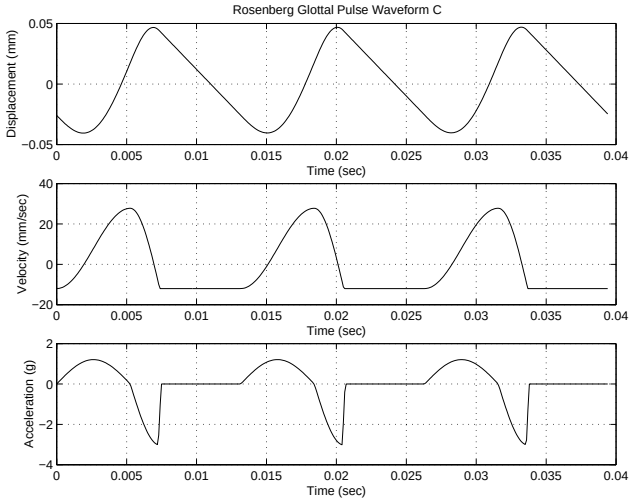

Figure 8. Displacement, velocity and acceleration curves for an EL excitation based upon the Rosenberg glottal pulse type C [3], scaled to 1 g rms acceleration. This excitation results in a very unnatural, metallic speech quality.

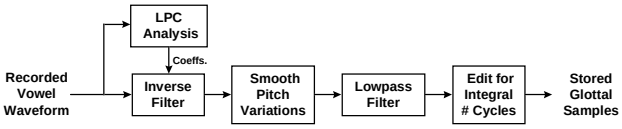

Figure 9. Block diagram of inverse filtering procedure to create lookup table for waveform generation.

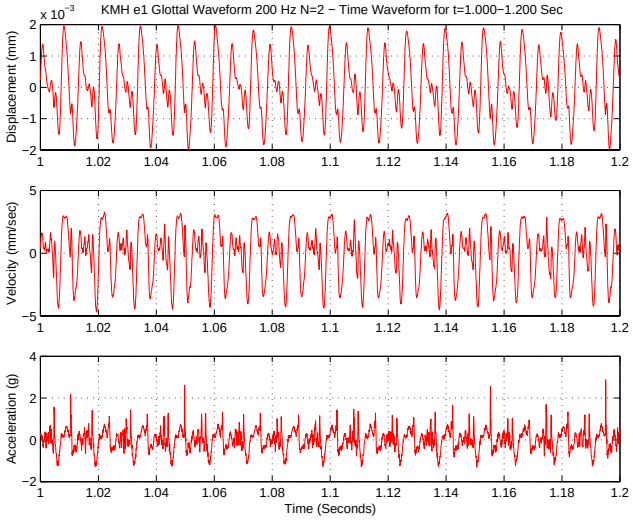

Figure 10. Displacement, velocity and acceleration curves for an EL excitation based upon inverse filtering of a recorded vowel, also scaled to 1 g rms acceleration. This excitation sounds much more natural, and even retains qualities of the original speaker. Time scale is extended to emphasize non-stationary nature.

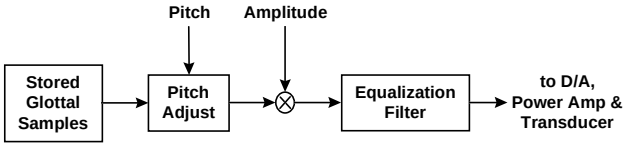

Figure 11. Block diagram of waveform generator for the new Electrolarynx Communication System.

The inverse filtering approach has interesting implications. If the user were to have a voice recording taken before the laryngectomy operation (hopefully well in advance before disease affects the voice), the EL could be customized to that voice. The user could therefore maintain some degree of individuality in the voice and hence reduce some of the hardship currently endured. Alternatively, the voice of a close relative might be adapted, or the user might select from a catalog of voices.

4. CONCLUSIONS

Considerable progress has been made in the design of source module components for the ELCS, which should enable the construction of a source-only EL prototype in the near future. This alone should offer a significant improvement over current EL devices. Work on the other modules is progressing. Of particular note is that a first human trial of a laryngeal nerve transposition recently took place at MEEI, which should enable work to commence on the processing of EMG signals to obtain pitch and amplitude control. If reliable control signals can be obtained, even more significant improvements to speech quality should be possible.

5. REFERENCES

[1] Kinsler and Frey, *Fundamentals of Acoustics, Third Edition.* Wiley and Sons, 1982.

[2] Cummings, K.E. and Clements, M.A. "Glottal Models for Digital Speech Processing: A Historical Survey and New Results", *Digital Signal Processing*, **5** 21-42 (1995).

[3] Rosenberg, A. "The Effect of the Glottal Pulse Shape on the Quality of Natural Vowels", *J. Acoustical Society of America*, **49** 2 (Part 2), 1971, 583-590.

[4] Ramalingam, C.S. and Kumaresan, R. "Voiced-Speech Analysis Based Upon the Residual Interfering Signal Canceller (RISC) Algorithm", *IEEE ICASSP 94*, p 473-6.